Imperial College London
Department of Electrical Engineering

Brain Machine Interfaces: Low Power Techniques for CMOS Based System Integration

Lieuwe Berend Leene

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Engineering, May 2016

Declaration of Originality

I hereby certify that all materials presented in this thesis which is not my own work has been properly acknowledged.

Lieuwe B. Leene

Copyright

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Lieuwe B. Leene

Abstract

The emergence of miniaturized electronic sensors for recording neural activity is opening up new opportunities for better health care and understanding brain function. The precise instrumentation for sensing these signals has been developed extensively, but no implantable system available today is capable of providing a high density recording structures that can be scaled to accommodate the large number of electrodes and processing neuro-prosthetics need for functional limb replacement. The design of these systems is complicated by micro-volt levels of signal that contain convoluted mixtures of information. This demands highly accurate signal quantization and exhaustive processing that is constrained by the scarce power availability. The resulting difficulty in realizing viable solutions for chronic implants necessitates cutting-edge fabrication technologies and state-of-the-art circuit optimization techniques.

This thesis presents the understanding behind optimizing these instrumentation systems in order to maximize the simultaneous sensing capabilities of brain machine interfaces that can be implanted wirelessly into living systems. These analytics enabled this work to outperform state of the art in terms of delivering high precision at 56 dB SINAD with a sub 0.01mm^2 silicon footprint and a 800 nW power budget by employing novel time-domain circuit techniques. This advancement will enable BMIs to be integrated & minimutrized using nanometre CMOS with extensive digital processing capabilities that are capable of decoding neural signals without supervision such that therapy in a fully implanted fashion. Moreover by introducing distributed processing architecture this work is the first to allows scalable fully reconfigurable functionality at the instrumentation interface for complex algorithmic operations while maintaining a power efficiency of 2.7μW per MIPS.

Acknowledgements

I would like to express my sincere gratitude for the opportunities and guidance that Dr. Timothy Constandinou has given me. It has been a wonderful experience to peruse my passion for electronics and do research guided by his generous supervision. He has made working with the next generation neural interfaces group has been a genuine privilege.

I also extend my thanks to my colleges at the centre for bio inspired technology for providing an amazing environment in B422. It is inspiring to see so much talent come together and find ways to work and have fun together. I especially want to thank Tor Sverre Lande for his the inspiration and insightful discussions.

Finally I would like to thank my friends and family for their support and encouragement.

Contents

Abstract ii

Acknowledgements iii

Contents v

List of Tables xi

List of Figures xiii

List of Algorithms xxiii

1 Introduction 1

1.1 Motivation 2

1.2 Research Objectives 6

1.3 Thesis Outline 7

2 Background 10

2.1 Electrical Sensing of Neural Activity 10

2.2 A Technology Perspective 14

2.3 Integrated Recording Systems . . . 25
2.3.1 Power Density Target for High Channel Count Recording . . . 25
2.3.2 Emerging Technologies . . . 34
2.4 Summary . . . 36

3 Neural Recording Front End Design 38

3.1 Architecture for Neural instrumentation . . . 39
3.1.1 Instrumentation Requirements . . . 40
3.2 Amplifier Principles for Miniaturization . . . 41
3.2.1 Noise limited Amplifiers . . . 43
3.2.2 Chopper Stabilized Amplifiers . . . 47
3.2.3 Bandwidth Limited Amplifiers . . . 50
3.2.4 Circuit Implementation . . . 53
3.3 Analogue Signal Conversion . . . 57
3.3.1 Capacitive array miniaturization . . . 58
3.3.2 Model based topology selection . . . 60
3.3.3 Circuit Implementation . . . 67
3.4 System Level Abstraction . . . 74
3.5 Conclusion . . . 80

4 Neuron Processor Interface (NPI) 82

4.1 Processing at the interface . . . 83

4.1.1 Methods for Neural Signals . . . 88
4.2 Resource Constrained Classification . . . 93
4.2.1 Spike Detection & Filtering . . . 94
4.2.2 Recursive Variance Decomposition . . . 97
4.2.3 Template Matching using K-means . . . 99
4.2.4 Complexity Evaluation . . . 100
4.3 System Architecture . . . 103
4.3.1 Distributed System . . . 105
4.3.2 Processing Core . . . 108
4.3.3 Testing Platform . . . 116
4.4 Conclusion . . . 120

5 Time Domain Techniques 122

5.1 Principles for time domain processing . . . 123
5.1.1 Sub-threshold Ring Oscillators . . . 125
5.1.2 Time Domain Sensor interface . . . 131
5.1.3 Time Domain analogue filter . . . 139
5.2 Analogue Signal Classification . . . 145
5.2.1 Feature Selection . . . 147
5.2.2 Mixed Signal Implementation . . . 149
5.2.3 Validation . . . 153
5.3 Conclusion . . . 157

6 Conclusion **159**

6.1 Original Contributions 161

6.2 Future Work 162

A Publications **164**

B Nomenclature **166**

B.1 List of Symbols 166

B.2 List of Abbreviations 168

C Supplementary Results for the Resource Models **170**

C.1 Topology Selection 170

C.2 System Level Abstraction 175

C.3 Transistor level implementation 178

D Supplementary Results for Classification Methods **182**

D.1 Signal Characteristics 182

D.2 Implementation structure 185

D.2.1 Introducing noise and LFP content 185

D.2.2 Analogue detection operator 186

D.2.3 Digital detection operator 186

D.2.4 Ω_3 optimization 187

D.2.5 K-means with centroid splitting 188

D.2.6 Recursive variance decomposition 189

Bibliography 191

List of Tables

2.1 Process parameter scaling based on different modalities with respect to the technology feature factor α 17

2.2 Performance specifications for integrated neural instrumentation systems found in the literature. † integrated power management, ⋆ integrated silicon probe, △ commercialized system, ◇ integrated stimulator, ○ integrated chemical readout. . 28

3.1 Summary of the target specifications for the analogue instrumentation system. . 41

3.2 Theoretical figures for NEF for various amplifier topologies. ⋆ N is the number of stages sharing the structure. 44

3.3 Summary of performance specifications of the proposed instrumentation topology and other bio-signal chopper stabilized amplifiers found in the literature. 57

3.4 Summary of performance specifications for the ΔΣSAR data converter and other oversampling/noise shaping data converter structures found in the literature. . . 73

4.1 Estimation on memory and computational resource requirements for each algorithm. 101

4.2 Overview of instruction sub-components. 113

4.3 Summary of performance specifications for the NPI system and state-of-the-art specialized integrated processing architectures. ⋆ Reconfigurable topology. . . . 115

4.4 Section of Instructions and recorded outputs from μC structure with the associated machine code. 118

5.1 Performance specifications of the proposed time domain instrumentation topology and those found in the literature. 145

5.2 Overview of Detection & Classification performance in green for data sets from [1] for different methods. ⋆ Requires off-chip Supervision. † White noise is also added at -20dB of the signal power. 157

List of Figures

1.1 Recorded activation map of a mouse brain after the mouse was allowed to explore a novel object for 90 seconds . . . 3

1.2 Electronic assistance in homotopic root repair for restoring bladder function after bladder decentralization due to spinal cord or spinal root injury . . . 4

1.3 A parylene self-locking cuff electrode with flexible cable that is wrapped around nerves 1 mm in diameter. . . . 4

1.4 A 70 mg stimulation device that is 2 mm in diameter for closed-chest stimulation of cardiovascular tissue that aims to replace conventional pacemakers that have increased risk due to battery replacement. . . . 5

2.1 Instrumentation for probing neural electro-physiology intra cellularly (above) and extracellularly (below). . . . 11

2.2 Representation of measuring neural activity with extracellular recording due to ion displacement. . . . 12

2.3 Example of coupling sensitivities with respect to different instrumentation parameters illustrated as membrane oscillations (blue), postsynaptic potentials (grey), action potentials (red) . . . 13

2.4 Exponential growth in research driven technologies. . . . 14

2.5 Primary factors influencing development of neural interfaces. . . . 16

2.6 Typical transistor parameter degradation due to technology miniaturization. . . 19

2.7 Different classes of processing modalities with respect to topology where supervised processes are in white and adaptive processes are in grey. 21

2.8 Extracted correlation in local field potentials for different frequency bands in the human superior temporal gyrus. 22

2.9 Chip micrograph of high density active probes fabricated with a 0.6μm CMOS process. 23

2.10 Neuronal recording yield over time measured from multiple cortical implants that use high density micro-wire arrays of which trails M, N, C, K and O where movable arrays. 24

2.11 System power density with respect to sensing area for state of the art recording systems. 26

2.12 Neural recording system from the literature showing the fabricated and system level implementation. 29

2.13 Neural recording system from the literature showing the fabricated and system level implementation. 30

2.14 Neural recording system from the literature showing the fabricated and system level implementation. 31

2.15 Neural recording system from the literature showing the fabricated and system level implementation. 31

2.16 Neural recording system from the literature showing the fabricated and system level implementation. 32

3.1 Illustration of the spectral power density characteristic for a typical neural recording with the associated frequency bins. 40

3.2 Typical system blocks for analogue instrumentation systems. 41

3.3 Harrison amplifier topology. 42

3.4 Illustration of normalized flicker noise component dependency with respect to the inversion coefficient. 45

3.5 Proposed compact chopper stabilized neural amplifier topology. 48

3.6 Analytic model of a two stage Miller compensated amplifier. 51

3.7 Proposed structure high loop gain amplifier stage utilising complementary transconductive input and the high gain current conveyor topology. 52

3.8 Schematic showing circuit implementation of the proposed compact neural amplifier. 54

3.9 Physical implementation of amplifier using a 6-metal 0.18 μm CMOS process measuring $75 \times 82\,\mu m^2$ in size. 55

3.10 Post layout simulated results of the proposed instrumentation circuit. 56

3.11 Post layout simulated results using periodic steady state analysis to evaluate the closed loop gain of the instrumentation circuit. 56

3.12 Schematic of a conventional N bit SAR ADC with the split capacitor at position M. 58

3.13 Numerical solution to Equation 3.13 relating the capacitive DAC area requirement with the DAC resolution (N) and the position of the split capacitor before capacitor K. 60

3.14 Schematic of the proposed $\Sigma\Delta$ assisted SAR ADC topology for achieving a more compact configuration. 61

3.15 Summary of the FOM ($P_{sys}/2^N\, f_s$) for each topology with respect to different resolution requirements. 65

3.16 Figure of merit dependency of the proposed ΔΣSAR topology with respect to design parameters K1 & K2. 66

3.17 Estimation on the expected figure of merit for a target resolution and varying SAR precision. The red star and blue circle indicate the target and measured performance respectively. 67

3.18 Schematic configuration of the top level control for the ΔΣSAR data converter. 68

3.19 Schematic implementation of the ΔΣSAR structure. 69

3.20 Physical implementation of ADC using a 6-metal 0.18 μm CMOS process measuring $93 \times 147\,\mu m^2$ in size. 70

3.21 Control loop used to perform calibration with a slow test signal at the ADC input. 70

3.22 INL Plots illustrating the mismatch artefact reduction due to calibration. 71

3.23 Measured THD and SNR of the fabricated data converter. 72

3.24 Testing setup used for characterizing the ADC. 72

3.25 Measured Figure of Merit as a function of oversampling ratio. 74

3.26 Multistage amplifier configuration using the series G to adjust the allocation power and area. 75

3.27 Resource allocation for analogue power and area using the parameters $G_T = 500$, $\alpha = 0.3$, and $\beta = 0.05$. 76

3.28 Normalized resource improvements for α with respect the case when $\alpha = 1$ for each configuration. 77

3.29 Analogue resource relations with respect to different supply voltages. 78

4.1 Impact of technology on R_A analogue (green) and R_D digital (blue) processing resource requirements extrapolated from a 180 nm CMOS technology under constant field scaling. 85

4.2 Analytic number of digital operations available with respect to different technologies (red) with references to the normalized performance of image processors (blue). 87

4.3 Estimated resource requirements for different classes of algorithms use for processing neural recordings found in literature. 90

4.4 Extracted frequency characteristics of digital filter used in Algorithm 4.1. 96

4.5 False alarm rates normalized by true positives for data sets with different background activity. 97

4.6 Illustration of centroid evolution over several iterations. 100

4.7 RVD and template based classification for data sets with $-26\,dB$ background activity. 101

4.8 RVD and template based classification for data sets with $-20\,dB$ background activity. 102

4.9 RVD and Template based classification for data sets with $-16\,dB$ background activity. 102

4.10 Illustration of the proposed distributed μC array for homogeneous program execution at the sensor interface. 105

4.11 System architecture for NPI sensing platform with digital interfaces annotated. . 106

4.12 Physical implementation of NPI system using a 6-metal 0.18 μm CMOS process. 107

4.13 Organization of the distributed execution unit detailing components and the interconnect. 109

4.14 Physical implementation of execution unit using a 6-metal 0.18 μm CMOS process . . . 112

4.15 Power dissipation with respect to specific operations for the same operand A=113 & B=114 in randomized order. . . . 114

4.16 Block diagram of the instrumentation platform developed as framework for real-time applications. . . . 116

4.17 Digital waveform of the internal data bus BIT 1-8 as new instructions are being loaded into the device using the clocked Latch and Configure signals. . . . 118

4.18 Graphical user interface used for configuring the NPI system showing test data. 119

4.19 Test bed used for characterization with various components illustrated. . . . 120

5.1 Voltage supply relationship with respect to the bandwidth and linearity requirements with respect to different technologies . . . 124

5.2 Model impulse functions for $\phi(t)$ and $A(t)$. . . . 126

5.3 Schematic of current regulated ring oscillator with capacitively couple noise source. . . . 127

5.4 Simulation results outlining the dependency of parameter dynamics as a function of oscillator phase . . . 128

5.5 The compensation effect of M1 on the ISF for capacitively couple noise sources with reference to Figure 5.3 . . . 129

5.6 Time domain instrumentation topology for low noise voltage to time-domain conversion. . . . 131

5.7 Transistor level implementation of the phase domain integrator structure with phase detector feedback. . . . 134

5.8 Simulated transient behaviour of the differential oscillator and the generated digital output. 135

5.9 Transient noise simulation result of the 180nm CMOS time domain instrumentation topology with a $6\,mV$ peak to peak sine input at $1\,kHz$. 138

5.10 Closed loop time-domain analogue filtering structure 140

5.11 Simulated transient behaviour of the differential oscillator and the time encoded digital signals internal to the feedback loop. 140

5.12 Schematic sub-blocks for first-order time domain analogue filter 141

5.13 Bandpass time-domain analogue filter which is cascaded to realize a 4^{th} order TD-BPF. 141

5.14 Karnaugh map associated with the subtracting XOR type PWM signals Q & X from D 142

5.15 Transient noise simulation of proposed instrumentation amplifier with time-domain filter structure with a $6\,mV$ peak to peak sine input at $1\,kHz$ 144

5.16 Transient noise simulation result of the proposed instrumentation topology in 65nm CMOS with $2\,mV$ peak to peak sine input at $2\,kHz$ 144

5.17 Illustration of Nyquist rate feature extraction and using feature enhancement in order to operate at sub-Nyquist rates. 146

5.18 Illustration of feature dependency for windowed neural spikes for data sets from [, roman] with $-16/, dB$ background noise and $-20\,dB$ added white Gaussian noise. 148

5.19 System abstraction showing the configuration for pre-filtering detection and re-configurable integrators. 150

5.20 Illustration of closed loop control for finding point of maximum variance. 150

5.21 Illustration of closed loop control for tracking two-centroids with the Ω_1 feature. 151

5.22 Delay line configuration for evaluating the absolute difference between the asynchronous time domain signal D and the registered centroid position X. 152

5.23 comparing PCA and Ω feature distribution for the Difficult2 data set. Ground truth for spike classes annotated as cyan, maroon, yellow and blue for false positives. 153

5.24 Ω_2 and Ω_3 classification for data sets with $-26\,dB$ background activity. 154

5.25 Ω_2 and Ω_3 classification for data sets with $-20\,dB$ background activity. 154

5.26 Ω_2 and Ω_3 classification for data sets with $-16\,dB$ background activity. 155

5.27 False alarm rates normalized by true positives for the analogue detection. 156

C.1 Summary of the performance merit for each topology. 174

C.2 Summary of the performance merit for each topology with a $4\times$ reduction in unit capacitance. 174

C.3 Summary of the performance merit for each topology with a $2\times$ reduction in reference voltage and a $4\times$ reduction in unit Capacitance. 175

C.4 Normalized resource redistribution dependency in low gain settings $G_T = 200$. . 177

C.5 System area requirement for two-Stage/single-Pole (red) and three-stage/double-pole (green) amplifier/filter structures. 177

C.6 Detailed dependency with regard to system supply voltage and the precision of the two-stage/single-pole (red) and three-stage/double-Pole (green) amplifier/-filter structures. 178

C.7 Transistor sizing for the two stage operational amplifier with bandwidth boosting implemented in 0.18μm CMOS. Sizing units shown are in microns. 179

C.8 Transistor sizing for the fully differential $\Delta\Sigma SAR$ switched loop filter implemented in 0.18μm CMOS. Sizing units shown are in microns. 179

C.9 Transistor sizing for the differential low noise complementary instrumentation amplifier implemented in 0.18μm CMOS. Sizing units shown are in microns. . . 180

C.10 Transistor sizing for the fully differential time domain instrumentation amplifier implemented in 0.18μm CMOS. Sizing units shown are in microns. 180

C.11 Transistor sizing for the time domain instrumentation high pass filter structure implemented in 0.18μm CMOS. Sizing units shown are in microns. 181

C.12 Transistor sizing for the differential time domain filter implemented in 0.18μm CMOS. Sizing units shown are in microns. 181

D.1 Example of synthetic data derived from typical neural recordings. 183

D.2 Filtered components of spiking activity and detection operator output. 183

D.3 Comparison of digital and analogue pre-filtering methods. 184

D.4 Comparison of digital and analogue detection methods. 184

D.5 Colour coded spike-waveforms for 16-sample spike windows where the three classes of neurons are coloured (ref,blue,green) and the false positives labelled (black). 184

List of Algorithms

4.1 Spike Detection and Alignment 95

4.2 Recursive variance decomposition 99

4.3 Incremental K-Means classification 100

Chapter 1

Introduction

The ability to tap into the electrical activity of the central nervous system holds the promise of personalised health-care with unparalleled capabilities such as restoration of vision or paralysis [2]. The development of implantable electronics that enable these capabilities are a key component to any future of bio-electronic medicines termed as electroceuticals that exhibit superior spatial and temporal precision than the chemical counterpart. This follows a vision for electronics that can be distributed throughout the body to treat a plethora of neurological conditions and psychological disorders by attaching to nerve endings to decode and modulate neural signalling [3]. These devices provide a means to model neurological activity and find a basis for diagnosing of brain related diseases. However the human brain alone distributes the activity of 8×10^{10} neurons over a $1200\,cm^3$ volume with intricate structures where current recording systems can only observe a minute fraction of isolated activity with centimetre scale devices. Making these systems practical and useful requires more effective ways to distribute the recording of activity. This requires techniques for miniaturization towards minimally invasive implantable recording devices.

> "Todays best brain-computer interface systems are like two supercomputers trying to talk to each other using an old 300-baud modem"
>
> 2016 - Phillip Alvelda [4]

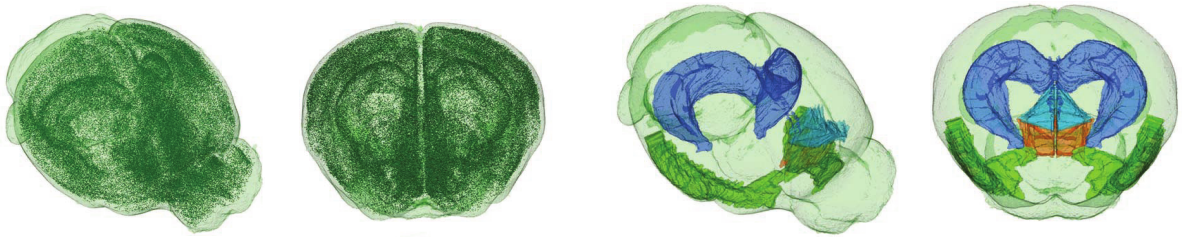
One exemplary effort directed at achieving better sensors for decoding neural activity is currently a $60 million project under a Neural Engineering System Design (NESD) program that is lead by the Defence Advanced Research Project Agency (DARPA). This is one of eight aspects under the brain initiative that is lead by research in the US which signifies the explicit importance of developing better interfaces to enable new technologies. From a data processing perspective the principle issue with current systems is the limited information rates that can be extracted from chronic neural recording devices.

1.1 Motivation

Brain machine interfaces have long been seen imagined as a solution to incapacitating neurological injuries like paralysis, spinal cord lesion, stroke, and brain trauma. However such a solution has eluded science for a number of decades. The functional interaction between electronics and neural tissue dates back to 1870 when Gustav Fritsch and Eduard Hitzig first questioned the muscular contractions due to electrical stimulation of the motor cortex [5]. This extensive development took many years in part because modulating neural signalling electronically in well controlled manner requires exquisite precision and correctly decoding the brain activity is a challenging processing problem that requires a intricate understanding in underlying brain function. The first success for interfacing electronics with neural circuitry is that of the cochlear implant which was approved by the American Food and Drug Administration (FDA) in 1985 [6]. In 2015 these implants help 13 thousand individuals over come their hearing disability in the UK alone [7].

Today we face an unprecedented and increasing number of patients with neurological disorders which we have yet to find rehabilitating treatment for. In fact 350 thousand people in the UK lost their self-reliance due to a disabling neurological condition [8]. Many of these cases are the result of unsuccessful rehabilitation after stoke, spinal cord/brain injury or our inability to treat progressive neurodegenerative conditions such as amyotrophic lateral sclerosis.

While these disorders still challenge our scientific understanding, the application of machine assisted intervention for neural repair and rehabilitation has demonstrated promising results with invasive implants [9] and non-invasive robotics [10]. Such experimental results encourage us to continue these efforts but it is also evident that this domain requires substantial improvements towards clinical solutions equivalent to that of deep-brain stimulators, pacemakers or the cochlear implants. In retrospect the guiding imperatives currently perused by leading research for better bioelectronic medicines in humans hope are aimed at meeting a number of neuroscience achievements by 2020 [3]. These efforts can be classified with five essential components;



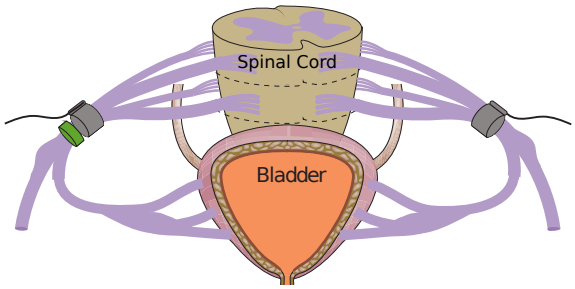
(a) Visualization of detected cells with activity-induced expression of fluorescent protein based tracers

(b) Anatomical segmentation of the active structures.

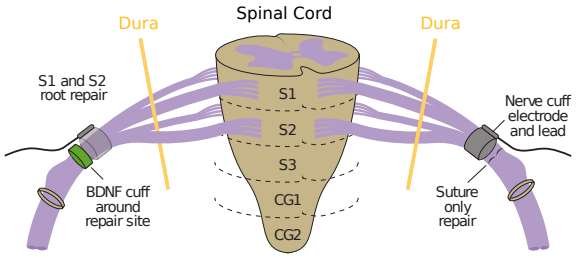
Figure 1.1: Recorded activation map of a mouse brain after the mouse was allowed to explore a novel object for 90 seconds from [11]

Structural & functional mapping - Revealing the signalling interconnect for organ-centric nerve wiring and functional signalling patterns provides a baseline for the diagnosis and understanding of nervous system. Similar to that in Figure 1.1 where the learning process of a rat brain is investigated with regard to functional patterns. This objective includes establishing correlations between organ function and neural activity. In order to catalogue these details we require tools for high-resolution imaging of fibre anatomy and taxonomy at micrometer resolution. The scale of exploring inter- and intra-species variation in neuroanatomy necessitates the collaboration through documenting a library or repository of tracers for visualization and standardized data collection. Moreover we need to find methods for precise identification and referencing nerve structures in a clinical environment.

Sensing organ functions - Finding better means to support the close-loop functionality of



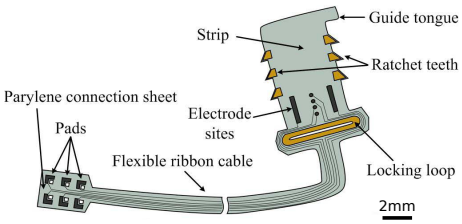
(a) Subsection of neural anatomy of the sacral contributions to the pelvic nerve used in restoring bladder function.



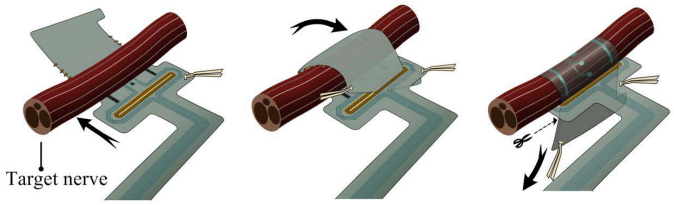
(b) Extradural electrodes situated to accommodate modulating neural activity of the pelvic nerve and assess neurally evoked bladder emptying.

Figure 1.2: Electronic assistance in homotopic root repair for restoring bladder function after bladder decentralization due to spinal cord or spinal root injury from [12].

bioelectronic factors in the body and organ behaviour is essential. Due to natural development and biological adaptation we require detailed understanding of physiological dynamics and thereby inferring organ function from marker variation. Here we need microscopic sensors that monitor and survey a variety of markers that allow us to characterize the changes in physiological markers in relation to organ and neural activity. Introducing low risk and highly reliable sensors is crucial for performing well defined and chronic data collection on patient variance. Here Figure 1.2 exemplifies such a sensor interfaced with the lower spinal cord to study the efficacy of neuro-reconstruction techniques to recover bladder control [12].



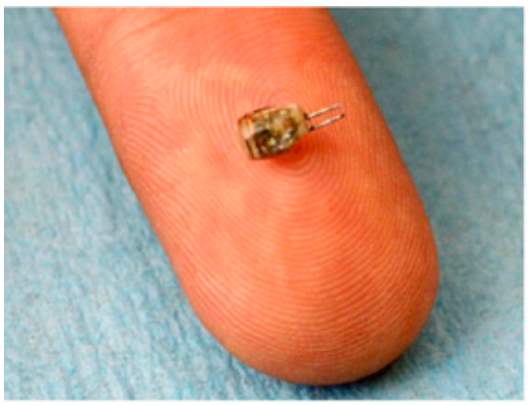
(a) Annotated components of the cuff electrode.



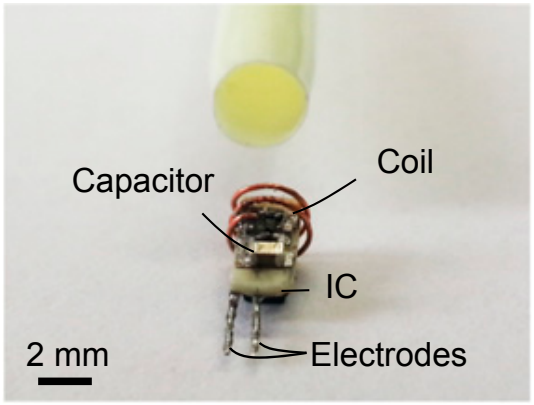
(b) Illustration of the implantation/self-locking procedure

Figure 1.3: A parylene self-locking cuff electrode with flexible cable that is wrapped around nerves 1 mm in diameter and provides an interface for active electronics from [13].

Electrodes for visceral nerves - Improving electrode-based interfaces for reliable recording and modulation is a proven enabler for many clinical applications. Miniaturization of cuff-electrodes is necessary for interfacing small nerves that are $<100\,\mu$m in diameter adjacent to organs that provide greater specificity for treatment. As shown in Figure 1.3 these are complex structures that require delicate application in a clinical setting. Furthermore high-density arrays



(a) Photograph of a wireless electrostimulator on a human finger.



(b) Miniaturized electrostimulator before epoxy encapsulation with system components annotated.

Figure 1.4: A 70 mg stimulation device that is 2 mm in diameter for close-chest stimulation of cardiovascular tissue that aims to replace conventional pacemakers from [14].

that exhibit shape-adaptable contacts could allow us to further maximize signal-to-noise and reliability. These platforms will need algorithms in order to track time evolving activity across multiple fascicles.

Signal imaging and actuation - Exploring other biophysical techniques that allow us to decode and actuate neural activity with less invasive means may result in more effective means for clinical applications with lower risk factors than cortical implants. This includes expanding on the current optogenetic methods to distribute excitatory or inhibitory activity in highly specific neuronal populations. Scaling electromagnetic and ultrasonic imaging to more practical form factors may also enable improved sensing modalities. Using nano-particles to enable remote interrogation by aforementioned sensing modalities also requires evaluating in-vivo reliability.

Visceral control modules - Enabling implantable sensors like that in Figure 1.4 requires integrating the interface with power management, wireless connectivity, and processing capabilities in highly miniaturized configurations. In particular these platforms need to provide modular support that enables recording, stimulation and blocking experiments for proof of concept. Developing the support for characterizing organ specific neural activity on-chip facilitates closed loop control without intervention. Moreover the capacity for signal processing and pattern recognition will improve bandwidth management for sensors with larger channels counts.

1.2 Research Objectives

Contemporary objectives for sensor design used in neuroscience applications can be approached from a number of perspectives. One could focus on wireless connectivity of body sensor networks, the chronic behaviour of miniaturized electrode interfaces, Analysis methods for distributed neural activity, or exploring organ specific functional indicators. All of which reflect how the research imperatives can be addressed in the domain of electronics. This work will specifically focus on the miniaturization and optimization of integrated electronic instrumentation. This work will form a basis for maximizing the capacity of implantable systems in a general sense irrespective of their functional purpose. In particular we raise the question: how does one enable a wireless implant to record from thousands of channels without being constrained by the bandwidth of telecommunication? Moreover how do we grantee that the integrated chip uses the absolute minimum amount of power and silicon area such that the system becomes viable for implantation? Can we find better ways to miniaturize these devices with existing technologies and what are the current limitations?

Here we attempt to approach the design of brain machine interfaces specifically with the objective of integrating thousands of channels on a single CMOS chip. This perspective implies not only that we need to find ways to miniaturize the electronic circuits used for sensing but also find processing architectures that perform highly efficient signal analysis irrespective of channel count. This reflects our consideration towards a fixed communication bandwidth that results from the limited power budget. The implication is that the neural recordings need to be compressed by representing only high level features of collective activity. As the number of channels increases our processing capacity must proportionally increase to accommodate the increased compression rate.

More explicitly, this work aims to develop a fully integrated instrumentation system for brain machine interfaces that is functionally far more capable and much more power efficient than conventional systems though analytic optimization methods and circuit level techniques.

1.3 Thesis Outline

The elementary components to these systems are discussed in Chapter 2. We review how neural activity is sensed electronically and the physiological components that affect these measurements. In addition we address how modern technologies are advancing the capabilities of neural recording systems to give insight to future prospects for BMIs. In particular we will discuss the associated impact CMOS integration has on sensor characteristics based on trends in on going efforts. Finally we overview brain machine interfaces that are currently presented in the literature by reviewing the current considerations in designing these sensors.

The principle discussion presented in Chapter 3 is based on the analysis of instrumentation circuit design and the optimization there of that enables us to achieve state of the art sensing efficiency. The methodology discussed informs us how resource efficiency is dependent on both topology and CMOS process parameters. In particular by introducing the underlying resource relations for amplifiers and data converters we can present a perspective on their system level implications. This discussion addresses the impact of circuit topologies to leverage the fundamental relationship between resource requirements and the performance specifications. In association we propose several topologies in each domain that can excel at maximizing performance with the minimum area and power requirements specifically for these low frequency signals. Finally these results are extrapolated to introduce the dependency on voltage scaling and higher order filter configurations. This reveals observable relations and limits with regard where our expectations lie given the direction of miniaturizing instrumentation structures.

Chapter 4 continues this notion of resource optimization by abstractly quantifying the extent we can perform processing on chip under the assumption that the design should dedicate as much resources on analogue instrumentation as on signal processing. This leads to the impression that depending on the construction of the algorithm there is an increasing amount of capacity for distributed digital processing as these systems employ more modern CMOS fabrication processes. Here we introduce an efficient implementation of two spike classification algorithms in association with a detection operator to quantify the resource requirements for methods commonly used in the literature. With this context we introduce a scalable microcontroller

topology the enables distributed processing across an arbitrary number of channels and executes reprogrammable algorithms with exceptional efficiency when large channel counts integrated on chip.

Chapter 5 uses the observations from the foregoing discussions to identify two components that allow integrated sensors to substantially improve performance. The first is using time domain topologies to realize a instrumentation topology based on digital logic gates that do not diminish performance due to transistor imperfections. By employing sub-threshold oscillator structures in amplification circuits the topology can achieves near ideal noise efficiency similar to conventional sensing but with improved characteristics. This structure exhibits significantly increased input impedance and reduced power dissipation than a more conventional realization would. The second component is that we propose performing adaptive classification of spike signals in the analogue domain in order to avoid digital processing data at the Nyquist frequency. This takes advantage of reduced feature bandwidth in neural recordings. Finally the different approaches to neural classification with minimal hardware requirements is evaluated in the analogue and digital domain. The proposed mixed signal method shows that we may opt to use an analogue integrator for every cluster extracted without needing analogue to digital conversion.

Chapter 6 summarizes this work by considering how the proposed architectures can be improved upon to approach better system integration. Reflecting on the current capabilities of the proposed system we conclude that much effort is still needed for realizing chips that are capable of directly decoding neural activity. While proposed platform already exhibits the necessary flexibility to accommodate a number of scientific inquiries there are two directions that are clearly exemplified with regard continuing research objectives. This outline can summarized in terms of the following technical details listed below.

- **Chapter 2**
  - Fundamental aspects for sensing of neural activity electronically and the trends in current BMI technologies.
  - Review of the fully integrated BMI systems presented in recent literature.
- **Chapter 3**
  - Chopped amplifier structure for the miniaturization of power efficient instrumentation and the design considerations there of.
  - Power efficient oversampling ADC topology for compact signal quantization with minimal area requirements.
  - System level resource models that addresses the configuration/optimization of amplifiers, data converters and digital processing.
- **Chapter 4**
  - Rationale for in-channel processing and the design of template-matching and PCA based methods for that context.
  - Scalable processing architecture based on software-defined instrumentation and the implementation there of.
  - Linux-based ASIC interfacing platform that enables real-time data visualization and high level instrumentation protocols.
- **Chapter 5**
  - Introduction for using sub-threshold ring oscillators as resilient time-domain memory that allows analogue processing with ultra-low supply voltages.
  - VCO based instrumentation topology for nano-metre CMOS capacitively coupled amplifiers and filter structures.
  - Develop a basis for performing adaptive neural spike detection & classification in the analogue domain.

Chapter 2

Background

In order to detail how we might be able improve the current approach of using electronic sensors to record neural activity. This chapter will review the primitive components to sensing brain signals and associate them with research developments that are promoted though the miniaturization in microelectronics. The discussion will give insight to trends like Moore's law that allow us to envisage what to expect from BMIs several years beyond current state-of-the-art and reveal challenges that still need to be overcome. This chapter is organized by introducing the basics of sensing of neural activity in Section 2.1 followed by a discussion on technology impact in Section 2.2. Finally we summarize what current systems look like today with respect to their capacity for decoding brain function in Section2.3.

2.1 Electrical Sensing of Neural Activity

The basis for interacting with neural activity electronically is derived from charged carrier exchange that couples electronic devices to the intracellular or extracellular potentials found in neural tissue and its fluids. Neurons found in the human brain signal information throughout the body by releasing various types of neurotransmitters at synaptic connections. Because the release of these chemical species is mediated by triggers from the localized membrane

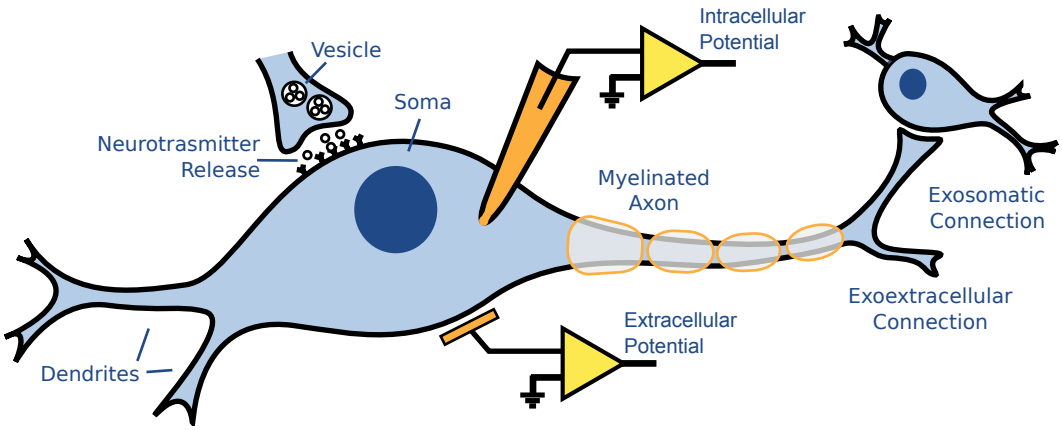
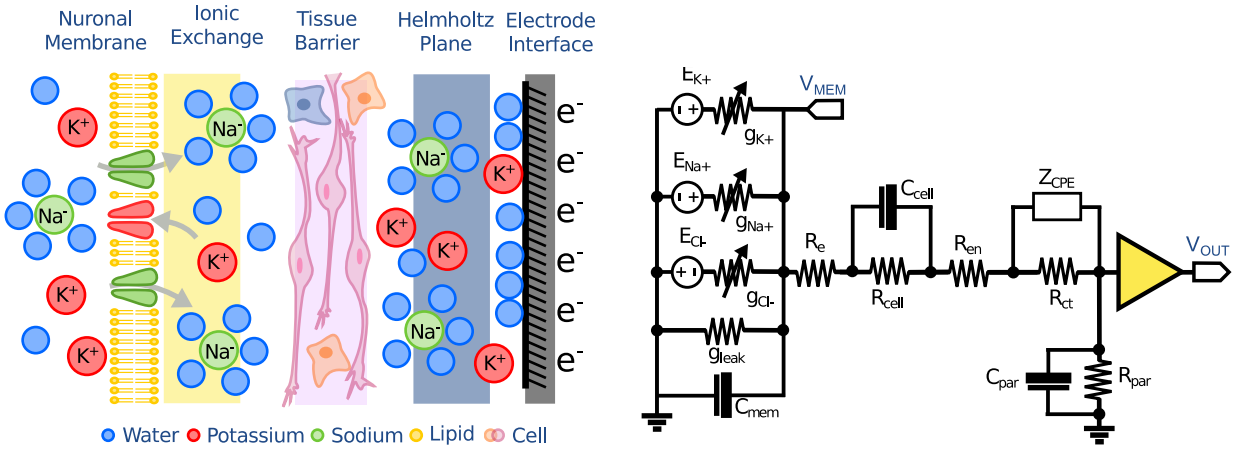


Figure 2.1: Instrumentation for probing neural electro-physiology intra cellularly (above) and extracellularly (below).

potential. A neuron's chemical activity can be inferred sensing voltage fluctuations at the cell membrane. Individual neurons will exhibit a variety of dynamics across the their cell membrane, the most prominent of which is the spiking response due to an excitatory threshold crossing in it membrane potential. These dynamics also include membrane oscillations and sub-threshold responses which can be detected but are substantially less energetic. However all of these phenomena derive from the interaction of gated ion-channels in these excitable membranes that enable the transmission of neural activity across the body of the neuron [15]. The specificity of this behaviour arise from specific neural characteristics that play a vital role in brain structure. For instance pyramidal neurons found in the hippocampus with slow-firing characteristics gain high place specificity with increased plasticity while fast-firing neurons have low selectivity but hypothetically promote neural network stability [16]. Plasticity refers to how easily a neuron changes it synaptic weights to increase sensitivity to trigger a response from the incident action potential of specific neuron. Evaluating factors relating to response latency, tuning and extracellular waveforms of individual neurons allows us to study their role in the network passively or in response to external stimulus.

The small signals generated by migration of ionic species is historically studied using the patch clamp method where an amplifier directly senses intracellular potential by means of a pipette sealed to the cell's membrane [17]. This is illustrated in Figure 2.1. In addition we could perform extra cellular recording where an electrode is placed in close proximity to the cell. A critical difference separating these two measurements is that extracellular recording is significantly



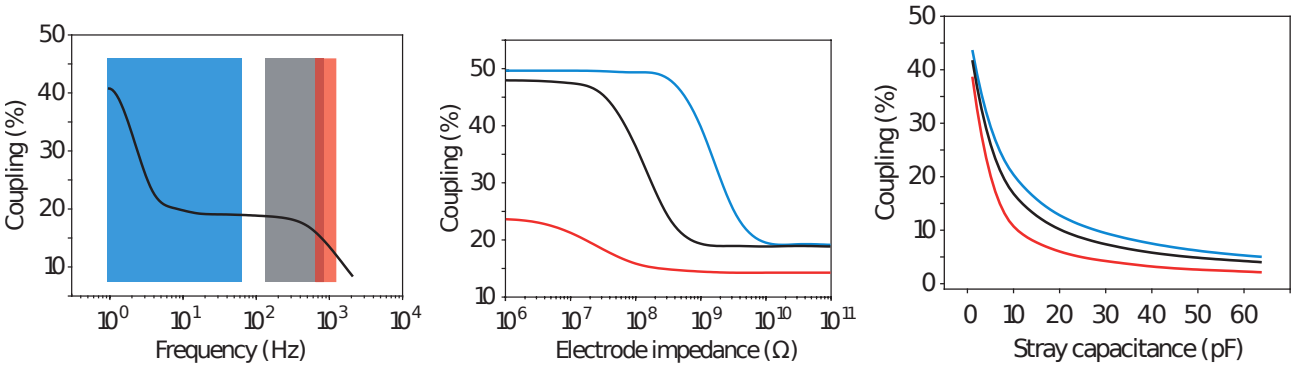
(a) Physiological representation of how ion migration across the membrane induces signals at the electrode interface.

(b) Electronic equivalent model for representing the transmission of cell membrane conduction dynamics.

Figure 2.2: Representation of measuring neural activity with extracellular recording due to ion displacement.

effected by the electrostatics of surrounding tissue. In particular intermediate neurons, genial cells, or astrocytes due to tissue scarring will attenuate the signal making it more difficult to detect activity. The defining advantage on the other hand is that embedding a shank with multiple electrodes will record the activity of numerous neurons without needing a clamping or sealing procedure. This allows close inspection of localized neural circuits and their interaction instead of the intricacies of single cell mechanisms. Either method however will allow the use of ion or protein sensitive electrode membranes for precise analysis of cellular fluid constituents.

Figure 2.2 illustrates the physiological and circuit equivalent understanding of how conduction dynamics at the membrane are coupled to an extracellular electrode. This coupling defined as V_{MEM}/V_{OUT} is related to what signals appear on the voltage output of the amplifier from which we want to infer how ionic species are being modulated at the cell membrane. That is we want to detect the time dependent or reactive variations in ionic membrane conduction g_{Na-} and g_{K+} which will be loaded by various components before appearing at the input of the amplifier. This is particularly important when considering electrode composition in order to couple to the input of the amplifier. One such consideration would be avoiding faradaic charge injection from the electrode that may result in the disassociation of molecules that form harmful by-products [18]. The physiological representation of spiking activity gives the idea that upon membrane stimulus the activation of sodium channels displaces charge inwards. This



(a) Coupling dependency for various signal frequencies.

(b) Impact of electrode impedance for the different signal bands.

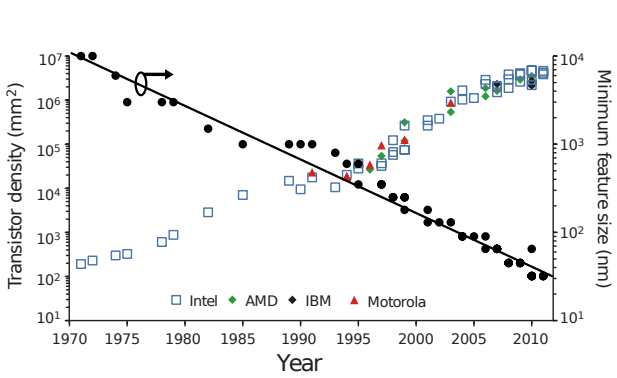
(c) Impact of stray sensor parasitics for the different signal bands.

Figure 2.3: Typical coupling sensitivities with respect to different instrumentation parameters illustrated as membrane oscillations (blue), postsynaptic potentials (grey), action potentials (red) from [20].

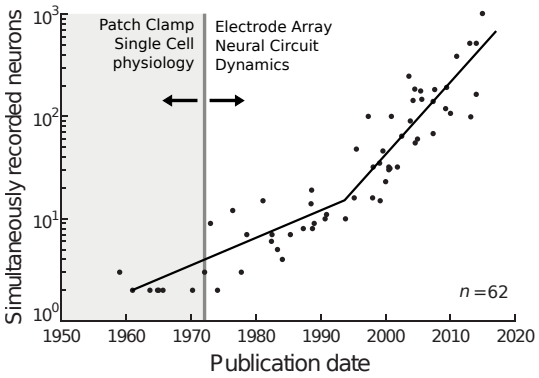
induces a localized recess in charge near the electrode before the potassium channels activate to initiate the repolarization of the cell. The effect of any intermediate tissue can be modelled by lossy coupling represented by the impedance R_e, R_{cell}, and C_{cell}. Similarly the electrode response can be modelled using Randeles circuit by extracting the equivalent parameters R_{ct} and Z_{CPE} [19].

In this respect we observe that the response observed at the amplifier is both dependent with respect to frequency and stay loading from biological or electronic elements in the signal path. Particularly the high frequency content or spiking activity is inhibited from coupling effectively through the tissue due to shunt resistance and stray capacitive loading. This also highlights that because the impedance of the electrode is proportional to its surface area when we reduce its size we must make sure to proportionally reduce the parasitic loading from the instrumentation in terms of C_{par} and R_{par}.

These relations are well understood and summarized in Figure 2.3 following a review on multi-electrode array technologies in [20]. The nature of this analysis similarly extends towards larger scale recordings like electrocorticography and electroencephalography where we increasingly subject the signals to a larger tissue barrier which limits propagation of electrostatic dipole effects. It also clarifies why low-frequencies tend to be more a prominent basis for analysis in the case of less invasive methods because they are less effected by the attenuation effects. In



(a) Trend in the number of integrated transistors per chip over the past 40 years. [22]



(b) Trend in the number of neuron recorded over the past 40 years updated from [23].

Figure 2.4: Exponential growth in research driven technologies.

contrast consider the advantage related to high spacial resolution at higher frequency bands as the measurement of single unit activity is isolated from distant sources by means of this attenuation. This will infer that recordings associated with analysing action potential activity are more invasive but will lead to better confidence in observing specific functional and temporally precise diagnosis of brain structures.

2.2 A Technology Perspective

The prolific impact developments in the microelectronics industry is compelling enough that its prediction, Moore's law, has come to represent an established achievement for integrated electronics [21]. The perpetual reduction in cost and sustained performance enhancement has become a strong driving force that many technologies exhibit similar growth characteristics in exponential progress simply by association with CMOS integrated circuits. While it could be considered a self-fulfilling prophesy, the expectation of progress has had a profound impact on science and engineering where research can be directed at objectives five to ten years ahead of contemporary technology to address emerging challenges. As a result leading research is not only associated with advancing our current understanding but also finding new ways to augment the methods used to validate these studies.

In 2015 Moore's Law celebrated its 50^{th} year and in Figure 2.4 we look back at its uncanny accuracy. While not evident, CMOS as a technology has had to address numerous hurdles like gate leakage, thermal limits of packaging, and the break down of mesoscopic carrier transport. The exponential growth in computational capacity per Watt in particular has played a mayor role in enabling mobile devices that are seeing new opportunities in point of care medical systems. Cost reduction is most certainly not the only factor that is driving the translation of CMOS to enhance a technology like patch clamp electrophysiology. The application of miniaturized electronics to a field like neuroscience introduces a considerable amount of added-value as 'Moore than Moore' scaling benefits beside digital processing power. These include increased sensor sensitivity, material diversification, sensor in package integration, or specific manipulation of quantum effects. The characteristic growth in the number of simultaneously recorded neurons in Figure 2.4 is a exemplary indicator how Moore's law is accelerating development in neuroscience. Upon the introduction of using microfabrication techniques to create electrode arrays on glass in 1972 [24] there has been a rapid growth in the number of simultaneously recorded neurons used in experiments. This growth accelerated again in the late 1990s when silicon based micro-fabrication techniques for micro-machined probes that established unprecedented capabilities. This trend can be interpreted to reflect that current systems are still unable to over come the complexity of how information in encoded within neural activity such that these application seek to acquire more recordings and measures. Nevertheless a substantial progress has been made over the past decade where only recently we have started to see successful real-time prosthetic control with a few degrees of freedom [25].

Figure 2.5 outlines the external drivers for innovation and more effective neural interface systems. As shown there is a engineering domain that primarily revolves around technology translation and pushing the capabilities of current integrated systems. The research driven factors from a bio-medical standpoint come from developing better models of the human body and finding new ways to interact with electronics or interpret signals. A large part of bio-medical research effort for interfaces is in order to explore basic science where we take the brain initiative [26] or the human brain project [27] as example. The nature of this research allows us to envision more effective ways to perform chronic treatment of neurological disorders and

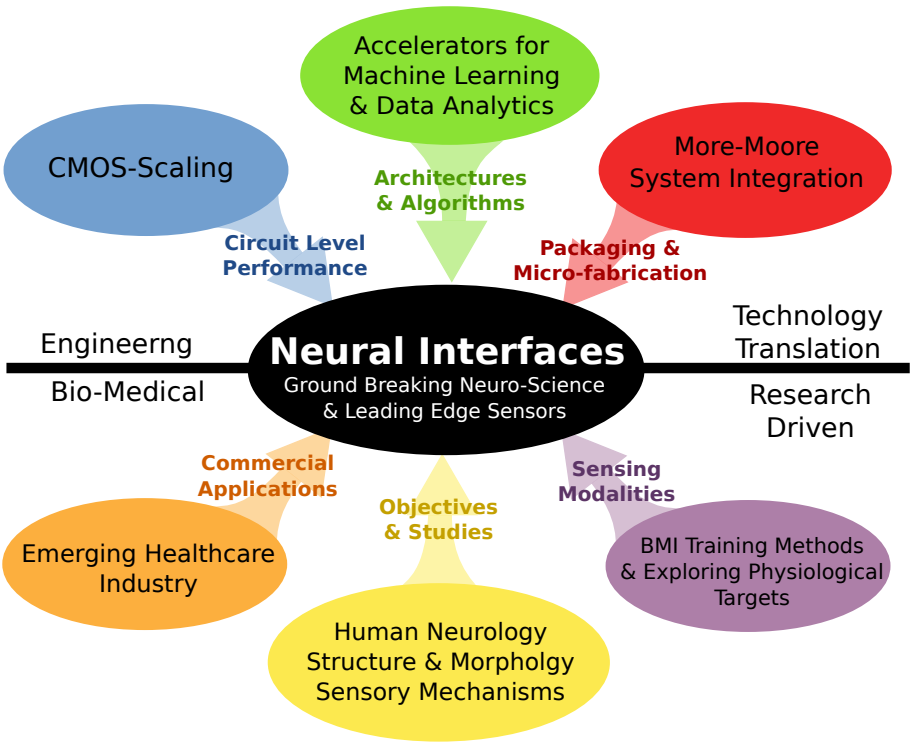


Figure 2.5: Primary factors influencing development of neural interfaces.

improving diagnostics by interfacing electronics with the human nervous system in the form of electroceuticals [28] or optogenetics [29]. While there remain many questions in regard to what the most effective means is to interact with the nervous system there are very clear short term goals associated with limb replacement and real time prosthetic control [30]. Although our focus here is with regard to the fundamental challenges associated with electronic sensors and instrumentation of neurons it is important to acknowledge their role in facilitating of high level objectives associated with these efforts. The impact technology has on sensors systems in particular can be derived from three factors; CMOS-Scaling, specialized processing architectures, and More-Moore System integration. The most obvious is that of the CMOS-Scaling which primarily correlates digital performance with the decrease in technology feature size. Because power dissipation is directly related to the switching losses used for clocking operations in digital systems. Reducing the physical size of devices implies a reduction in transistor capacitance C_{gate}. This well known relation for power dissipation; $Power = N_{tran}\, C_{gate}\, V_{DD}^2\, f_c/2$ tells us for given supply voltage V_{DD} and clock frequency f_c, the complexity of our digital system will infer how many transistors N_{tran} are used or equivalently how much capacitance is switched. In fact we can accurately anticipate to what extent different CMOS parameters will change under

Table 2.1: Process parameter scaling based on different modalities with respect to the technology feature factor α from [31].

Description	**Parameter**	**Constant Field**	**Maximum Speed**	**Multi Core**	**Adiabatic Scaling**
Gate Length	L_{gate}	$1/\alpha$	$1/\alpha$	$1/\alpha$	1
Wire Geometry	W, L_{wire}	$1/\alpha$	$1/\alpha$	$1/\alpha$	α^2
Supply Voltage	V_{DD}	$1/\alpha$	1	1	1
Gate Capacitance	C_{gate}	$1/\alpha$	α	$1/\alpha$	1
Clock frequency	f_c	α	α	1	$1/\alpha$
Transistors per Core	$N_{tran}/core$	α^2	α	1	1
Core Density	N_{core}/A	1	1	$1/\alpha$	α
Power Density	$Power/A$	1	α	1	1
Digital Throughput	$f_c\, N_{tran} N_{core}$	α^3	α	α	α

different scaling laws as technology progresses to finer transistor sizes [31].

The nature of the scaling laws in Table 2.1 reflect the different limitations for processing capacity in terms of α using different scaling modalities. Constant field scaling has been the predominant drive for improved performance that extends from the 1970s. Until the late 1990s where the threshold in particular was preventing V_{DD} to be scaled proportionally due to off-leakage in the deep sub-micron domain. This implied a combination of constant field and maximum frequency scaling was employed to increase processing throughput. While average power densities of integrated chips increased it became more evident other solutions were needed. This led to the commercialization of multi-core scaling in the mid 2000s that physically distributed the performance across a multitude of integrated circuits. This was known to be limited Amdahl's law [32] where the bottleneck transforming software to become highly parallel in execution. Anticipating this inevitable bottle neck the current direction for performance enhancement is looking for other scaling modalities that exploit the availability very large number of specialized transistors structures. Adiabatic operation for charge recovery has in fact been studied as a power reduction technique for over a decade [33]. This is currently being proposed in association with the processor in memory solution to address some of the data intensive problems [34].

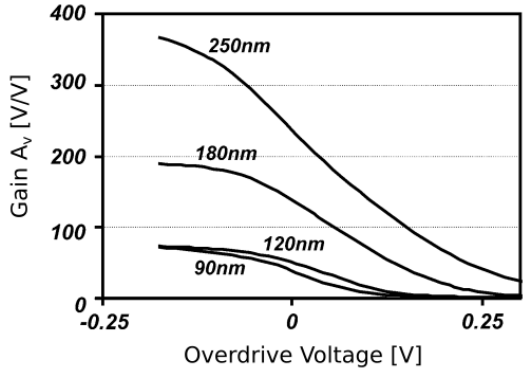
$$U_{Adiabetic} = \frac{C\, V_{DD}^2}{2} \cdot \frac{\tau_{RC}}{t_s} \tag{2.1}$$

It is interesting to note the fundamental concept behind this motivation as detailed in [31] in attempt to explore new mechanisms that reduce energy dissipation in switching systems. We can derive Equation 2.1 when we consider the power dissipation of a digital switching circuit $U_{Adiabetic}$ when instead of a fixed supply voltage we have a fixed supply current. We observe that if the charging phase t_s is much larger than the circuit's time constant τ_{RC} the dissipated energy is reduced and charge is recycled with other logic blocks. Moreover if this is configured with a resonant inductor L for energy recovery we fundamentally expect significant improvements over conventional CMOS in power dissipation based on fundamental principles in nano-meter CMOS [35]. This has conceptually already been applied to neural stimulation circuits [36] as wireless sensors are typically powered by a resonant power source this technique's prevalence is expected to be more frequent.

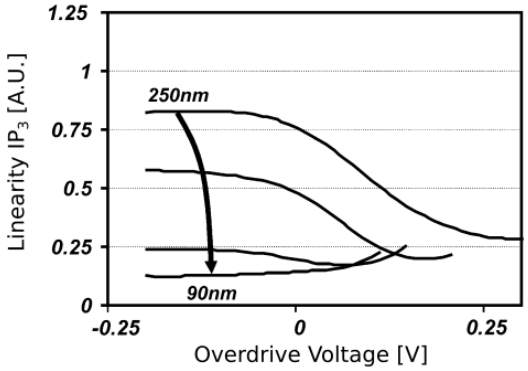
While the fundamentals of devices and processing architectures continue to evolve, where do we find analogue sensors in this changing context? Because the disparate nature of analogue sensing is continuous in operation the improvement in switching characteristics of the transistors will not be reflected. Fortunately the concept of semiconductor based instrumentation is well abstracted to the extent that we can predict the performance of analogue sensors with equivalent accuracy as that of digital circuits. In particular it turns out that many process parameters do not significantly affect the power dissipation of analogue sensors which we aim to aggressively minimize [37].

$$Power \propto kT \cdot SNR \cdot BW \cdot \frac{1}{V_{DD}} \frac{\Gamma}{gm/I_D} \tag{2.2}$$

Equation 2.2 embodies the main dependencies with respect to power dissipation for sensing signals electronically with a certain bandwidth BW and uses constants kT and q to represent the Boltzmann energy and the electron charge respectively. Here gm/I_D and Γ represent transistor efficiency in terms of its transconductance per ampere and noise excess factors respectively. With respect to the different technology scaling we can see two elements that will change as process parameters change. That is the voltage supply which scales somewhat inversely to that of digital dependency and the transconductance efficiency. This efficiency factor



(a) Illustration of transistor gain dependency on technology and gate overdrive voltage.



(b) Illustration of transistor linearity on technology and gate overdrive voltage.

Figure 2.6: Typical transistor quality degradation due to technology miniaturization from [37].

generally is equivalent the the transistor sub-threshold slope in the low-power domain which ideally approaches a theoretical value of $ln(10)\,kT/q$ independent of technology. This factor is as important to analogue as it is to digital operation so we expect a constancy in good transconductance efficiency for most CMOS technologies. The supply scaling however directly deters all-analogue sensor systems from migrating to more advanced processes because it may imply increased power dissipation if we naively attempt to use the same topology.

In addition to an increased power budget it is typical to see trends similar to those illustrated in Figure 2.6. This shows that using more advanced technologies may result in the minimum size transistor to behave less like an ideal transconductance element than its previous generation. This manifests itself by reducing amplifier gain or distorting filter linearity that prevent linear amplification of small signals. Ideally we want predictable relations between the small gate voltage variation and drain current of a device but this is becoming more difficult to achieve due to the dependency on the drain potential and short channel scattering effects. In addition these scaling laws will also imply variability in impurity doping concentrations have a proportionality to the scaling factor α. In fact we can estimate its effect on threshold voltage variance σ_{Vth} using the process parameters t_{ox}, N_A, t_{ox}, and $L_{eff}W_{eff}$ which represent the transistor's gate oxide thickness, channel doping concentration, and effective channel area resprectively [38].

$$\sigma_{Vth} \approx 3.19 \times 10^{-8} \left(\frac{t_{ox}\, N_A^{0.4}}{\sqrt{L_{eff}W_{eff}}} \right) \propto \alpha^{0.4} - \alpha^{1.4} \tag{2.3}$$

Most terms in Equation 2.3 will correlate to some extent between similar analogue transistors due to the fact when that they are placed close to one another and will experience similar statistical distribution in their parameters. It is not possible to prevent the over all variance for degenerating as the sensitivity literally approaches discretization at quantum levels. This has lead to the prevalence of digital calibration and pre-distortion techniques to counteract these shortcomings and still take advantage of increased bandwidth of analogue circuits. In addition this trend also explains the recurrence of mixed signal systems in modern sensor systems where the availability of immediate signal processing can effortlessly be taken advantage of. Dealing with transistor imperfections for circuits is one of the biggest effort in current analogue circuit design theory [39].

This brings us to the importance algorithms and machine learning for sensor systems in particular. It is becoming more common that modern systems have an abundance of data and processing capacity at its disposal. This has generally lead to 'smart' sensors that can extract an adaptive set of indicators or refine signal structure at the source instead of using supervision to fine tune any changing characteristics for each sensor exhaustively. For BMIs in particular the input data has unknown mixtures of information that we can only generally assume as relevant by probing the right sections of brain structure in the area of implantation. Dealing with this decoding problem can be quite demanding in the context of recording from large ensembles of neurons. If we inevitably want to achieve a mobile solution for brain controlled prosthetics it would be imperative to develop specialized hardware and processing topologies that can achieve efficient computation on a implantable platform [40].

The classes of methods being applied to neural sensors are very rich where some focus on extracting various characteristics in recordings while others focus on reinforcement learning associated with subjects adapting to a prosthetic [45]. Figure 2.7 depicts the typical structure for methods used in neural signal decoding with references and annotated automation. We note that at least in the literature there is a general trend to wards full automated feature mapping and representation discovery because a priori understanding on signal encoding is not needed at the cost of being less resource efficient. This is a highly diversifying field because there are many challenges that remain to be addressed effectively such as the effect if long term behavioural

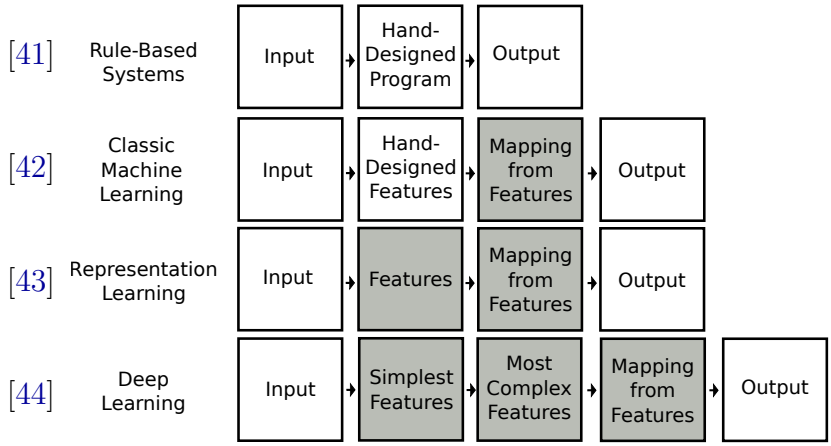


Figure 2.7: Different classes of processing modalities with respect to topology where supervised processes are in white and adaptive processes are in grey.

dynamics on cognitive state variables [46]. Even assuming spike rate based cortical recordings are used it is difficult to assert to which decoding method these systems will converge to in the long term. More typically novel methods approach improving performance by contributing to the framework of signal analysis. Typically this is not done in conjunction with the hardware or computational requirements which has led to some disparity with respect to methods that are feasible for integration and those that are not. Fortunately we can also find many promising hardware realizations that do allow the acceleration or an increase efficiency for these decoding methods. Although the more progressive deep learning methods still struggle with finding compact configurations [47] initial results appear promising with respect to power efficiency [48] due to digital scaling. However this is a common trend among integrated classifiers where power dissipation is on the order of several microwatt but due to complexity such a system requires several mm^2 of silicon area for operating on tens of channels [49] [50].

While these methods represent a corner stone for BMIs, it can be argued that carefully accounting for the spacial distribution of neural activity within brain tissue plays a much more significant role towards information extraction than the intricacy of the decoding method. In fact over the past decade, while we have been able to record from many more neurons and use more advanced processing, the rate of information extracted from neural recording has only marginally doubled to 2 bits per second [51]. In fact this paper argues that if adjustment of electrode positioning is allowed these rates can be fine tuned to achieve much better information rates experimentally. This is still a significant improvement over non-invasive recording

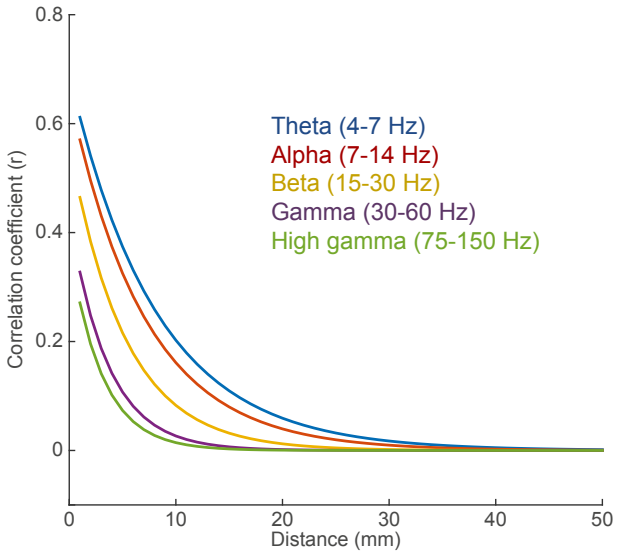
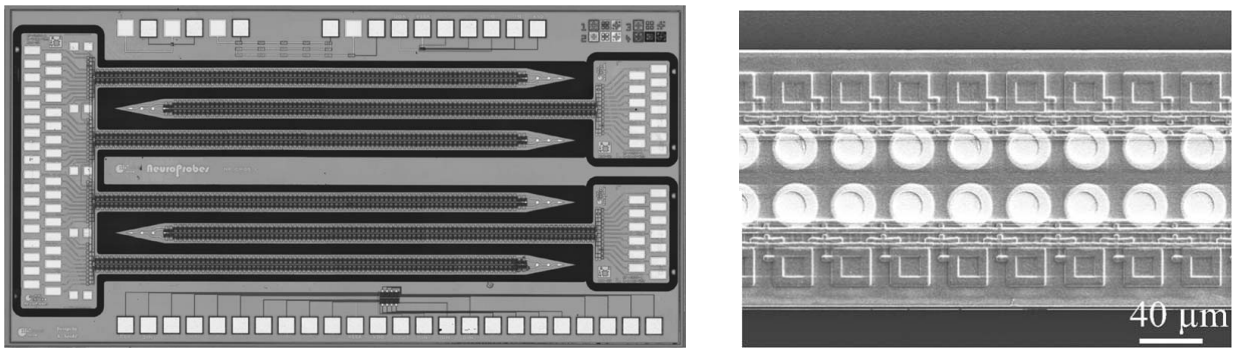


Figure 2.8: Extracted correlation in local field potentials for different frequency bands in the human superior temporal gyrus.

that is generally limited to less than 0.5 bits per second. A main consideration is illustrated in Figure 2.8 where low frequency neural activity are spatially correlated due to the propagation of local field activity [52]. Implantable devices need careful consideration towards electrode distribution to prevent high correlation in recorded data. These physiological parameters are highly varied in different sections of the brain and the motor cortex in particular exhibits high single unit correlations that impedes effective recording [53]. As such electrode distribution across the cortex of the brain is revealing to be critical aspect to implantation and approaching a more spatially distributed system design.

With that in mind we can interpret the current trend for monolithic integration of active electronics and neural electrodes. Fundamentally we know that in order to enhance the basic tasks seen in BMI experiments, a significantly larger number of neurons may be needed for recording. This is both from a reliability point of view as neurons will 'drop out' as tissue ages and inferring more degrees of freedom requires even better signal integrity. Moreover we are not definite in our current understanding of information encoding in relation to the anatomy of the brain. However scaling the current paradigm for neural recording platform will not allow implantation of micro-wires with external recording for very high channel counts. Both in terms of surgical feasibility and chronic stability of the system. The use of silicon probes with integrated CMOS devices have demonstrated superior viability for chronic implants with high density electrodes,



(a) Shank structure for wafer-level fabrication that enable numerous post-processing procedures such as platinum or iridium deposition.

(b) Close-up of the protruding iridium oxide electrodes and integrated electronics below.

Figure 2.9: Chip micrograph of high density active probes fabricated with a 0.6μm CMOS process from [56].

flexible structures, drug delivery [54], and hermetically sealed micro-packaging [55]. In extension these devices are significantly more scalable due to wireless capability and on-chip compression or data analysis which can be attributed to translating advancements in system integration. In particular the fine control over micro-fabrication structural or chemical composition is enabling chronic stability with less abrasive lesion when inserted into neural tissue.

Figure 2.9 shows the structure of current active probe technology where circuits can be integrated underneath the distributed array of contacts that are exposed to the tissue [56]. Such a shank is capable of monitoring activity from multiple cortical layers of the brain simultaneously corresponding to various structural and functional neurons that exhibit different dynamics at each layer. In fact it typical to see electrode structures use very dense configurations such that we can be selective with regard to which neurons are decoded though electrode proximity. Currently the head stage extruding from the shank is of considerable size due to the contacts to external circuitry and possibly analogue to digital converters. Ideally this can be minimized by redesigning the instrumentation and allowing for only a few contacts that connect to a inductive link. It is important to point out that freely moving cortical probes with a reduced footprint appears to be the most significant factor for preventing haemorrhaging and the formation of scar tissue that impedes chronic utility of such a device [18]. We can attribute significant improvements in chronic electrode behaviour with functional probes that leverage chemical agents in combination with micro-meter size incisions to reduce tissue damage [57].

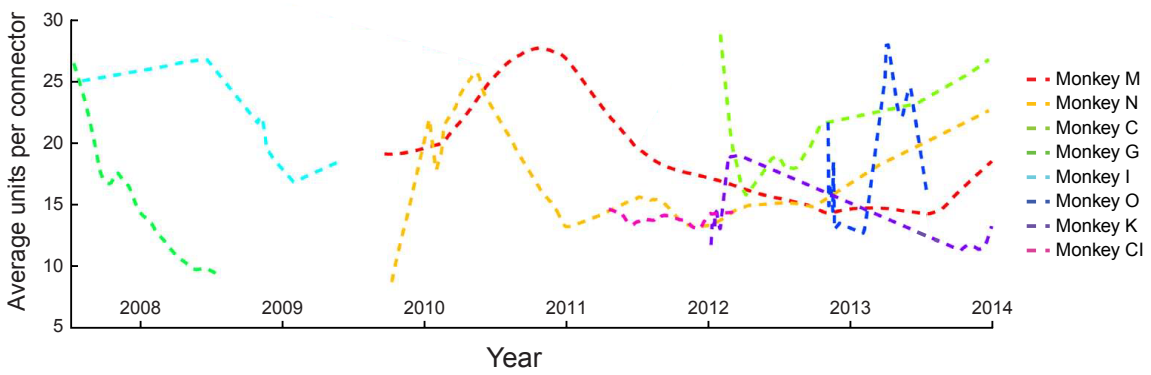


Figure 2.10: Neuronal recording yield over time measured from multiple cortical implants that use high density micro-wire arrays of which trails M, N, C, K and O where movable arrays from [58].

To a certain extent it is beneficial to have active recording electronics to be on the same order of scale as electrode structures. Such that electrodes can form in relatively arbitrary configuration from a fabrication standpoint and our data processing capacity would scale proportional to the number of recording sites in the system.

As illustrated in Figure 2.10 many of the earlier penetrating electrode structures can experience detrimental loss of recording capacity in chronic settings. The integrity of electrodes over the course of several years is scrutinized by a number of practicalities during and after implantation. Researchers have found effective methods to recover most of the recording activity by adjusting the electrode shank by a few micrometers or introducing neural growth factors after the primary inflammatory response from invasive surgery has passed [58]. In fact this remains a major challenge for integrated modules because once these are implanted we lose substantial access to manipulate these devices. Many of these techniques must be mediated by accompanied micro-mechanical structures that coordinate with the electronic system. This aspect represents the leading edge of biomedical device integration where chronic solutions still struggle with attempting to achieve a multitude of capabilities beside simply recording activity.

While it is easy to identify the demand for miniaturized packaging with electronic, mechanical and chemical sub-systems that are hermetically sealed for long term stability. The fabrication methods play a major role in enabling wafer-level or scalable fabrication techniques that are a crucial requirement for cost-effective commercialization. Silicon MEMS allow well established techniques to be applied to fabricating medical devices [59] but also exhibit drawbacks in

long term bio-compatibility associated material composition and structural stiffness. Emerging techniques in 3D-printing [60] or distributed mesh structures [61] may address these challenges but have yet to be evaluated with respect to long-term stability to the same extent that we understand the failure modes in silicon MEMS structures.

2.3 Integrated Recording Systems

Here we shall discuss the novelties presented recent publications to review some of the typical system level design considerations. These works present state-of-the-art specifically in the context of fully integrated systems. This implies that the design considers a multitude of functionalities simultaneously to enable isolated operation for implantation.

2.3.1 Power Density Target for High Channel Count Recording

The trend towards increased recording capacity emphasises the importance for careful sensor design that maintains adequate thermal dissipation. Particularly when current systems extensively perform processing on chip. It remains to be the case that electrode configurations like the Utah array [62] are distributed over a volume of cortex while sensing take place on a two dimensional array of instrumentation circuits. Clearly fabrication technologies could provide 3D integration of sensor circuits. However this may actually exhibit worse thermal capacity than 2D integration due to reduced surface area per recording unit. The key requirement is then that integrated devices should realize very high 2D density in order to probe the different layers & columns of cortical tissue.

In order to reveal the challenge of managing thermal capacity experienced by state of the art recording systems we have surveyed a number of sensor systems published over the past 5 years. This is summarized in Figure 2.11 which illustrates how well high density recording is achieved

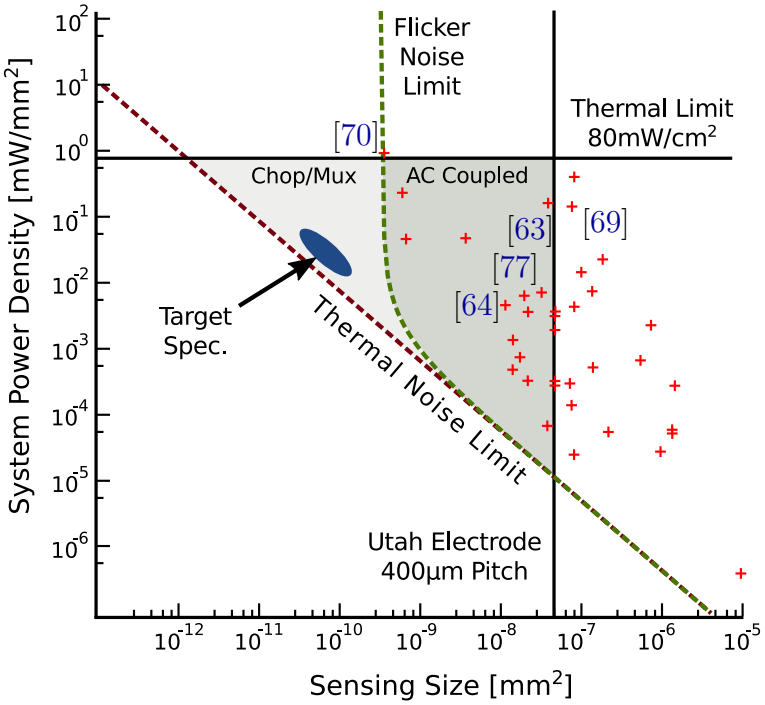


Figure 2.11: System power density with respect to sensing area for state of the art recording systems.

with respect to the thermal budget of 80mW/cm^2 [66]. In particular the system power density is considered with respect to the sensor density. The power density is evaluated in terms of the total system power dissipation and the fabricated silicon die size which typically dissipates its heat into the surrounding tissue. On the other hand the sensor size is taken with respect to the size of integrated electrode pitch or the size of the corresponding instrumentation circuits in order to emulate the achievable electrode pitch.

Probably the most interesting aspect of this depiction is the constraint imposed by the thermal noise limit. This results from the minimum current dissipation from each sensing circuit in order to achieve $5\boldsymbol{\mu}\mathrm{V}_{\mathrm{rms}}$ input referred noise given a 5 kHz bandwidth. A strict amount of power needs to be dissipated irrespective of the instrumentation size resulting in a power density inversely proportional to sensor area. The closer a system is to this limit the better the NEF it will exhibit. Unless the supply voltage, noise requirement, or bandwidth is readjusted the system power density cannot fall below this limit. As a consequence, managing the thermal budget is an imperative design consideration for future systems that target extreme miniaturization.

This is usually required in order to minimize device impact on biological tissues. For such a scenario, reducing the acquisition bandwidth may become a predominant reason for using LFP recordings over spiking activity. This could be the only way to allow several thousand channels to be integrated onto a sub-millimetre size CMOS device without thermal concerns.

The flicker noise sources for active readout transistors also plays an important role that can impair such achievements. If we assume transistors are equal in size to that of the sensing area then it should be expected that below a certain size the flicker components inhibit signal detection. Conventional readout topologies can not achieve very high densities due to flicker noise if chopping or electrode multiplexing techniques are not utilized. However such structures guarantee DC blocking behaviour which is desirable for safety regulations. It is well known that chopper instrumentation can achieve substantially smaller configurations but degrades instrumentation input impedance [67]. The system in [68] shows how the use of chopping can achieve 50$\boldsymbol{\mu}$m electrode pitch for exceptional spacial resolution with 768 active recording channels. This particular configuration performs quantization at the sensor interface using a ramp ADC and a open loop amplifying structure to minimize the number of analogue components in the system. A similar structure is employed in [69] with the addition of digital feedback in order to introduce high-pass filtering behaviour. This key modification allows robust in-vivo recording that is not impaired by electrode drift or large fluctuation in local field potentials that are typically not present en-vitro.

Another means to accommodate higher sensing densities is achieved by actively multiplexing the electrode array and only selecting the most informative electrode locations for the decoding task. This allows the work in [70] to interface over 26 k electrodes while simultaneously recording from 1024 units for the en-vitro study of cell cultures. By embedding SRAM memory cells within the electrode switch matrix this system presents a highly versatile MEA platform that can perform extensive signal conditioning. The main challenge is that the system complexity results in a 76mm^2 die size which is difficult to translate towards an implantable solution. For this reason the work in [65] focuses explicitly using a silicon probe as substrate to allow 52 simultaneous recording channels from 455 electrodes. This approach seems to be one of the most promising for future BMI systems as the electrodes can be fabricated together with the electronics using

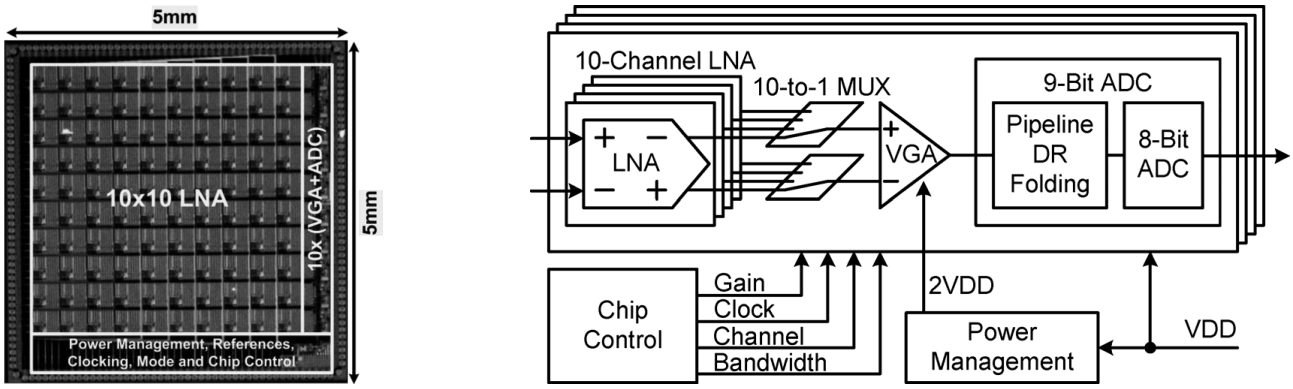
Table 2.2: Performance specifications for integrated neural instrumentation systems found in the literature. † integrated power management, ⋆ integrated silicon probe, △ commercialized system, ◇ integrated stimulator, ○ integrated chemical readout.

Reference		Intan [74]	Yoon [75]	Je [76]	Gielen [65]	Rabaey [64]	Chan [63]
Year		2012	2012	2013	2014	2015	2016
Technology	[nm]	-	250	180	180	65	180
Supply	[V]	3.3	0.9	0.45	1.8	1	1.8
Power	[μW]	830	3.96	0.96	27.8	3.6	9.1
Channels		64	16	100	455	64	200
Area	[mm^2]	0.47	1.56	0.25	0.19	0.075	0.067
Bandwidth	[Hz]	10k	11k	10k	6k	8k	10k
Highpass	[Hz]	0.1-1k	0.1-1.2k	0.25	1-500	10-1k	0.1
Noise	[μV_{rms}]	2.4	4.8	3.8	3.2	7.5	4.07
Note		△	⋆	†	⋆	†◇	○

well established microfabrication [71].

In addition to these high density recording systems a number of promising circuit techniques have been proposed that may allow considerable improvements in performance. For instance the application of bulk switching in [72] is capable of reducing flicker noise without sacrificing the DC blocking characteristics. On the other hand [73] suggests removing the capacitive coupling structure altogether by extensively utilizing digital feedback which could achieve very high recording densities if electrode multiplexing is also incorporated. Overall this trend towards higher recording densities is expected to remain a sustained effort where both fabrication and circuit techniques will play a key role for miniaturization.

Here we highlight the digital electro physiology interface chips provided by Intan technologies as one of the few commercialized integrated instrumentation systems [74]. These chips are the base line for numerous other advanced neural sensors. The mixed signal architecture it self can still be identified in newer systems. The on-chip configuration is relatively straightforward where there are 16-64 channels with differential inputs that are amplified and filtered by 3^{rd} order low-pass filter. In addition to a combination of 1^{st} order analogue and 1^{st} order digital high-pass filter are used to reject low-frequency content. The digital back-end directly clocks the data converter as slave to sample the various channels. The key component to these systems is



(a) Fabricated device in 180 nm CMOS.

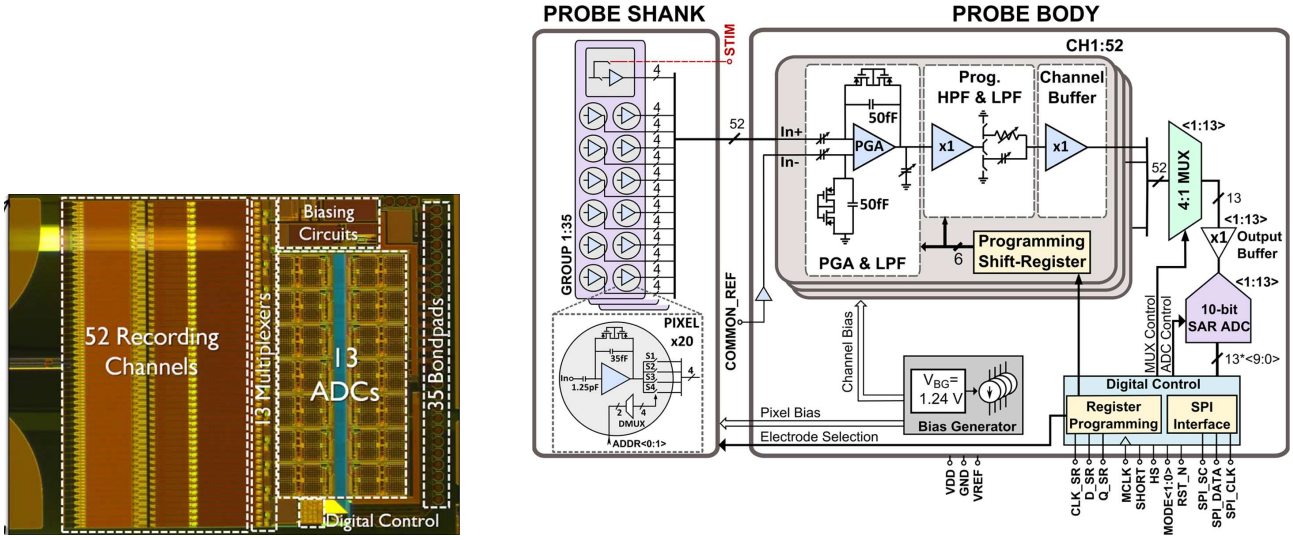
(b) Integrated system architecture.

Figure 2.12: Proposed neural recording system from [76] showing the fabricated and system level implementation.

the robust simplicity and guided application that has accelerated the research for many groups.

The multi channel device in [75] is some of the earlier work that focuses explicitly on the miniaturization of analogue recording. This work was later improved upon in [76] where a dual sample and hold structure was proposed to mitigate the class-A power-bandwidth trade off associated with the buffer driving the ADC by using two separate capacitor arrays. This system can be seen in Figure 2.12. More importantly this system introduced on-chip power management to that would allow a telemetry module to power the implant. There is also an attempt here for using the dynamic range at different supply voltages more effectively to marginally reduce the system power budget. By using a charge pump to drive the supply on stages that have a large output swing the low voltage component have reduced power requirement.

The authors of [65] introduce the conceptual hurdles and advantages of using a fully integrated active probe that is fabricated by post processing standard CMOS wafers. This system can be seen in Figure 2.13. Here the active electrodes are demonstrated to deal with the cross-coupling of high density electrode configuration and enabling a significant amount of multiplexing for electrode selection. Because we see the use of analogue voltage transmission across the chip the discussion addresses a number of limitations due to the limited drive capability of the active electrode structure. We may be quick to suggest alternative structures that does not suffer from such complications but it is difficult to avoid open-loop behaviour will not allow good



(a) Fabricated device in 180 nm CMOS.

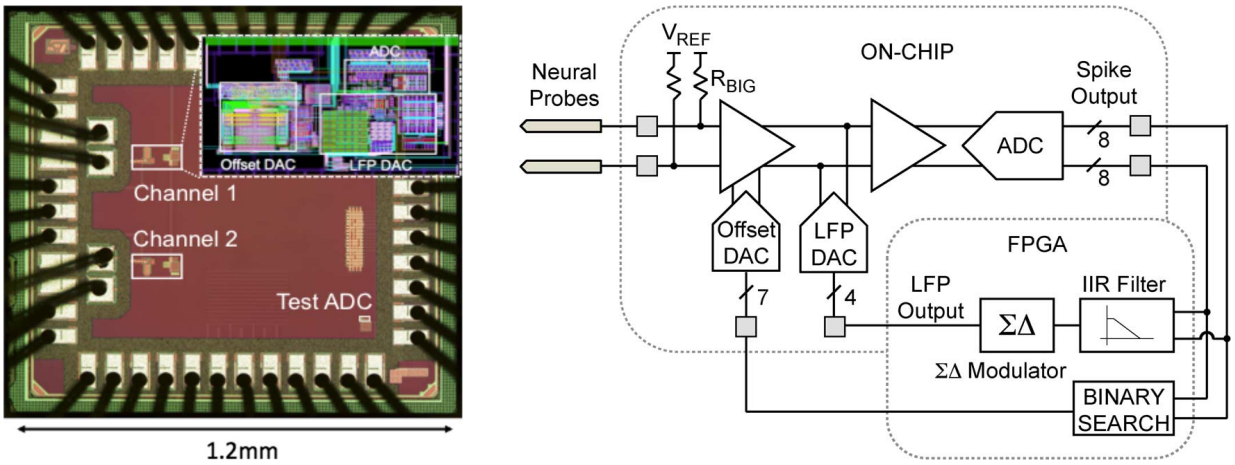
(b) Integrated system architecture.

Figure 2.13: Proposed neural recording system from [65] showing the fabricated and system level implementation.

linearity or matching drawbacks.

The work in [77] exemplifies an alternative to closed loop amplifier structures by performing direct quantization. This system can be seen in Figure 2.14. Similar to oversampling data converters this approach resolves a number of issues like the limited linearity of analogue blocks or variance in their characteristics. However these complications can now be seen in the feedback structure where the digital to analogue conversion requires fine linearity calibration in the feedback and most of the digital processing has a considerable dynamic range. This approach does well to leverage deep sub-micron CMOS.

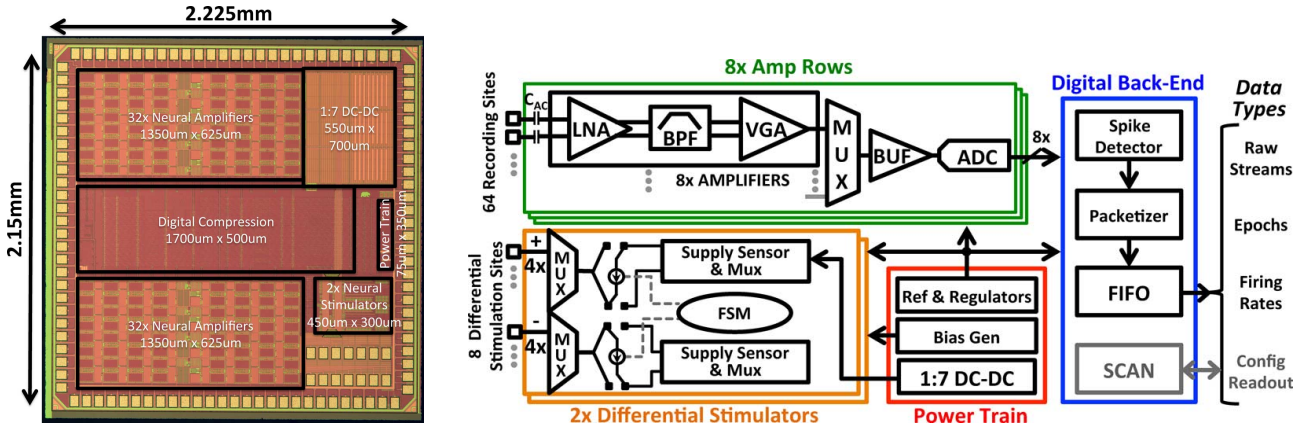
The system presented in [64] by the same authors as is exceptional because it demonstrated a fully integrated system on chip solution that can perform closed loop stimulation while processing neural activity. This system can be seen in Figure 2.15. Moreover the system integrates power regulation that adaptively scales the stimulator supply in order to maximize the efficiency of current stimulation. The embedded digital processing is capable of extracting spikes that are detected by a non-linear energy operator and evaluate long term firing rates as a form of data compression. Moreover the 65 nm CMOS technology allows this system to achieve a respectable power budget while performing extensive digital processing.



(a) Fabricated device in 65 nm CMOS.

(b) Integrated system architecture.

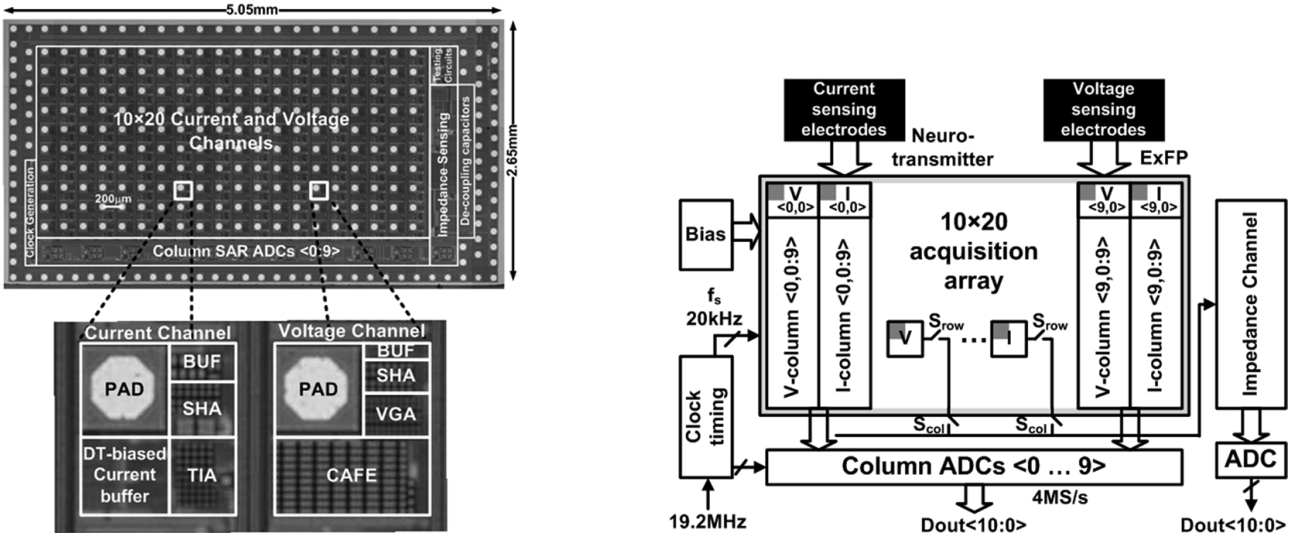
Figure 2.14: Proposed neural recording system from [77] showing the fabricated and system level implementation.



(a) Fabricated device in 65 nm CMOS.

(b) Integrated system architecture.

Figure 2.15: Proposed neural recording system from [64] showing the fabricated and system level implementation.



(a) Fabricated device in 180 nm CMOS.

(b) Integrated system architecture.

Figure 2.16: Proposed neural recording system from [63] showing the fabricated and system level implementation.

The acquisition platform in [63] merges the capability to perform electrical and chemical sensing of biological cultures. This system can be seen in Figure 2.16. The discussion on the different sensing modes motivates discreet-time biasing of the current feedback. This is followed by careful transistor sizing accordance to the detailed noise analysis that attempts to maximize the power efficiency. This work also reviews a number of other works with respect to their performance and draws our attention towards the difficulty of achieving good linearity and input impedance.

Generally we find that systems published in the recent literature are pressed to achieve good power efficiency in the analogue and digital domain because this component are well understood as a critical factor. The focus for innovation is typically the introduction integrating more axillary circuits that improve the viability for wireless implantation and wafer-level fabrication with electrodes. We do find that typically the compactness is not viewed as stringent because it is presumed to be largely technology dependent. Moreover integrated processing methods are strictly evaluated with respect to their algorithmic complexity and not the memory requirements which are equally scarce resources.

The fully integrated realization of neuromodulation ICs has been an outstanding endeavour for a number of research efforts that is rapidly becoming more successful. As demonstrated by

Medtronic's pioneering work [78] that is one of the few to realize neuromodulation devices with FDA approval. Similar closed loop systems extensively use LFP signals for seizure prevention have been particularly successful at realizing integrated solutions. For instance the system in [79] uses extensive segregation of different LFP frequency bands from 0.5 to 30 Hz as features which allows a support vector machine in the digital baseband to to perform classification on EEG activity. While this is not quite as invasive a solution as the Medtronic device this work illustrates that the processing architecture plays a key role in on-chip resource requirements. Generally the level of invasiveness will reflect in the system's latency to detect events in neural activity while exhibiting more challenging instrumentation requirements for EEG systems when compared to ECoG or intracortical recordings. In fact the newer implementation in [80] uses a non-linear basis SVM engine to improve its hardware efficiency resulting in an accuracy improvement from 84% to 95%.

To some extent the integrated form factor allows these SOCs to deliver substantially increased sensing & stimulation capabilities when compared to benchtop alternatives. In fact the overall experimental complexity is considerably less and environmental noise sources are less likely to perturb an integrated system. For instance the system in [81] uses 64 recording and stimulation channels for cortically implanted electrode arrays to deliver electrical stimulation therapy based on closed loop control. This type of active neuromodulation can improve clinical efficacy in freely moving rats. Earlier results [82] indicated that such an approach can reduce epileptic seizures by 90% in addition to improving chronic viability owning to a reduced implant size and wireless capability. This particular system uses phase synchrony in specific LFP frequency bands to trigger current mode stimulation with minimal feedback latency. A similar wireless system in [83] also uses LFP energies to deliver stimulation but proposes to take advantage of log-based signal encoding in neural activity to improve energy efficiency. This is just one of many examples where the physiology of neural activity can be used to allow more effective processing methods. The target application here is using deep brain stimulation to treat essential tremor and Parkinson's disease using a programmable PI controller.

An isolated number of works have used neural spiking activity to realize closed loop stimulation control of neural activity on SOC platforms. This is possibly due to the large stimulation arte-

facts that can disturb the instrumentation front end but also because most target applications are directed at external motor control. The system in [84] suggests adiabatic charge-recycling may potentially realize better efficiencies. However the recorded spike rates are not directly used to adjust the stimulation pattern without off-chip intervention. On that note the system in [85] utilizes the spiking rates from classified neurons to infer bladder volume and thereby provide a well informed condition for stimulation. This should indicate that peripheral nerve control is more adequate for this class of devices simply due to the lack of information in local field fluctuations. Instead chemical measures or the associated nerve activity is used for closed loop control.

2.3.2 Emerging Technologies

It is becoming apparent that increasing the capacity current BMIs requires us to find more efficient means to perform signal extraction as well as finding more effective interfacing strategies. For this reason we will highlight a few of the promising technologies that will hopefully make substantially more effective systems in the future.

Advanced CMOS Technologies

A number of recent BMI publications show a growing interest for using advanced CMOS technologies in order to accommodate more digital processing capabilities on chip. This is particularly relevant for closed loop neuromodulation that needs more sophisticated diagnostics to perform therapeutic feedback. To some extent the high channel count recording systems also necessitate extensive processing for various types of signal compression. As a result there are a number of opportunities associated with using nano metre CMOS processes. For example the 0.25 V neural processor in [86] is able to perform feature extraction on quantized recordings with exceptional power efficiency in part due to the 65 nm technology. Combining this with existing sensor interfaces that operate at very low supply voltages [87] [88] may result in an order of magnitude improvement in power dissipation when compared to previously proposed systems.

It is important to point out that there are other challenges that prevent conventional instrumentation structures to deliver precise sensing at these technology nodes. Particularly when attempting to achieve a compact configuration with good linearity at these reduced supply voltages [89]. It should not be surprising that the use of oscillators have been particularly successful to leverage the digital design style [90]. Our group recently demonstrated that oscillator based structures, while high digital in nature, can allow exceptional performance for filtering time domain signals [91]. In fact the concept of encoding signals in the time domain has been proposed in a number of recent works [92] [93] [94]. This is motivated by asynchronous processing capabilities for sparse neural activity that could drastically reduce power dissipation [95]. However this is still an ongoing effort where the current realizations show exceptional dynamic range but have not yet been able to demonstrate the same noise efficiency conventional methods. This is partly because such benefits can only be realized when both instrumentation and signal processing is performed using the time domain signal modality. Some further argument can be made that the reduced parasitics and smaller geometries from this trend make chopper circuits substantially more viable. This is because the input capacitance can be reduced if the closed loop gain is maintained relatively large. Such a reduction should translate towards boosting the sensor's input impedance to hundreds of mega ohm. Again this works in favour of VCO topologies because they inherently exhibit excessive open-loop gain.

Chemical Sensing

In addition to the electrical activity that can be recorded from neural populations, chemical markers such as dopamine concentrations can play a key role in utilizing neurochemistry to refine the diagnostic fidelity of BMIs. The neurochemostat is a recent innovation that allows a means to perform closed-loop regulation of endogenous neurotransmitters [96]. Because a number of neuropathologies relate explicitly to deficiencies in specific neurotransmitters this development can enable new therapeutic strategies that probe different pathways using rich neurochemistry. While chemical sensors face other challenges associated to the electrode composition & interface that can can degrade sensitivity in chronic settings [97]. However there are a number of promising means to prevent such degradation such as anti-fouling coatings [98].

Moreover a number of existing sensor platforms already allow simultaneous sensing of electrical and chemical activity with hundreds of different channels to study neurodegenerative diseases [99].

Optical Sensing

The use of optics in implantable sensors is possibly one of the newer themes in brain machine interfaces as a result of the enabling success in optogenetics. Such sensors may provide essential clinical tools to precisely guide neurosurgery as well as new high resolution imaging tools for brain structures [100]. That said there are also efforts to perform label-free imaging sensors that use the polarization of reflected light which does not require optogenetic transfection [101]. While there is a great amount of functional flexibility, implantable optics face a demanding challenge in relation to developing compact devices. This is exemplified by the implantable prototype in [102] which uses a coupled fibre to deliver optical stimulation from a battery powered system. This is primarily an outstanding challenge for recording activity from large volumes [103] that has the potential to be integrated along side more conventional recording circuits [68].

2.4 Summary

In hindsight to the trends in BMI innovation there is an definite expectation with regard to seeing more integrated probes with better processing capacity. Particularly because active probe integration solves many problems associated with electrode configuration and scalable data processing that is not limited by communication links. In some cases like closed loop stimulation it may even be necessary in order to avoid inconsistent signal latency which neural plasticity is highly sensitive to [104]. Moreover there are not many clear answers towards optimizing area and power of analogue instrumentation. Particularly when dealing with noise limited systems. It is more typical to see strategies proposed for solving problems that are

emergent from the implementation which is difficult to isolate from the fundamental aspects for signal amplification in a well defined manner. We also note that the number of neural recording system with integrated processing is very limited in terms of reconfigurability. In contrast to the work done with image processing ASICs that allow very efficient acceleration of various algorithms [105–108]. The reconciliation of software and algorithm developments is an important aspect which these specialized systems have yet to accommodate.

Chapter 3

Neural Recording Front End Design

This chapter focuses on the multitude of questions associated with the mixed signal design for multi channel integrated neural recording systems. As a result, a significant section will be directed at developing an abstract understanding of how design parameters influence the various design challenges. This discussion will clarify the key limitations for these systems and propose how they can be mitigated or efficiently designed for. In the scope of integrating a large number of recording channels together, clearly understanding how each resource trades for another is crucial for optimizing a complex system. Optimization methods found in the literature typically assume a certain configuration which limits to what extent improvements can be made [109]. However here we specifically identify abstractions that allow us to consider the impact of different topologies and filter structures simultaneously. This should enable a much boarder sense of optimization that will reflect in the improved performance characteristics demonstrated here.

We will focus on elaborately evaluating the dominant resource requirements with respect to noise, mismatch, quantization, and functional aspects for signal conditioning together that is mostly implementation independent. In addition we propose several circuit implementations based on this analysis that present highly efficient and compact instrumentation. The corresponding abstractions that we use attempt to realize clarity respect to underlying dependencies. This should allow better analytic models that make the limiting factors appear obvious and

reveal means to circumvent specific constraints with alternative techniques. For example we may be interested to know when it is worthwhile to put certain functions in the digital domain in terms of the CMOS technology parameters. Approaching the ideal instrumentation structure in such a scenario remains highly desirable for constrained applications. Thus conforming to the technology parameters could reveal that conventional methods do not deliver the most effective solution.

The chapter is organized as follows; Section 3.1 describes the general problem statement related to the analogue front end which is followed by the associated amplifier design considerations in Section 3.2. The method for improving the analogue to digital conversion is outlined in Section 3.3. These results are then collected in Section 3.4 to evaluate the impact of system level parameters as a function of resource requirements.

3.1 Architecture for Neural instrumentation

The analogue dimension of neural recording system can be broken down into two objectives for signal conditioning that will maximize the performance of the proceeding digital signal processing. The first is related to getting adequate signal quantization by amplifying the signals to full input range of the data converter without corrupting the signal of interest. The second objective is performing some kind of filtering that removes noisy or irrelevant components in the recording and only captures the relevant signals of interest.

As depicted in Figure 3.1, the input spectrum of a typical in-vivo electrode recording can be classified using a few frequency bands. The energy from extracellular spiking activity is primarily concentrated around $300\,Hz$ to $6\,kHz$ and is characteristically intermittent resulting in a distinct difference between the average and instantaneous spectral power [110]. This characteristic is also present in the LFP band to a lesser extent. From an electrical standpoint the design constraints are derived from the tolerated noise levels in each frequency band to maintain a proper signal to noise ratio. As a consequence it important to specify the signal to noise ratio

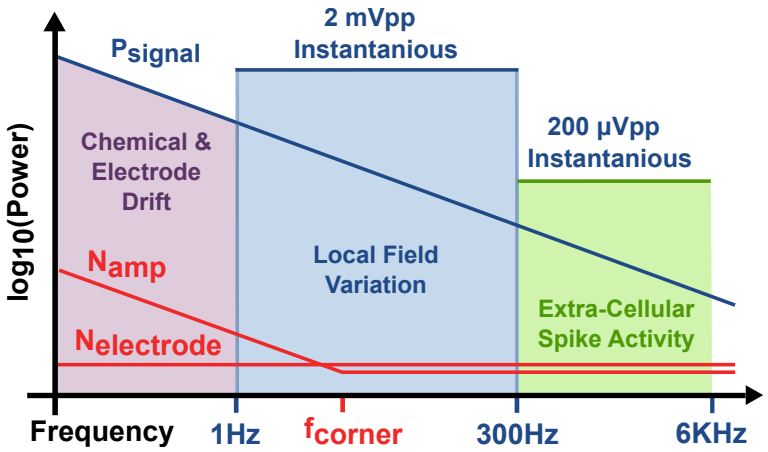


Figure 3.1: Illustration of the spectral power density characteristic for a typical neural recording with the associated frequency bins.

in terms of noise density opposed to integrated noise figures as digital processing accuracy is not limited by the later term. Here we should also note that the electrode spectral noise power $N^2_{electrode} = 4\,kT\,R_{en}\Delta f$ depends on the resistive component of the electrode impedance. This is typically matched by that of the amplifier noise characteristic N_{amp} so that no excess power is wasted and is expressed in terms of the electrode resistance R_{en}, Boltzmann energy kT, and the frequencies of interest Δf.

3.1.1 Instrumentation Requirements

This kind of electrical sensing can be broken down in the a number of system blocks each of which perform an essential operation to this process. These are shown in Figure 3.2 and consist of an amplifier, a filter, a sampler, and a quantizer. Occasionally one circuit can combine multiple of these operations together depending on the construction. Table 3.1 presents the overall performance requirements that should be demonstrated when these components are integrated together. These are also the specifications that we will target as the design is being considered in the following discussion. The reasoning behind these specific requirements are mainly related to conventional signal acquisition given the bandwidth and noise requirements. Moreover these seem to be sufficient for most decoding/characterization methods hence similar figures can be found in most BMI publications.

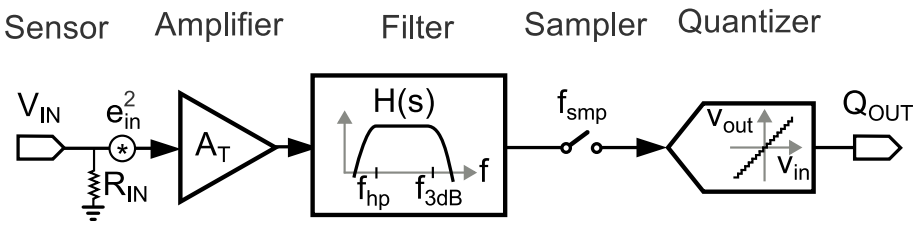


Figure 3.2: Typical system blocks for analogue instrumentation systems.

Table 3.1: Summary of the target specifications for the analogue instrumentation system.

Parameter	Symbol	Specification
Integrated Channels		64
Supply Voltage	V_{DD}	<1.8V
Power Dissipation	P_{SYS}	<5 μW
Diff. Signal		5 μ - 5 mVpp
Common Signal		50 mVpp
CMRR/PSRR		>80 dB
Input Referred Noise	e_{in}^2	<5 μVrms
Total Gain	A_T	>40 dB
THD at max input		>40 dB
3dB Bandwidth	f_{3dB}	6 kHz
High pass frequency	f_{hp}	<1 Hz
Sampling rate	f_{smp}	25 kS/s
Input Impedance	R_{IN}	>50 $M\Omega$ @ 1 kHz
ADC Resolution	$ENOB$	12 bits
Active Area		0.01 mm^2

3.2 Amplifier Principles for Miniaturization

The principle design considerations for neural instrumentation have been well established particularly with regard to the Harrison topology [111] that been widely adopted in many systems and shown in Figure 3.3. Objectively the optimization techniques have become both more specialized and specific for maximizing the average signal to noise ratio in the LFP or EAP bandwidth with the absolute minimum power budget. Interestingly due to the use of more advanced CMOS technologies there is a persistent trend towards sub-threshold operation. This is motivated by trading in the excess transistor bandwidth for improved current efficiency that measured in terms of the achieved transconductance per dissipated ampere of current. In fact this is purely a result of maximizing the individual transistor performance with respect to the speed efficiency product [112]. This is expressed in Eq 3.1 using f_T, U_T, v_sat, μ as the transition

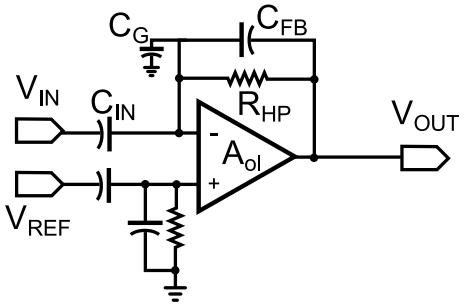


Figure 3.3: Harrison amplifier topology.

frequency, thermal voltage, velocity saturation voltage, and carrier mobility respectively.

$$\max_{IC}\left\{f_T\,\frac{gm}{I_{DS}}\right\} = \frac{v_{sat}^2}{4\pi\,\mu\eta\,U_T^2}\,\frac{1}{IC_{max}} \approx \frac{22}{IC_{max}}\left[\frac{THz}{V}\right] \quad Where \quad IC_{max} = \left(\frac{L_{sat}}{L_{tech}}\right)^2 \tag{3.1}$$

Here L_{sat} is a technology independent BSIM6 parameter that reflects the impact of ballistic carrier transport during velocity saturation and normalizes the minimum feature length L_{tech} for a specific technology as an effective length. The implication of Equation 3.1 is that the transistors for optimized low frequency instrumentation amplifiers are exclusively in the sub-threshold regime because f_T is always in excess with respect to the signals of interest. The subthreshold operation results in each transistor's transconductance being defined as $gm = \frac{I_{DS}}{\eta U_T}$ which only depends on drain current. Instead of noise optimization though the overdrive voltage, V_{ov}, the topology can only reduce noise by removing non-amplifying transistors or biasing them with reduced drain current when compared to the input transistor(s). This reflects the need for a different design methodology as the input referred contribution is dominated by how the total amplifier current distributed to all the transistors. At least in the small signal sense the key requirement is that the amplifying transistors dissipate all the current while biasing/non-amplifying transistors dissipate relatively very little.

In principle due to the under-determined nature of transistor level design the optimization methodology is initially constrained by one of the most important objective characteristics. This could be low noise, wide bandwidth, good linearity, etc. Hence this discussion will digress by distinguishing the design considerations for noise or bandwidth limited amplifiers as separate

cases. This should reveal some key relations with respect how power efficiency is achieved. For each case we evaluate the implications with respect to different resource requirements.

3.2.1 Noise limited Amplifiers

This discussion is guided by the leading challenge for instrumentation systems which is maximizing efficiency while maintaining good linearity. For this reason a noise efficiency factor (NEF) was first introduced in [113] and is expressed in Equation 3.2.

$$NEF^2 = e_{in}^2 \frac{I_{tot}}{U_T \, 4kT \, \omega_{3db}} \tag{3.2}$$

This figure represents a normalized efficiency or in other words it evaluates how much extra current is dissipated by a particular circuit when compared to an ideal bipolar junction transistor for the same noise performance. Here e_{in}^2, I_{tot} and ω_{3db} represent the input referred noise power, the total current dissipation and the -3dB bandwidth in radians respectively. NEF reflects how well a particular topology achieves efficient amplification for a particular noise floor and thus it inherently trades off with a multitude of other parameters. Here we shall use it as design parameter that reflects the chosen transistor level topology. With this in mind, we propose the following reformulation from Equation 3.2:

$$e_{in}^2 = \frac{kT}{C} \frac{NEF^2}{\eta \, A_{cl}} \frac{I_{in}}{I_{tot}} \quad \text{where} \quad C = \frac{gm}{\omega_{3db} \, A_{cl}} \quad \text{and} \quad gm = \frac{I_{in}}{\eta \, U_T} \tag{3.3}$$

This result leads to:

$$gm = \omega_{3db} \frac{kT}{e_{in}^2} \frac{\zeta}{\eta} \quad Equivalently \quad I_{in} = \omega_{3db} \frac{q \, U_T^2}{e_{in}^2} \zeta \quad \text{where} \quad \zeta = NEF^2 \frac{I_{in}}{I_{tot}} \tag{3.4}$$

Note that this relation is exclusive to noise limited characteristics and implies nothing with regard to the output load or linearity conditions. Moreover there is a fundamental requirement

Table 3.2: Theoretical figures for NEF for various amplifier topologies. ⋆ N is the number of stages sharing the structure.

Topology	NEF	Minimum Vdd	Reference
Single Transistor	η	V_{th}	-
Differential Pair	$\eta\sqrt{2}$	$V_{ds}+V_{th}$	[116]
Complementary Pair	η	$2\,V_{th}+2\,V_{ds}$	[117]
Common Reference⋆	$\eta\sqrt{\frac{1+N}{N}}$	$V_{ds}+V_{th}$	[118]
Common Bias⋆	$\eta\sqrt{\frac{2}{N}}$	$(1+N)\,V_{ds}+V_{th}$	[115]

for transconductance with respect to noise and an implementation related factor ζ. This factor represents the noise efficiency of the topology and the slope factor η that tells us about the transistor performance as a fundamental process parameter. Numerous techniques for improving NEF can be found in the literature. As a generalization these can be put into two categories. The first reducing the transconductance of non-signal amplifying transistors using degeneration such that their input referred noise is minimized [114]. The second approach is AC coupling the amplifier's input signal to biasing transistors such that the total transconductance is increased and the current efficiency is improved. Interestingly because this factor relates to current efficiency the NEF can be smaller than 1 or exceed the efficiency of a BJT using a stacked mixer structure that reuses the same biasing current for multiple amplifiers [115]. This hints at the fact that NEF should be normalized to the voltage supply but in some sense these structures trade off dynamic range for power efficiency. Theoretical NEF figures for some of the primitive low noise topologies are listed in Table 3.2 assuming biasing transistors have negligible contribution and taking V_{th} as the NMOS & PMOS threshold voltage.

These relations highlight the fact NEF primarily dependent on the chosen topology and less sensitive to the actual transistor design after optimization. Choosing a topology for the instrumentation amplifier with respect to its ideal NEF performance is significantly more effective than starting with a particular structure and introducing resistive degeneration on transistors that should not contribute noise.

Also notice that the expression for noise in Equation 3.3 only has one degree of freedom and

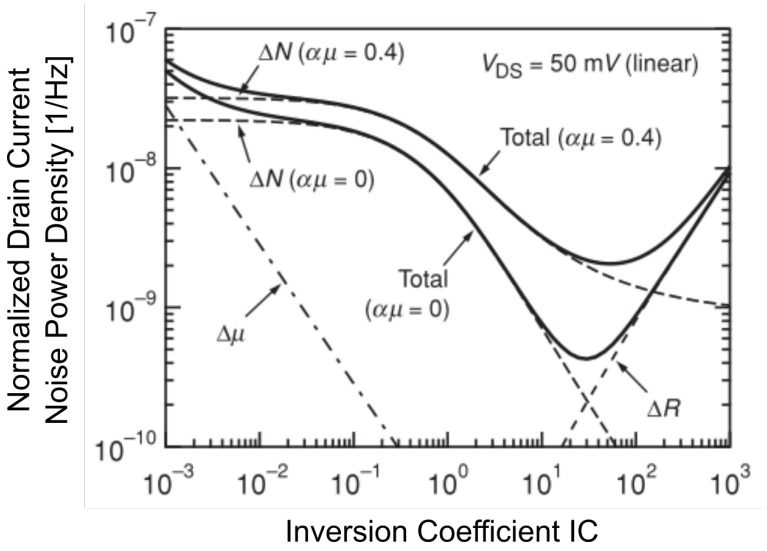


Figure 3.4: Illustration of normalized flicker noise component dependency with respect to the inversion coefficient from [119] Copyright 2006 by John Wiley Sons Reprinted with permission.

that is the ratio between the closed loop gain and capacitive load of the amplifier. This implies the 3dB bandwidth of the amplifier is fixed but its unity gain frequency is arbitrary. In fact by satisfying the relation for Equation 3.4 it is automatically the case the the equivalent noise density requirement is also satisfied. This is significant because we could allow the first stage to provide wide band gain and rely on a second stage to perform filtering. The second stage will have a capacitor gain product that is A_1^2 times smaller than if the fist stage had to perform filtering. This can has a large impact on analogue circuit area that is typically dominated by capacitors used for filtering and setting closed loop gain.

So far we have only considered the implication thermal noise requirements on the design. We must also address the flicker noise sources because neural signals have a lot of low frequency content. Moreover because flicker noise sources concentrate the noise power at the lower frequencies, the total noise profile inside the LFP frequency band can be dominated by this type of noise. The nature of flicker noise from transistor physics can be due to a number of phenomena; mobility fluctuation $\Delta\mu$, carrier density fluctuation ΔN, and changes in access resistances ΔR. Each of these phenomena will exhibit a $1/f$ frequency dependence when computing the input referred power spectrum. Typically for a given inversion coefficient IC only one of these phenomena will dominate the overall noise characteristic of a transistor. This is illustrated in Figure 3.4 which shows that ΔN is typically the leading cause for flicker noise generated

additively to the drain current. IC is a factor that indicates to what extent a transistor is operating in the subthreshold region by using the definition IC=$\mathrm{I_D}/(2\mu\, C_{ox}\, W/L\, U_T^2)$. This uses The more general parameters q, W, L, C_{ox} that represent electron charge, transistor width, transistor length, and gate oxide sheet capacitance respectively. The region of interest for biomedical circuits is typically when $IC < 1$ which exhibits good current efficiency and subthreshold operation. The phenomenological model corresponding to carrier density fluctuation ΔN component is expressed in Equation 3.5 after being referred to the transistor gate as an equivalent voltage noise density [119].

$$e_{fl}^2\,\Delta f = \frac{q^2\, kT\, \lambda\, N_T}{W\, L\, C_{ox}^2\, f}\cdot K_G \quad \text{where} \quad K_G \approx (1+\frac{\alpha\,\mu}{2})^2 \quad \text{for} \quad IC<1 \tag{3.5}$$

Here α and λ represent the coulomb scattering coefficient and tunnelling attenuation distance respectively. Notice that this expression has relatively weak biasing dependency in weak inversion contrast to the strong inversion region as shown in Figure 3.4. This trend follows very closely to the $(gm/I_D)^2$ characteristic which implies a fixed SNR for varying IC. The parameter N_T reflects the density of trapped charges at the oxide interface inside the transistor's conducting channel. Whether this parameter is consistent across various technology nodes is naturally put into question [120] but similarly there is evidence supporting that indeed this factor is process independent [121]. Now we should keep in mind that increasing the input transistor size will accommodate lower flicker noise but also result in increased noise. This is because of the signal loss when coupling C_{in} to C_{fb} that is loaded by the parasitic input capacitance of the amplifier C_g (see Figure 3.3). Keeping the ratio C_g/C_{fb} fixed as a δ, we can express the required input capacitance in Equation 3.6 in terms of general amplifier requirements using A_{cl}, K_F, f_{cor} as the closed loop gain, flicker charge density, corner frequency respectively.

$$C_{in} = \delta A_{cl}\, C_g = \frac{3}{4}\delta\, A_{cl}\, C_{ox}\, W\, L = \frac{3}{4}\frac{K_F\, A_{cl}\, \delta}{C_{ox}\, e_{in}^2\, f_{cor}} \quad \text{where} \quad K_F = q^2\, kT\lambda\, N_T\, K_G \tag{3.6}$$

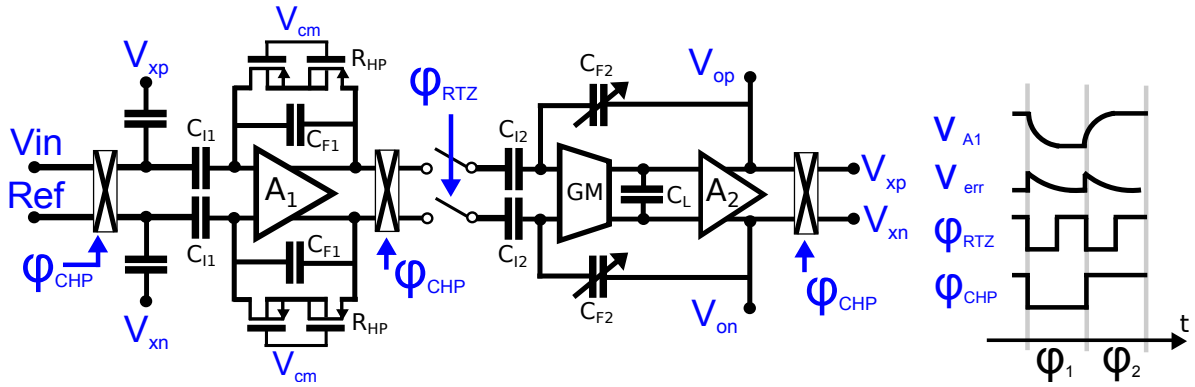
This expression indicates that attempting to achieve all desirable characteristics; small e_{in}^2, small f_{cor}, large A_{cl} simultaneously in a single amplifier structure comes at the cost of a very

large input capacitance that scales proportionally to all factors. This representation suggests the Harrison topology has limited flexibility for improving input capacitance as the only solution appears to be minimising $\frac{KF}{C_{ox}}$ through CMOS process selection. Moreover δ cannot be made arability small as it will more typically be bounded by the minimum feedback capacitor C_{fb}. This need to be large enough to set the high pass pole location at sufficiently small frequency to prevent the resistor R_{fb} from introducing noise inside the signal band which has a integrated power of $\frac{kT}{C_{fb}}$ [111]. Not to mention that the resulting size of the input transistors can be very large for this particular topology.

3.2.2 Chopper Stabilized Amplifiers

Alternatively we can apply chopping techniques to deal with these noise requirements which is used extensively in bio-signal instrumentation systems [67]. By up modulating the signal to a higher frequency before amplification, the flicker noise is added to the usual near-DC band which no longer coincides with out input signal. The output is then demodulated to recover the input. The difference is that the flicker components now lie at the chopper frequency f_{chp} which is typically out of band and can be rejected easily. This eliminates the requirements from Eq 3.6 on the input capacitance and shifts the focus to rejecting up modulated aggressors at higher frequencies. We suggest using keeping the sampling and chopper frequency coherent because it allows low order FIR filter reject all up modulated harmonics. For instance by chopping at the half the Nyquist frequency ($f_s/2$) or odd multiples of it (i.e. $[2n+1]\, f_s/2 \mid n \in \mathbb{Z}$) will fold chopper harmonics onto $f_s/2$. The resulting filter are quite relaxed because of the large fractional bandwidth in the transition band that separates our signal bandwidth f_{3dB} from f_{chop}. In this particular case we employ a sampling frequency of $25\, kS/s$ and use a chopping frequency at $37.5\, kHz$ to achieve this functionality [122]. Conveniently any common mode signals from the sensor or analogue supplies are also rejected using this configuration because they will appear at the chopper frequency.

In addition to basic chopping functionality, the performance can be further improved by pro-



(a) Chopper stabilized topology employing ripple reduction, enhanced settling and impedance boosting.

(b) Clock signals ϕ_{CHP} and ϕ_{RTZ}.

Figure 3.5: Proposed compact chopper stabilized neural amplifier topology.

viding closed loop feedback to actively cancel aggressors on top of filtering the resulting up modulated aggressors. This can be achieved in multiple ways and in some cases could improve linearity. One possible technique is using a DC-servo loop and another is performing ripple rejection both of which remove different components [123]. Here we will consider the implementation of three such techniques that improve chopping performance that specifically have negligible power and area requirements. The considerations made here will be similar to that of [124] [125] but with explicit focus on area reduction.

Figure 3.5 shows the proposed configuration that promises a significant reduction in input capacitance and the required silicon area. This configuration has two gain stages where the first stage A1 is a wideband low noise stage and the second provides A2 low pass filtering as motivated by Section 3.2.1. This enables the rejection of flicker noise from the first stage completely and effectively shifts the corner frequency of the second stage by gain of first stage squared. Moreover this the configuration does not require auxiliary integrators provide feedback on the capacitive feedback network around A1 that would lead to increased complexity.

The pseudo resistor across A1 in this configuration provides closed loop rejection of low frequency noise below the high pass pole with the time constant $\tau_{HP} = C_{F1}R_{HP}$. The noise components in the band from f_{HP} to f_{CHP} will be a mixture of flicker and thermal noise that are up-modulated by the chopper proceeding A1. This is because A1's corner frequency will lie inside of this band after sizing the input transistors such that C_G is about 5% of C_{IN} which

usually leads to a target area of about $100\,\mu m^2$.

It is important that the gains of A1 and A2 are carefully selected because this configuration only provides a first order role off in terms of analogue filtering. It could be that f_{CHP} is not sufficiently outside of the $6kHz$ filter bandwidth resulting in some aggressors to appear on the output of A2. For this reason we require an aggressive high pass pole location to minimize this total up modulated power. More specifically we can say that the total noise contribution below the f_{HP} is mainly from R_{HP} which has a noise power of $\frac{kT}{A1^2\,C_{F1}}$ when referred to the input of A1. The main concern here is that we have to sacrifice a small amount of dynamic range on A2 to prevent distortion. Although this power is quite limited, we need make sure the FIR filter can reject up-modulated noise components effectively.

In addition we take advantage of the reduced output referred sampling noise a the input of the second stage that scales by f_{3dB}/f_{CHP}. This is because most of the sampling noise will lie outside of the filter bandwidth. The size of C_{I2} can be reduced to alleviate the slewing errors due to the band limited behaviour of A1. In addition the parasitics at the output of A1 will pre-charge before C_{I2} is connected reducing the settling error due to the active nodes switching at the input and output.

A common concern for chopper stabilized circuits is the resistive element of each chopper which in this case is seen at the input of the amplifier. This resistance is due to the switching capacitor C_{I1} that is continuously dissipating dynamic current. This can partially be compensated for by performing positive feedback from the output to assist in cancelling the dynamic current through C_{PF} [123]. However this will rely on the matching of the capacitor ratios $\frac{C_{I1}}{C_{F1}}\frac{C_{I2}}{C_{F2}}$ to be equal to $\frac{C_{I1}}{C_{PF}}$. This can be quite challenging if small configurations are desired that do not need exhaustive calibration. The use of a high precision ADC makes this somewhat easier because the total gain A1·A2 does not need to be as large implies smaller ratios and better matching. Evaluating this resistance in terms of the switching capacitance will result in the expression in Equation 3.7.

$$R_{in} = \frac{1}{2f_{CHP} \cdot (C_{I1} + C_{par} - \frac{C_{I1}}{C_{F1}} \frac{C_{I2}}{C_{F2}} C_{PF})} \tag{3.7}$$

This dependency should indicate that if the dynamic switching current cannot be well matched due to parasitics or variability the next objective would be to reduce the total switching capacitance. From our discussion however it appears that reducing the input capacitance is limited by the psuedo resistive noise that induces aggressors at the chopper frequency. This constraint can be mitigated using a distributed amplifier structure that splits A1 into two identical sections. This should be configured such that the second stage has its high pass pole and corner frequency proportionally larger than the first stage but scaled the gain of the prior stage. However such a topology is more constrained by parasitics that worsen the settling errors in $0.18\,\mu m$ CMOS. In addition the poor control of psuedo resistive characteristics does not allow this to be convincing solution [126]. The feasibility may be more favourable in more advanced technology nodes. Considering a value of $1\,pF$ for C_{I1} we expect slightly over $20\,M\Omega$ without positive feedback and approximately $200\,M\Omega$ with 10% matching. This may be acceptable in either cases depending which type of electrode is used but generally any thing above $100\,M\Omega$ is satisfactory for most scenarios.

3.2.3 Bandwidth Limited Amplifiers

Biomedical instrumentation has the advantage that the slowly varying signals prevent most implementations from facing problems due to limited bandwidth. The exception however lies with the stage that drives the input capacitance of the ADC and the settling time during sampling can be quite challenging [127]. Particularly when multiple channels are multiplexed to the same data converter. In some sense there is an similarity when we look at NEF and bandwidth efficiency because they are strongly dependent on maximizing transconductance efficiency.

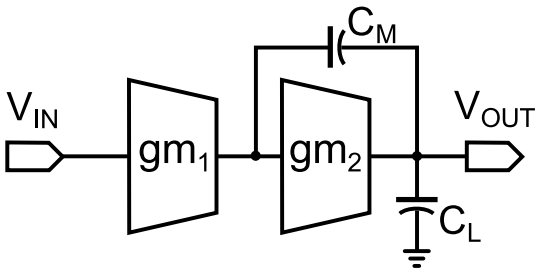


Figure 3.6: Analytic model of a two stage Miller compensated amplifier.

$$FOM\left[\frac{MHz\,pF}{mA}\right] = \frac{f_{UGF}\cdot C_L}{I_{tot}} \quad \text{for diff. pair} \quad FOM = \frac{10^3}{4\pi\,\eta U_T} \tag{3.8}$$

Strictly stated in Equation 3.8, a bandwidth constrained circuit should minimize the total current consumption I_{tot} for a given unity gain bandwidth f_{UGF} and capacitive load C_L. It is typical to find dedicated structures out side of the signal processing chain that drive the ADC input capacitance and focus specifically on maximizing the FOM by employing current recycling, adaptive biasing, and positive feedback techniques. The challenge here is efficiently introducing these techniques while also preserving the capability for full output swing, stability and particularly low distortion. The later is likely the most challenging and demands high loop gain that is generally not found in adaptive single pole structures if full output swing is also required. With that said, two stage Miller compensated topologies can provide an excellent solution to this problem because high gain in the second stage will suppress a number nonlinearities excited by the input stage. Further more the capacitive coupling of the output to the input of the second stage implies the settling speed is limited by bandwidth of the second stage. This allows the configuration to simultaneously provide filtering and settling while sharing many of the biasing and feedback elements. Using the model shown in Figure 3.6. We can show that sampling induced kick back from the ADC at V_{out} has negligible in pact on internal integration node as it is inversely proportional to the product $A_{cl}\cdot\frac{gm2}{gm1}$ where $gm1$ and $gm2$ are the transconductance of the first and second stage. This is derived from evaluating a step response due to discharging the output load C_L which has the Laplace domain response as Equation 3.9.

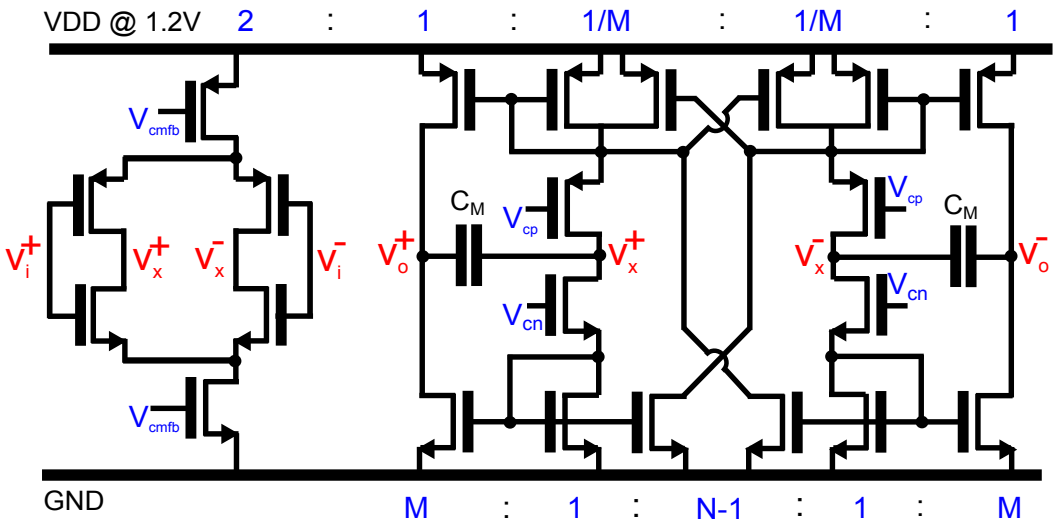


Figure 3.7: Proposed structure high loop gain amplifier stage utilising complementary transconductive input and the high gain current conveyor topology. Transistor sizing is detailed in Figure C.7.

$$H_{step}(s) = \frac{s^2}{s^2 + (\omega_2 - \frac{\omega_1 C_M}{A_{cl} C_L})\, s + \frac{\omega_1 \omega_2}{A_{cl}}} \quad \text{where} \quad \omega_1 = \frac{gm_1}{C_M} \;,\; \omega_2 = \frac{gm_2}{C_L} \tag{3.9}$$

Figure 3.7 shows the proposed circuit implementation of the two-stage amplifier used inside the second instrumentation stage in Figure 3.5. This structure has the advantage of providing very high loop gain across the Miller capacitor and allows full output swing due to the positive feed back structures in the current mirrors. The PMOS mirror provides high gain by cancelling the $1/gm$ transresistance of from the diode connected pair leaving the high impedance node and the NMOS mirror provides positive feedback to speed up the transient behaviour on the PMOS side. When this structure provides closed loop gain larger than 20 dB it is sufficient to rely on the NMOS current mirror for stability. In fact this is equivalent to a feed-forward stabilization technique that by passes high frequency signal lag induced by the pole at the PMOS side. However when good phase margin is required at the unity gain frequency stability becomes more stringent. In this case we suggest introducing an additional capacitor across V_n & V_p to realize a zero that cancels the pole in the PMOS branch [128]. The zero will in fact boost the effective ω_2 from $N\,M\,\frac{gm_{M5}}{C_L}$ to $\frac{N\,M+M}{2-N}\,\frac{gm_{M5}}{C_L}$. The factor M in this structure has a rather interesting implication with respect to NEF. If M is large enough this topology will have a NEF equivalent to the complementary structure. However in effect the biasing current of

the intermediate branch is reduced when M is large which can move the parasitic poles in side the amplifier bandwidth. The apparent trade off between stability and NEF is unique to this structure but it is not challenging to have M=8 for low power applications.

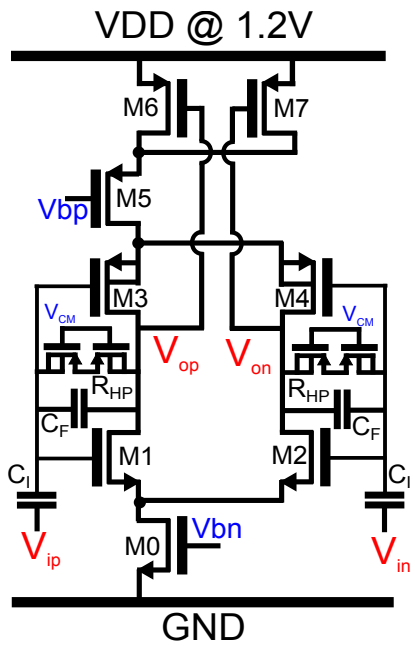
$$FOM = \frac{10^3}{4\pi\,\eta U_T} \cdot \frac{2\,C_L}{C_M\,(1 + M/K + 1/K)} \tag{3.10}$$

The components that improve bandwidth efficiency are detailed in Equation 3.10. Referring this back to Equation 3.3 however implies the noise is dominated by the capacitor that introduces the dominant pole of the system. The observation made here is that unlike the single stage topologies, the two stage configuration can trade off input referred noise for a better speed FOM by adjusting the $\frac{C_M}{C_L}$ ratio. The high level methodology applied here is replacing the C_L with a smaller capacitor that requires less power with the hope that stability can still be maintained by boosting transconductance power with positive feedback and current recycling.

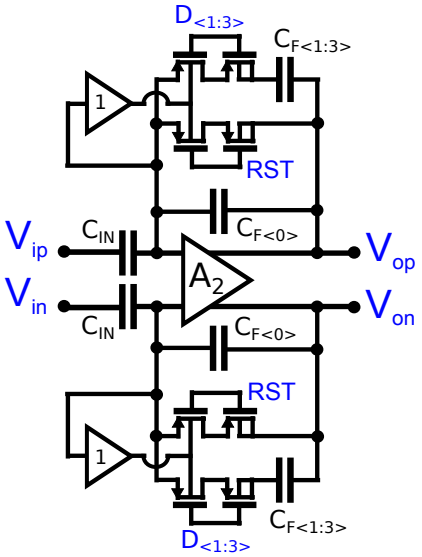
3.2.4 Circuit Implementation

Figure 3.8 shows the transistor level implementation of the topology used in Figure 3.5. The first gain stage is a highly compact complementary structure that exhibits exceptional noise performance. The second stage transistor implementation is the high gain two-stage topology discussed in Section 3.2.3. The variable gain configuration is facilitated by the digital controlled low leakage switches that connect a selected set of capacitors in feedback. This particular configuration provides more generic instrumentation of the 1 Hz to 6 kHz bandwidth. It is well known that the analogue filters introduce frequency dependent group delay near the pole locations which has been shown to degrade processing capabilities of spike sorting techniques [129]. By placing the high pass pole well inside of the LFP band the spike wave-forms exhibit less distortion due to analogue filtering and is instead filtered using linear phase filters in the digital domain that do not suffer from such drawbacks.

The reset mechanism on instrumentation amplifiers using pseudo-resistive elements is essential.



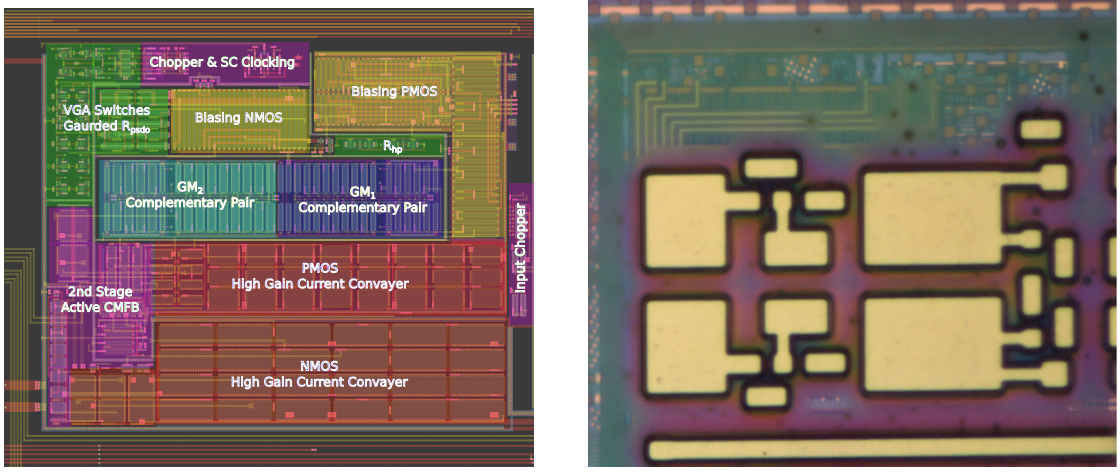
(a) Schematic of the topology used in the first gain stage. Transistor sizing is detailed in Figure C.9.



(b) Configuration of the feedback structure for reconfigurable gain and reset.

Figure 3.8: Schematic showing circuit implementation of the proposed compact neural amplifier.

Either during stimulation, start-up, or amplifier saturation the charge across the feedback capacitor must be neutralized before correct operation can begin. This mechanism allows the rejection of various distortion components that would other wise corrupt the latent signal integrity or digital signal processing. However there is an inherent problem with these reset switches due to the parasitic charge injection induced on the intermediate semi-floating nodes. Moreover if these elements are cascaded to increase resistance or dynamic range these sensitive floating nodes are also increased thereby building up more residue charge. A significant amount of charge can introduce a permanent reset artefact after reset as this charge redistributes internally inside the resistor. The proposed solution to this problem is by minimizing the floating nodes and guarding the floating N-Well from injected noise. This should allow a very large pseudo resistance for a sub-Hz high pass cut off frequency while maintaining exceptional reset characteristics. We minimize the resulting charge residue by absorbing the leaky diode currents and residues into the guarding amplifier. Now there will be some instantaneous off-set as the reset signal injects charge directly onto the feedback capacitor but this can be quite small when using small switches. The drawback here is that there may exist a very slow drift on the



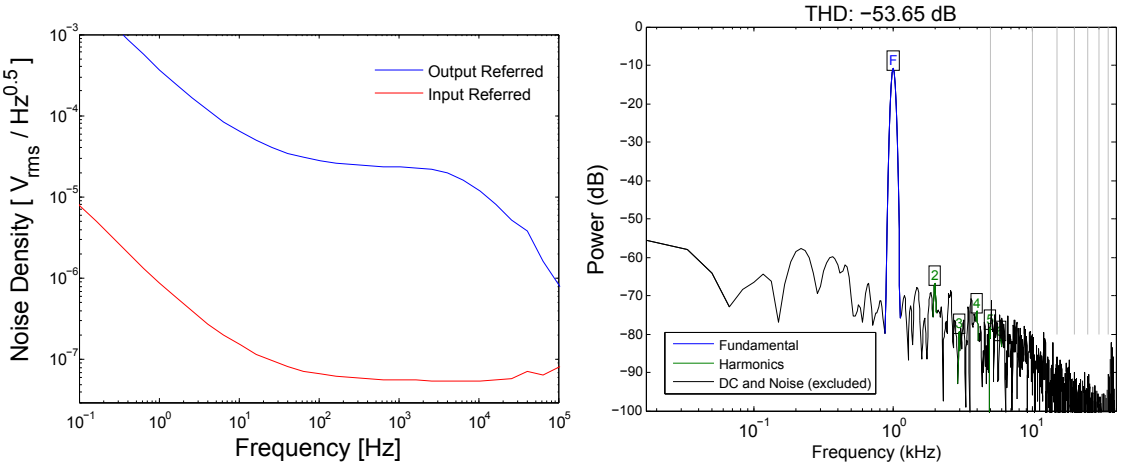
(a) Floor plan of the fabricated instrumentation amplifier.

(b) Chip micrograph showing the configuration of the feedback capacitors.

Figure 3.9: Physical implementation of amplifier using a 6-metal 0.18 μm CMOS process measuring $75 \times 82\,\mu m^2$ in size.

order of V/sec from the guarding amplifiers due to $V_{os}\,R_{diode}$. But simple digital assistance will suffice in eliminating this concern by periodically resetting the structure and cancelling the residue off-set. This re-introduces the high pass pole at a well defined location depending on the periodicity of the reset signal and reconstructing signal in the digital domain [130].

The floor plan for this implementation is annotated in Figure 3.9. The typical focus for analogue layout is achieving good matching for the input transistors and capacitors to minimise off-set or undesirable signal coupling. In this case the chopper introduces a lot switching that is difficult isolate from the signal so instead we focused on minimising parasitics of the clocked nets. The common mode feedback on the second stage uses a switched capacitor and wide band amplifier to ensure accurate common mode settling without deteriorating linearity. This is important because the ADC can be quite sensitive to the sampled common mode resulting in a reduced precision if there is an unexpected offset on the sampled output. Simulated performance of the implemented topology is shown in Figure 3.10. This compact configuration can achieve an input referred noise of 5.6 μV_{rms} over the specified bandwidth with a noise corner frequency of 20 Hz. The performance is detailed with a clear reduction in size can be observed when compared to other chopper systems in Table 3.3. The total gain is 421 V/V for this particular configuration which can be adjusted using the digital calibration bits integrated into the structure allowing



(a) Periodic steady-state analysis demonstrating input and output referred noise spectrum of the amplifier.

(b) Transient noise simulation demonstrating a $52\,dB$ THD for a $2.45\,mV$ *at* $1\,kHz$ sine input.

Figure 3.10: Post layout simulated results of the proposed instrumentation circuit.

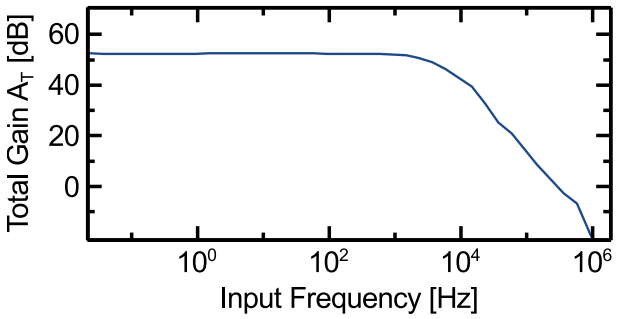


Figure 3.11: Post layout simulated results using periodic steady state analysis to evaluate the closed loop gain of the instrumentation circuit.

different gain and power settings. The maximum available gain setting is shown in Figure 3.11.

Overall the proposed implementation performs well for supply voltages larger than $1.1\,V$ where the limiting factor is due to the current biased complimentary input stage. This configuration necessitates a voltage overhead requirement of $2V_{TH} + 2V_{ov}$. However both of the gain stages are class-A which at exhibit relatively well behaved current transients on the supplies. Class-AB alternatives do not share this feature and are more prone to disturb neighbouring recording circuits. Minimizing the dynamic current dissipation should lead to better LDO performance and lower supply induced sensor noise when many channels are integrated together. This also motivates another aspect for using a wide-band amplifier configuration for the first amplification stage because it usually implies that the common mode will also have wide band regulation. This leads to better common mode rejection in the signal band due to additional

Table 3.3: Summary of performance specifications of the proposed instrumentation topology and other bio-signal chopper stabilized amplifiers found in the literature.

Parameter	Units	This Work	Markovic [125]	Makinwa [123]
technology	[nm]	180	40	65
Supply Voltage	[V]	1.2	1.2	1
Total Current	[μA]	1.05	1.67	2.1
Bandwidth	[Hz]	$<$ 1-6 k	$1-5k$	$0.5-1k$
Filter Order Roll-off	[dB/Dec]	20	20	20
Noise Floor	$[nV/\sqrt{Hz}]$	55	101	60
Noise Corner	[Hz]	20 Hz	100	<1
Dynamic Range	[dB]	58	69	64
Area	[μm^2]	$6.2 \cdot 10^3$	$7.2 \cdot 10^4$	$2 \cdot 10^5$
Area-Power-Product	[μW μm^2]	$7.3\,10^3$	$141\,10^3$	$420\,10^3$
NEF		1.08	2.5	1.66

loop gain.

3.3 Analogue Signal Conversion

Analogue to digital conversion remains to a crucial component instrumentation, particularly for full signal characterization. Even when considering the demanding constraints for integrated neural sensors, the prevalence of full spectrum signal characterization is ubiquitous in the literature. This is motivated by the efficiency and reliability of various digital processing methods that require very efficient signal conversion to discreet samples instead of processing recordings in the analogue domain. Typically the most valued performance criteria for such a system is the ADC power consumption. A Successive Approximation Register (SAR) ADC is commonly used for quantizing biomedical signals because it only dissipates switching energy that can be very small for slow sampling rates. The SAR topology is depicted in 3.12 and can be found extensively in BMI recording publications.

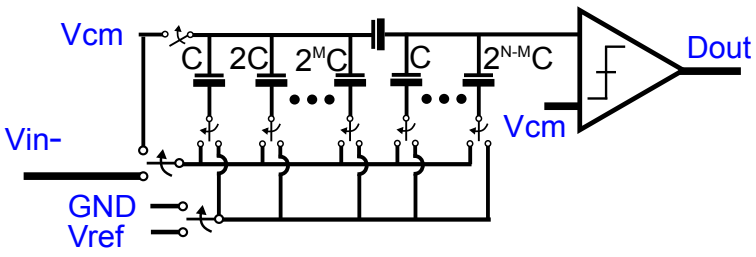


Figure 3.12: Schematic of a conventional N bit SAR ADC with the split capacitor at position M.

3.3.1 Capacitive array miniaturization

This discussion pays special attention to acquiring neural recordings that include LFPs while minimizing the required silicon area per sensor. This is motivated by wanting to integrate many sensors on chip for large arrays and secondly reducing any capacitive switching noise that can be quite difficult to reject in fully integrated systems. Recording LFPs and EAPs simultaneously will require increased ADC resolution so that the instrumentation dynamic range exceeds 60dB. Equivalently this means 10 to 14 bit precision is needed depending on the nonlinearity tolerances of the proceeding processing methods. This can be quite difficult it terms of the SAR specifications because the capacitor mismatch and sampling noise can prevent aggressive sizing on the unit capacitor. For a given ADC precision N, the SAR capacitor array will require $M \cdot 2^{N/M} + M$ unit capacitors C_{unit} where M is the number of equally split sections. By splitting the array into sections it should be obvious that the total capacitor requirement C_{Total} can be reduced to some extent. The quantization errors resulting from capacitor mismatch on the other hand is also closely related to these parameters. For given standard deviation σ and confidence interval CI we can use Equation 3.11 to make a simple estimate for the expected quantization error E_Q [131].

$$E_Q = V_{ref} \frac{\sum \Delta Ci}{2^N C_{unit} - \Delta C_{Total}} = V_{ref} \cdot \alpha(N) \frac{\sqrt{2^N} - 1}{\sqrt{2} - 1} \tag{3.11}$$

The above expression assumes no split configuration is used where α represents a scaling factor that is dependent on the number control bits for each sub-DAC, $\alpha(x) = \frac{CI\sigma}{2^x - CI\sigma\sqrt{2^x}}$. V_{ref} is the reference voltage by which the sampled input is normalized to arrive at the binary encoded

result. Now extending this formulation to include the dependency of M and bounding $E_Q < LSB/2$ in accordance with the required ADC precision leads to the expression in Equation 3.12.

$$\frac{1}{2^{N+1}} \geq \alpha(N/M) \sum_{k=0}^{M} \left[\sum_{i=0}^{M} \sqrt{2^i} \right] \cdot \left(\frac{\alpha(N/M)}{2^{N/M}} \right)^k \tag{3.12}$$

There several higher order terms with respect $\sigma\, CI$ not shown here because they have vanishing contribution as N is increased and require a numerical solution to the problem. Otherwise for M=2 and arbitrary placing the split capacitor K position in the array we can similarly reconstruct the equality from 3.11 in Equation 3.13.

$$\frac{1}{2^{N+1}} \geq \frac{CI\sigma(\sqrt{2^{N-k}} - 1)}{(2^{N-K} - CI\sigma 2^{\frac{N-K}{2}})(\sqrt{2} - 1)} + \frac{(\sqrt{2^K} - 1)CI\sigma}{(\sqrt{2} - 1)2^K(2^K - CI\sigma 2^{\frac{K}{2}})} \tag{3.13}$$

The standard deviation σ is closely related to the exact requirements for the whole capacitive DAC in terms of the total area and unit capacitor size. The dependency of E_Q is mainly subject to the variance due to the MSB capacitors and for each less significant bit (from MSB to LSB) the expected variance increases by $\sqrt{2}$ while its capacitive coupling decreases by 2. This is because $\sigma \propto 1/\sqrt{A_C}$ where A_C is the area of the capacitor. Clearly there is a process related figure of merit here that relates to the quality of capacitors since small capacitors with excellent matching will result in the best characteristics ADCs such that we minimize the % deviation per $\mu\, m^2$.

Figure 3.13 shows a numerical solution to the equality in Equation 3.13. This allows us to consider the effect of split capacitor positioning with respect to the optimal area allocation for the capacitor array. The visible plateau for small N represents the case when the design is bounded by the minimal unit capacitance. This is determined using the process documentation for the target 0.18 $\mu\, m$ CMOS technology that gives its mismatch specifications and minimum sizing. Generally split capacitor configurations are more sensitive to parasitics they can lead to more pronounced nonlinearities. However in some cases that the unit capacitor size limits the array size such that splinting the array is an effective solution for improving power dissipation.

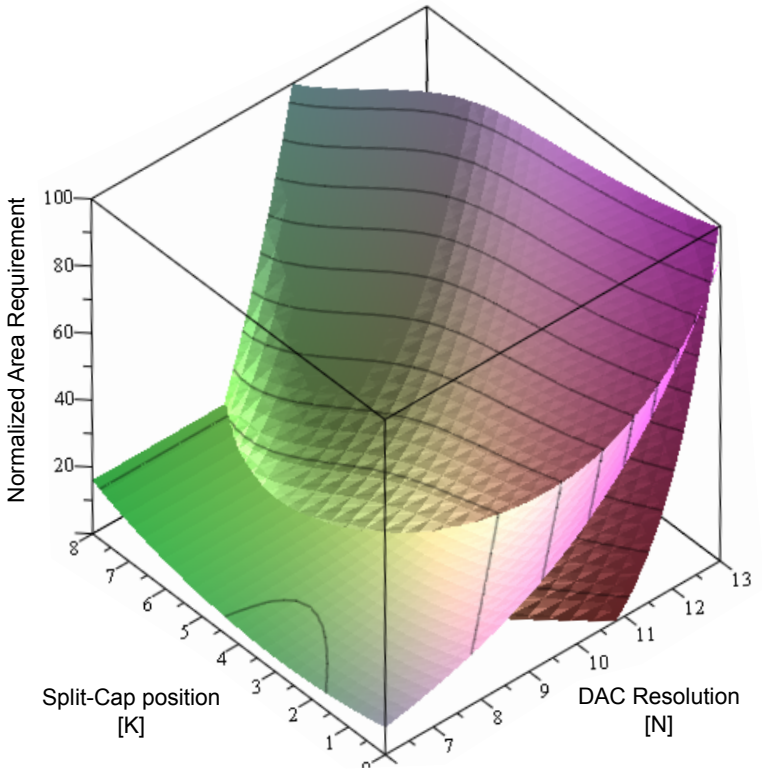


Figure 3.13: Numerical solution to Equation 3.13 relating the capacitive DAC area requirement with the DAC resolution (N) and the position of the split capacitor before capacitor K.

We reiterate that this also indicates that the binary weighted configuration without splitting maximizes area efficiency if we are not limited by sampling noise or minimal capacitor sizing. In addition a fully differential DAC counter intuitively reduces the minimum size if the switching method first detects polarity before applying successive feedback [132]. This is because the first quantization cycle does not depend on the capacitive division. This in turn means that the array can tolerate twice the mismatch error implying a 4 times smaller unit capacitance while only doubling the number of capacitors in the array.

3.3.2 Model based topology selection

From here there are multiple directions we can take in order to ensure efficient operation and simultaneously achieve a compact configuration. A common approach is to multiplex the SAR ADC to a large number of channels but this will also require the analogue stage driving the ADC to dissipate proportionally more power due to settling requirements on the sampling

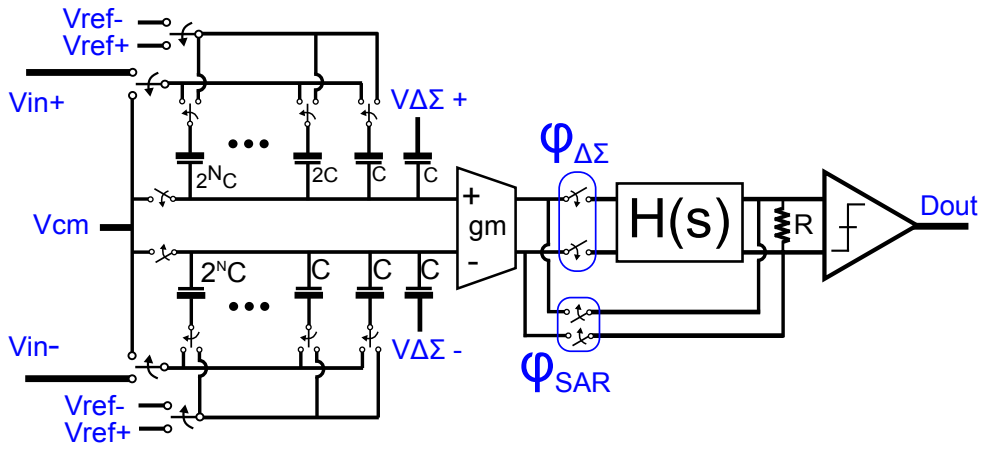


Figure 3.14: Schematic of the proposed ΣΔ assisted SAR ADC topology for achieving a more compact configuration.

capacitance. From a high level perspective, distributing the quantization effort into a large array of ADCs with staggered operation should lead to much more systematic power dissipation due to their uncorrelated operation. Opposed to using a single high speed ADC that requires a much higher clock frequency with stronger tones in the generated supply noise. Another SAR based alternative using calibration for the capacitive array such that it can specifically be designed with the smallest possible unit capacitors. Then we could correct any nonlinearity or quantization errors that arise from capacitor mismatch if the array is characterized precisely enough. This does require either foreground or background calibration modules to extract the individual capacitor weights. Because we aim to perform a number of processing techniques in the digital domain for characterizing neural recording, it makes sense for us to consider effective means to perform calibration.

The structure illustrated in 3.14 represents a hybrid topology based on SAR and sigma delta structures. The motivation is driven by the efficiency of SAR quantization for large signals and the compactness of high resolution quantization from sigma delta loops. The digital control will perform fully differential bottom plate sampling of the input which is then rapidly quantized to 2^N levels using the typical binary search. After the SAR operation the resulting residue left on the capacitive array is quantized using a sigma delta control loop that feedback on the nodes $V\Sigma\Delta\pm$.

There is a strict advantage over conventional sigma delta loops which is that the residue error

that needs to be quantized is reduced to $\frac{V_{ref}}{2^N}$ which can easily be designed to lie within the linear range of a differential pair. This negates having to use passive or active feedback to deal with transconductance nonlinearity and significantly improves the power efficiency by retaining a relatively simple control loop topology. Moreover as the feedback loop is typically responsible for small dynamic range of 30dB the requirements on clock jitter and decimation filtering is made more relaxed.

The more desirable advantage over a high resolution SAR is that the capacitive DAC may designed in a highly optimal configuration with as few bits as possible. This allows sizing that primarily focuses on suppressing parasitic effects with minimal sampling capacitance. As will be demonstrated this topology does not require an axillary calibration DAC or a pseudo random dithering source for performing mismatch correction. This is due to the capability that the internal sigma delta structure is in the same signal loop as the SAR operation and can trade-off bandwidth for increased noise rejection simply by adjusting the sampling frequency f_s. Naturally because this topology inherently needs a pre-amp stage for SAR conversion we should not expect the FOM to do better than low resolution SARs.

Intuitively one can think that when combining the two topologies the individual sources for power dissipation now scale with $2^{\frac{N}{2}}$. More specifically these sources come from the capacitive DAC and decimation filter. The components that do not have reduced scaling are related to the sampling noise and the thermal noise floor of the oversampling modulator. To demonstrate this quantitatively we will build an analytic model for the SAR and $\Delta\Sigma$ SAR topologies to demonstrate some of the inherent characteristics. This will also reveal the techniques for optimizing of the proposed structure.

$$FOM_{ADC} = \frac{P_{sys}}{2^N f_s} \tag{3.14}$$

Maximizing the performance indicator from Equation 3.14 will represent our objective function which reflects the efficiency by which each sample is converted into a digital code. Through the simplicity of this relation, any comparison primarily requires an accurate expectation for

power budget in terms of the required resolution or precision requirements.

$$P_{Ideal} = \underbrace{E_{search} \cdot f_s C_{unit} V_{ref}^2 (2^{N-2} + 2^2)}_{Capacitor\,Array} + \underbrace{2N(N+2) f_s E_{gate}}_{Register\,Logic} \tag{3.15}$$

Equation 3.15 Considers the primitive structure with an ideal comparator where E_{search} represents the average dissipation for binary search switching method and E_{gate} is the average gate dissipation per clock cycle. Both these parameters adjust to different core libraries or various switching methods that typically trade off efficiency for parasitic tolerance [133]. This ideal structure is extended by the requirements of either a dynamic latch comparator or an analogue pre amplifier that allows negligible comparator requirements at the expense of consuming a static current. The classic pre-amplifier approach also tends to deal with mitigating kick back noise but in general the straightforward application of classic kT/C relations conveniently give;

$$P_{amp} = 32\,\pi\,\ln(2) \cdot \underbrace{\frac{(U_T\,N\,2^N)^2}{V_{ref} \cdot \eta\,q}}_{Noise} \cdot \underbrace{A_{ol}\,f_s\,NEF^2}_{Bandwidth} \tag{3.16}$$

Here A_{ol} represents the gain provided by the pre-amplifier. Notice the very typical inverse relationship with respect to V_{ref} which motivates the use of the more efficient dynamic comparator structure. However evaluating the equivalent input referred noise of a dynamic structure accurately requires the a piece wise evaluation for different phases of operation and the respective stochastic integrals [134]. The contribution can be associated with two dominant sources, that of sampled noise;

$$\sigma_S = \frac{4\,kT}{3\,C_x\,F} + \frac{kT}{3\,C_x\,F^2\,H} + \frac{kT}{12\,C_x\,F^2\,H^2} \tag{3.17}$$

And noise contributed from transconductive elements;

$$\sigma_M = \frac{kT}{C_x\,F^2} + \frac{kT}{2\,C_x\,F^2\,H} + \frac{kT}{8\,C_x\,F^2\,H^2} \tag{3.18}$$

$$F = \frac{2\rho V_{th}}{V_{ref} - V_{th}} \qquad \text{and} \qquad H = \frac{V_{ref}}{2\,V_{ov}} \cdot 2\rho 1 + \rho \tag{3.19}$$

As before, this must be bounded by the acceptable quantization noise, $V_{ref} \cdot 2^{-N-2} = \sqrt{\sigma_M + \sigma_S}$, which give the values for C_x. Strictly there is a strong dependence on the input signal in order to evaluate the dissipated power but on average it is reasonable to approximate this to the capacitive switching energy of $P_{Latch} \approx f_s\, C_x\, V^2$.

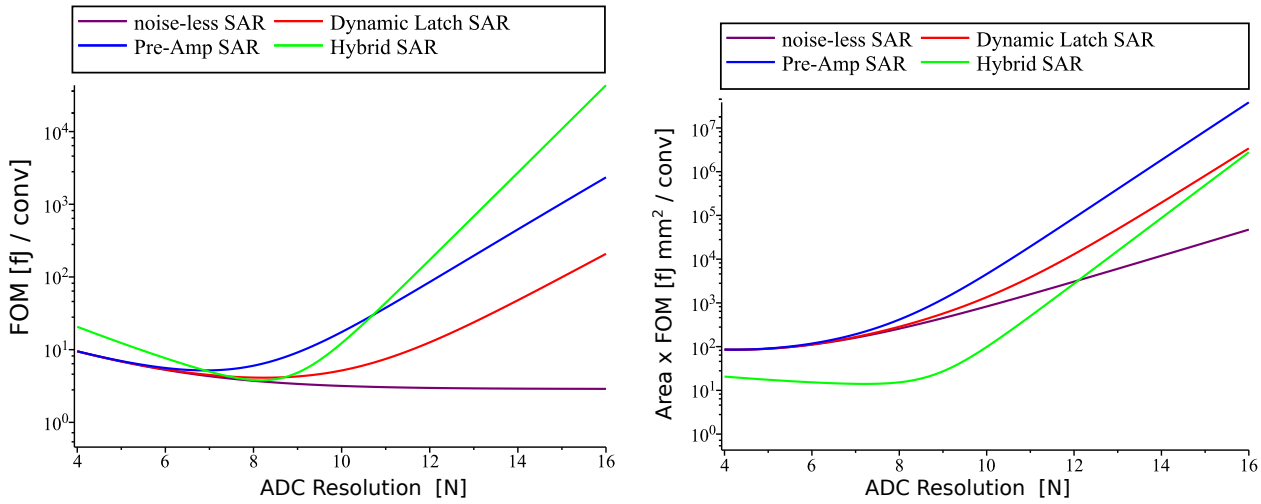
Now consider the components of the $\Delta\Sigma$ structure. Clearly it will follow closely to that of the pre-amplifier based relations with the exception that the primitive components from Equation 3.15. Instead this will scale with $N-K$ where K is the number of bits resolved by the sigma delta loop. Here two additional components will be accounted for, the first is the integrator and the second is the digital FIR that decimates the modulated residue quantization. A second order feed forward integration topology is chosen for $H(s)$ based on its efficacy of being applied to the configuration shown in Figure 3.14 and primarily minimizing the number of summing operators and coefficients prone to mismatch. For the sake of discussion we make the assertion that decimation noise rejection is bounded $K \leq (FIR)^{-1/2}$ in the case of a rectangular window for analytic clarity [135]. Furthermore, note that as we increase the SAR quantization the first stage will proportionally see a reduction of the input signal that needs to be accounted for to achieve the correct integration constant.

$$P_{int} = 32\pi \cdot \underbrace{\frac{(U_T\, 2^N\, NEF)^2}{q\, V_{ref}}}_{Noise} \cdot \underbrace{FIR\, f_s\, (1 + 2^{N-K})}_{Bandwidth} \tag{3.20}$$

And similarly the digital decimation filter will scale in the form of;

$$P_{fir} = \underbrace{2^K}_{OSR}\, \underbrace{(K + \log_2(K))}_{Quantization}\, fs\, E_{gate} \tag{3.21}$$

Collecting these terms for each topology will equate to expressions that typically have scalar dependencies on technology or implementation which we must make a set of reasonable as-



(a) Power efficiency scaling for various SAR structures

(b) Efficiency area product for various SAR structures

Figure 3.15: Summary of the FOM ($P_{sys}/2^N f_s$) for each topology with respect to different resolution requirements.

sumptions for. The literature will indicate numerous means by which each component can be reduced through specialized logic cells, adaptive comparator power allocation, or power saving switching methods. Our particular interest lies with the dependency on N that will imply the effectiveness of a certain topology for a given dynamic range requirement. In addition this familiarizes us with specific factors fundamental to power dissipation with respect to resolution.

Figure 3.15 presents the expected merit for each topology as the target resolution is varied. Without consideration for area, there is a clear power advantage for the dynamic SAR structure mediated primarily by the fact that the comparator does not have settling associated tolerance. This is the main reason why the pre-amp topology requires a proportionally increased bandwidth/power as resolution is increased. What stands out is that the $\Delta\Sigma$ structure has a power dependency $\propto 2^{3N}$ for achieving the required input referred noise in contrast to more conventional dependency of 2^{2N}. The mechanism behind this is due to the SAR quantization that reduces the signal input range which needs to be recovered to achieve the correct integration factors. Moreover the over sampling ratio increases simultaneously which has an overall multiplicative effect. Clearly the resolution of the SAR quantizer should only perform a few conversion that put the residue in the linear range of the loop filter and let the modulator perform most of the quantization effort. When all topologies are using the same unit capac-

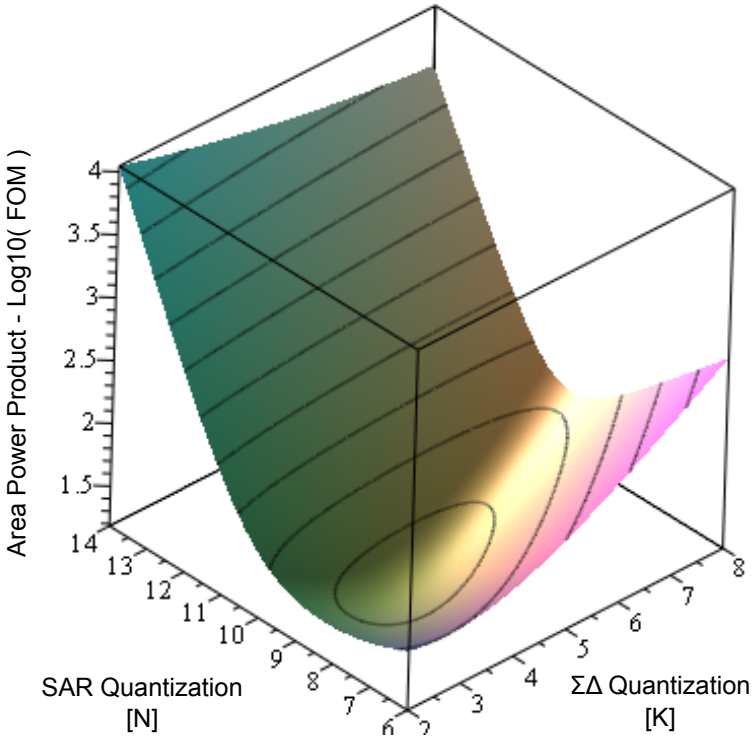


Figure 3.16: Figure of merit dependency of the proposed ΔΣSAR topology with respect to design parameters K1 & K2.

itor, this result demonstrates that for $N < 5$ & $N > 14$ the ΔΣ topology becomes strictly unfavourable in terms of power but performs comparably with respect to power efficiency for $N \approx 10$. Taking the FOM area product by considering the capacitors in terms of □ units the advantage of the ΔΣ topology becomes more obvious. For the precision significant to neural recording, $8 < N < 12$, the hybrid structure consistently grantees a more compact configuration by a factor of 10.

Considering the design space of the ΔΣSAR structure in more detail will expose a more optimal strategy for increasing FOM. Figure 3.16 exemplifies how the FOM behaves as either the SAR of sigma delta accuracy is increased. After the optimal basin at N = 9 & K=4 the best strategy for improving ADC resolution is by increasing SAR quantization at half the rate of the sigma delta increase in resolution. For reference a conventional ΔΣ modulator [136] is designed with the same target specifications and using the same design method to configure the OPAMP integrators and resistive input network. Such a configuration achieves 167 dB FOM_s irrespective of target resolution when we consider just the analogue power dissipation. In fact this figure is commonly achieved by state of the art [137]. As shown in Figure 3.17 the ΔΣSAR

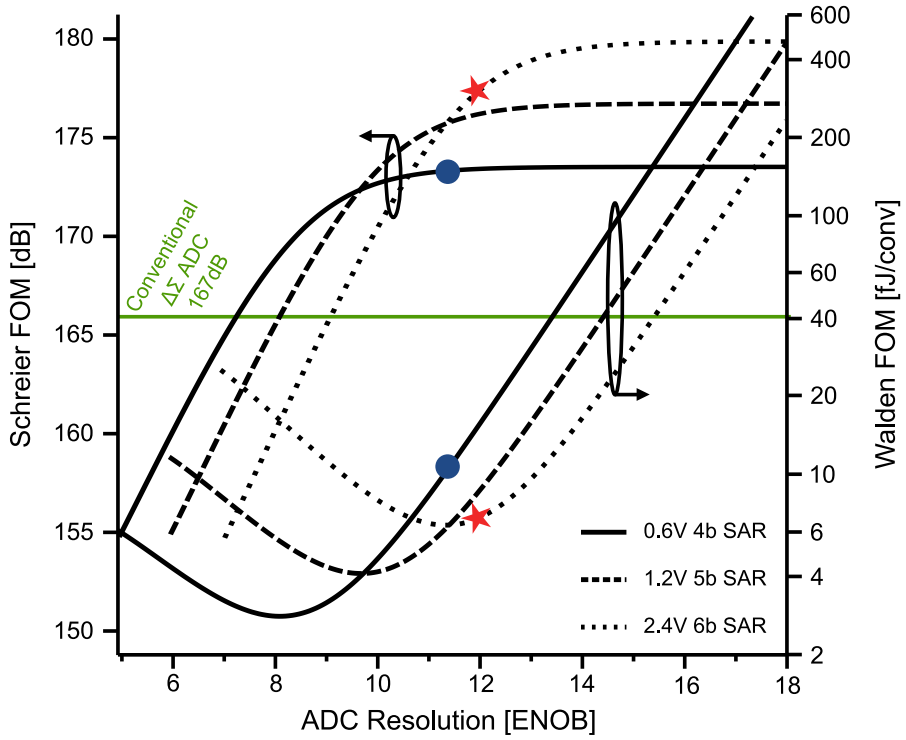
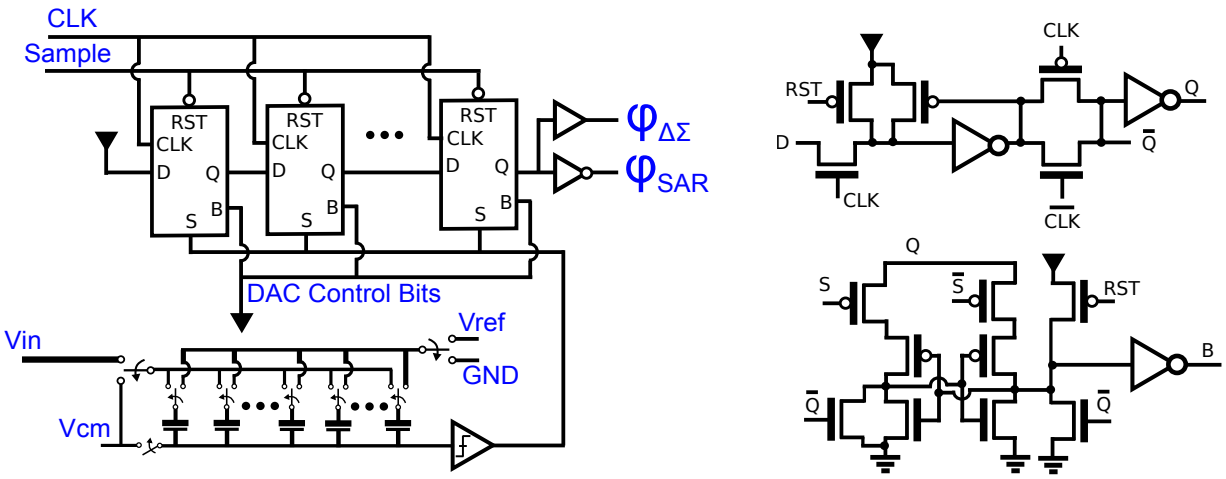


Figure 3.17: Estimation on the expected figure of merit for a target resolution and varying SAR precision. The red star and blue circle indicate the target and measured performance respectively.

configuration can theoretically achieve more than 4X better performance than conventional $\Delta\Sigma$ modulators for resolutions above 12 bits even when operating at lower supply voltages. This is because of the improved noise efficiency. Please refer to Section C for additional details regarding derivations and topology comparisons that are omitted here for clarity.

3.3.3 Circuit Implementation

Extending the conventional SAR structure to perform sigma delta modulation is achieved with relatively little changes to the overall topology. The main difference is that during the last phase of SAR conversion a register must be toggled that switches in the integrators intermediate to the comparator. Simultaneously the $V\Sigma\Delta\pm$ capacitors are directly connected to the comparator bipolar output instead of the common mode voltage VCM for differential feedback. This configuration is integrated on chip and performs 7 bits of differential SAR quantization with



(a) Converter FSM control

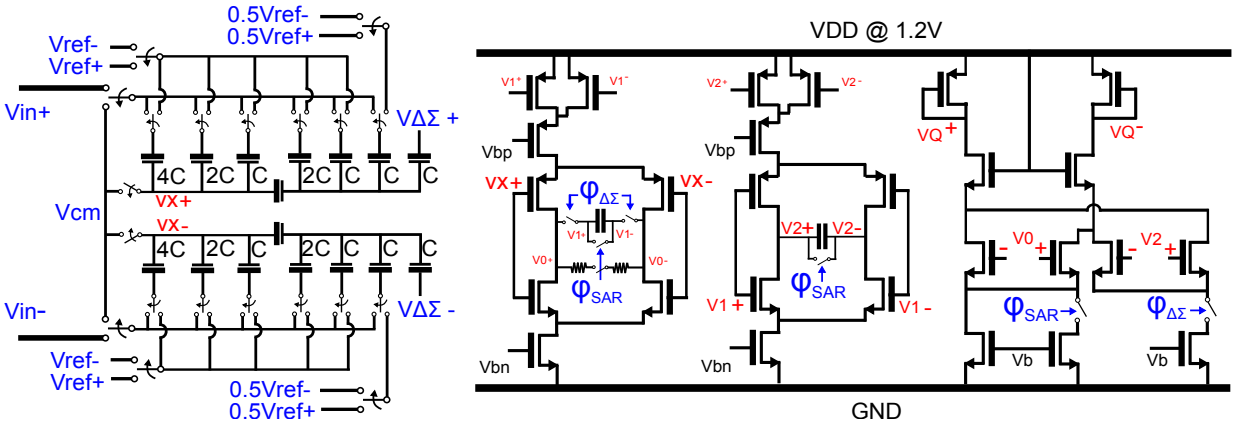
(b) FSM logic slice

Figure 3.18: Schematic configuration of the top level control for the ΔΣSAR data converter.

another 5 bits resolved by the noise shaping modulator with an over sampling rate of 32. At the system level, 4 analogue recording channels will be multiplexed to the input of the ADC which implies sampling rate of $100\,kS/s$ is required to sample each output at $25\,kS/s$.

Figure 3.18 shows the top level configuration of this data converter. By using a specialized register logic slice a small reduction in complexity is achieved in addition to the mitigation of timing issues typical with the conventional self clocking register configuration. This topology uses a bottom plate sampling strategy to neutralize the effect of parasitics and common mode comparator nonlinearities while operating at 1.2V with a 10MHz clock frequency. Although there are only $N - K + OSR$ active phases, settling the output of the recording amplifiers on to the capacitor array will require several cycles because of the band limited behaviour present in the driving stage.

The implementation of the capacitive DAC and second order feed-forward integrator are shown in Figure 3.18. This configuration also opts to scale the voltage reference for the LSB in order to reduce the total number of capacitor required. As the capacitor array is implemented using CMIM devices the 7 bit differential structure with a split capacitor for $M = 3$ will grantee 10.1b for a confidence interval of 3σ using Equation 3.13 and process documentation parameters that show a $8 \times 8\,\mu m$ has $0.23\,\%$ mismatch induced standard deviation. The reasoning for this configuration is that we are guaranteed $> 9.5\,bits$ without calibration and will allow $> 12\,bits$



(a) Split capacitor array with sigma delta feedback.

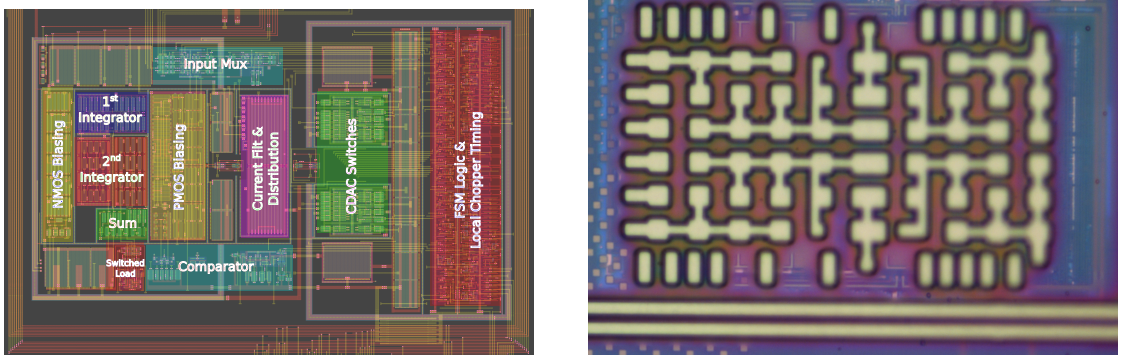
(b) Switched second order feed-forward integrator structure. Transistor sizing is detailed in Figure C.8.

Figure 3.19: Schematic implementation of the ΔΣSAR structure.

with calibration. For either case the accuracy is sufficient for recording LFP and EAP signals simultaneously. This result was also confirmed with monte-carlo analysis using foundry supplied PSP models.

The integrator topology primarily deals with the contrasting bandwidth requirement of the SAR operation and the sigma delta integration for the first stage. Particularly when taking the SAR decisions at the oversampled clock the first stage can only provide wideband gain if the capacitor is switched out and a resistive element is used instead. The circuit complexity can be dramatically reduced by using triode region transistors that regulate the PMOS biasing current for a well defined common mode. Because these transistor can be large in area they could slow down the maximum SAR speed. To avoid this the CMFB circuit is semi open loop during the SAR quantization leading to an increase bandwidth by using the common mode voltage that preserved on the integration capacitor. Also by switching the biasing current of the analogue summing stage a constant common mode can be presented to the comparator input thereby reducing any off-set disparity between the two operation phases.

Figure 3.20 shows the fabricated structure of the ADC. Since the capacitors are placed on top of the active circuits this floor plan distances the integrators and the MSB capacitors to physically isolate the digital switching noise sources. A number of shielding structures are employed to improve post layout performance. There include various guard rings and isolating N-wells but due to the proximity of the digital switching the most effective strategy is appropriately



(a) Transistor layout floor plan of fabricated device

(b) Chip micrograph showing the configuration of the capacitive DAC.

Figure 3.20: Physical implementation of ADC using a 6-metal 0.18 μm CMOS process measuring $93 \times 147\,\mu m^2$ in size.

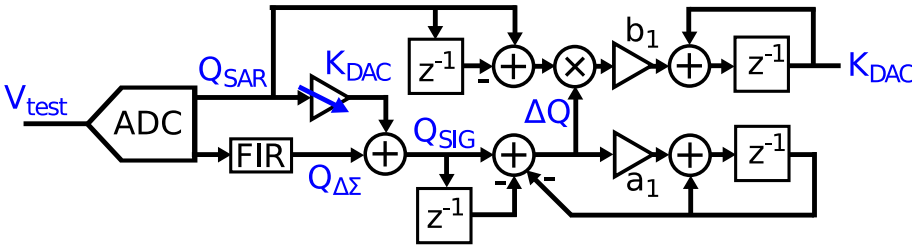
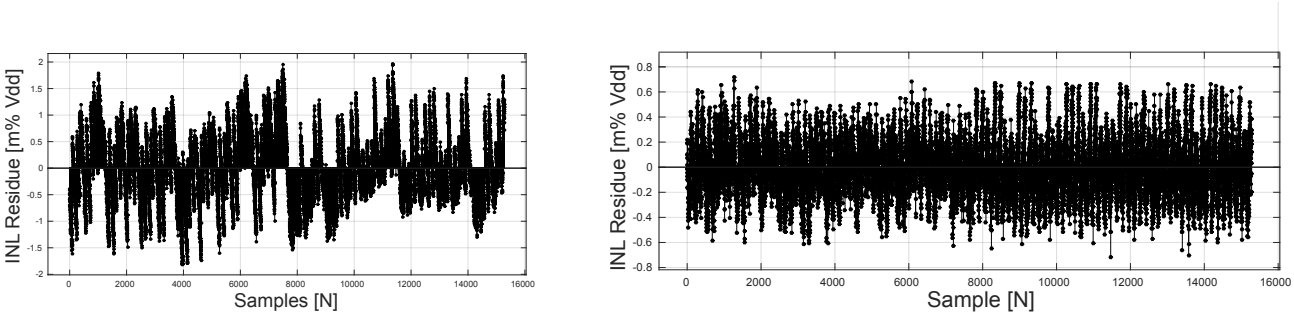


Figure 3.21: Control loop used to perform calibration with a slow test signal at the ADC input.

orienting fully differential structures in order to equalize the coupling components. Here metal layers 1-3 are used for transistor interconnect, layers 5-6 for the capacitive DAC, and layer 4 is interposed in order to shield the two sections while connected to the common mode voltage. This is because the transient fluctuations on V_{cm} are only due to mismatch and should be the most quiet reference in the system with large capacitive loading.

In order to take advantage of this structure we reveal two distinguishing characteristics that can not be found in either conventional topologies or other hybrid topologies. When the capacitive DAC is considered as a set of weights that need to be determined we realize that the derivative for slow varying signals is predominantly quantized by the sigma delta loop. With the exception when the SAR bits switch the quantization is independent of the mismatch in these weights. As a result all the mismatch coefficients can be accounted for with respect to the $\Sigma\Delta C$ capacitor.

The calibration technique discussed is abstractly represented by Figure 3.21 where there are two IIR control loops with the coefficients a_1 and b_1. In part this loop performs normal operation



(a) Uncalibrated INL from a slow ramp input

(b) Calibrated INL from a slow ramp input

Figure 3.22: INL Plots illustrating the mismatch artefact reduction due to calibration.

by evaluating the signal quantization Q_{sig}. This is done adding the SAR quantization with calibrated weights and decimating the oversampled residue with a 32^{nd} order FIR window quantized with 8 bit coefficients for each sample. Here a_1 simply has to be small enough to track the signal and reject noisy components to determine ΔQ. ΔQ represents DNL nonlinearities that are used to adjust the coefficients K_{DAC}. The multiplication operator is in fact a bitwise evaluation that indicates if a coefficient needs to be adjusted due to a correlation between ΔQ and a change in that bit. Hence b_1 needs to be small enough to prevent level dependent tuning and V_{test} should be a full range slow varying signal.

The improvement in INL is evident in Figure 3.22 due to the calibration mechanism with $a_1 = 1/4$ and $b_1 = 2^{-8}$. The close interaction between INL & DNL errors over the full dynamic range for a capacitive array in addition to the sigma delta loop's capability of quantizing $\pm 2\,LSB$ of the array allows this method to converge accurately. Here it is observed that the calibration improves the quantization accuracy by two additional bits.

Figure 3.24 shows the test bench used during device characterization. The saleae logic device is a digital probe that offers 100 MS/s digital signal acquisition for measurements of up to 10 seconds. Here the raspberry pi module simply provides real time interaction with the device configuration using automated spi control and a graphical user interface that will indicate ADC precision based on the selected operation. This allows us to tweak the operating conditions and find which noise sources are disturbing the configuration. The analogue bias I_{BIAS} is generated by a 2602A Keithley system source meter and fed in using a guarded triax cable. The differential input signals are generated using a Agilent 33522A arbitrary waveform generator and fed to

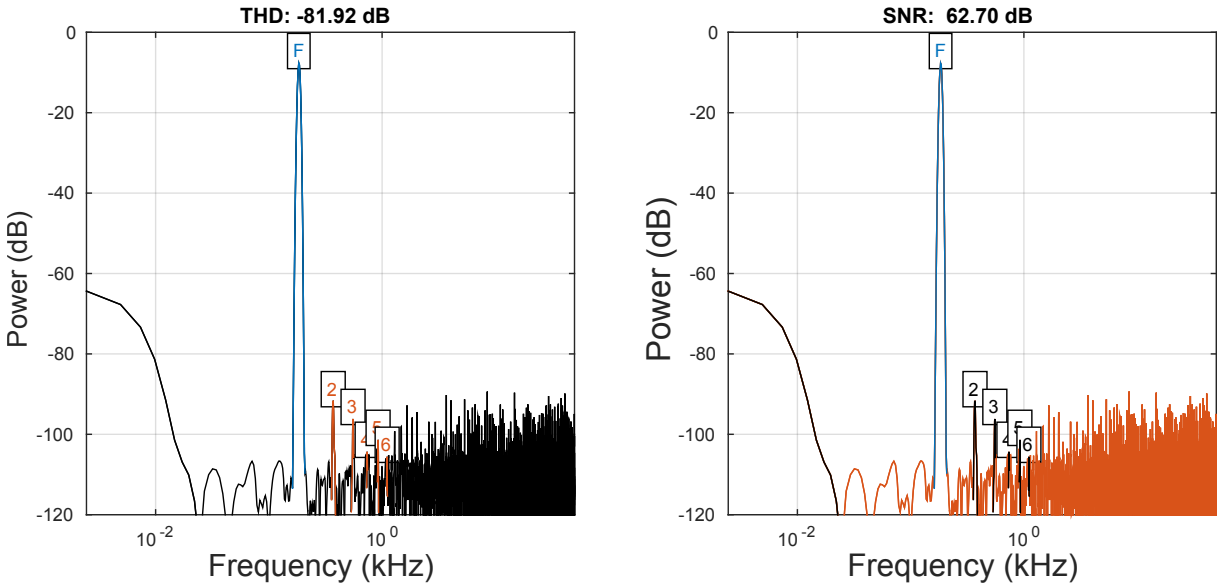
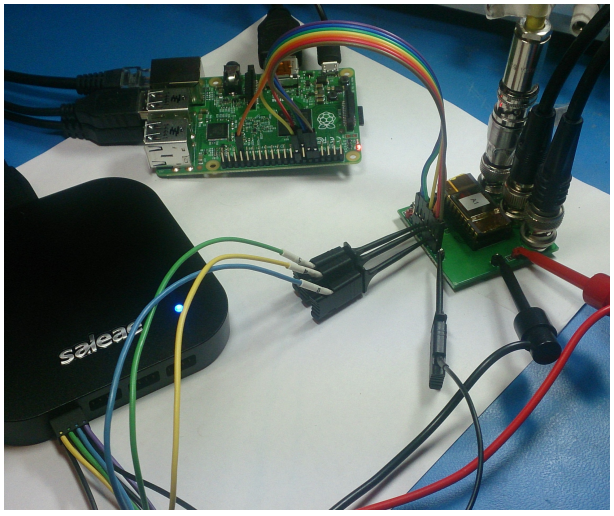
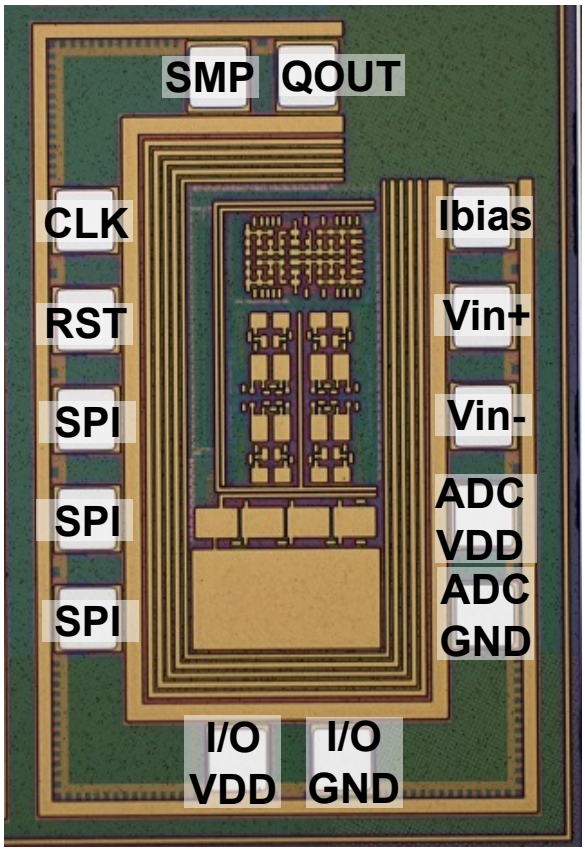


Figure 3.23: Measured THD and SNR of the fabricated data converter.



(a) Photograph of the mixed signal test bench.



(b) Chip photograph of the device under test.

Figure 3.24: Testing setup used for characterizing the ADC.

Table 3.4: Summary of performance specifications for the ΔΣSAR data converter and other oversampling/noise shaping data converter structures found in the literature.

Parameter	Units	This Work	Lo [138]	Roermund [139]
Technology	[nm]	180	65	65
Supply Voltage	[V]	1.2	1.2	0.8
Total Current	[μA]	12	13	1.7
Sampling Frequency	[kS/s]	200	8	16
ENOB	[bits]	11.3	17.5	14.5
SFDR	[dB]	86	105	87
Area	[μm^2]	93×147	400×180	600×300
Power-Area-Product	[$\mu W\ \mu m^2$]	$1.9 \cdot 10^5$	$1.1 \cdot 10^6$	$2.4 \cdot 10^5$
$P/(fs + 2^N)$	[fJ/conv]	10	29	6.6
$SNDR + 10log(BW/P)$	[dB]	166	180	177

the ADC input using BNC cables.

Table 3.4 outlines the characteristics of the implemented ADC configuration while comparing it to recent oversampling/noise shaping data converter publications. Figure 3.23 demonstrates the spectral characteristics of the quantization output for a input signal at half the full input range after calibration. In comparison to the analogue instrumentation, the resource related specifications are significantly larger. However note that there requirements are distributed over a number of channels as a result of multiplexing this structure.

The trade off with respect to residue over sampling in Figure 3.25 demonstrates that there is some flexibility with respect to sampling rate and SINAD performance. In addition this also clarifies that post-fabrication adjustments do not exhibit significant resolution improvements beyond the design point. This is related to the sampling noise of the capacitor array and the noise floor of the analogue integrators that need to be programmable for different oversampling ratios. At which point the decimation also has more strenuous requirements that may result in an inefficient resource overhead. Strictly stated it is significantly more efficient to reject noise with digital bandpass filtering selected frequency components than having the ADC resolve the signal beyond the target precision.

In the context of miniaturization the topology presented here follows closely to the expected improvement from the model for high resolution signal acquisition. We achieve nearly 12 bits of

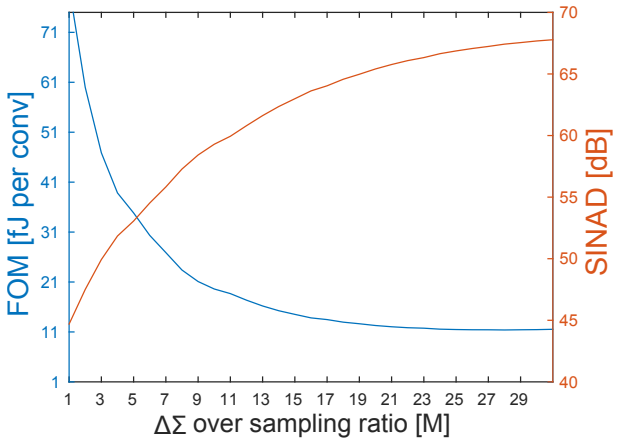


Figure 3.25: Measured Figure of Merit as a function of oversampling ratio.

quantization with a 6 bit equivalent capacitive DAC which is reflected in the compact design foot print. When compared to similar compact ADC implementations found in recent publications we observe a competitive power budget with again significantly smaller area requirement. Some additional digital processing is required opposed to the simplicity of SAR converters to take full advantage of the topology. However such hardware is typically readily available in systems that also perform spike sorting and neural signal classification.

3.4 System Level Abstraction

Numerous specifications such as ADC resolution and input referred noise of the instrumentation amplifiers relate directly to signal specific parameters. Moreover a particular processing algorithm would favour certain filter configurations of others in terms of signal conditioning. In multi stage systems however there is a significant amount of flexibility related to choosing gain for individual stages or their filter parameters that is indifferent to the resulting transfer function. Here we consider such a primitive N stage analogue processing chain and discuss the allocation of resources to gain insight to some of the high level the optimization for selecting a specific configuration. Such a configuration is shown in Figure 3.26.

$$G[n] = A_g\left(\beta + \alpha^n\right) \qquad \text{where} \qquad A_g = \sqrt[N]{\frac{G_T}{\prod_{i=1}^{N}(\beta + \alpha^n)}} \tag{3.22}$$

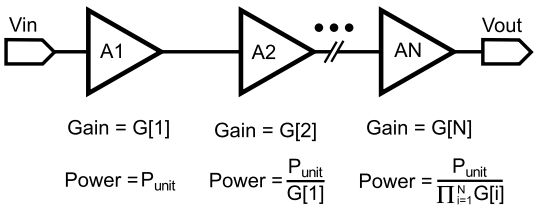


Figure 3.26: Multistage amplifier configuration using the series G to adjust the allocation power and area.

Consider a geometric series for the gain of each stage as expressed in Equation 3.22. Here G_T, α, β represent the total gain required, resource distribution factor, and a minimal contribution factor. The formulation is motivated by the fact that if α is one resources are allocated equally. This means every stage has equal gain but it also implies that the sum of all gain factors is minimal leading to a minimum amount of area due to the feedback capacitors. More typically designs will choose a smaller α such that most of the gain is situated at the first few stages. This allows some reduction in power in the proceeding stages because of the reduced noise requirement. β simply allows us to specify that a fraction of the total gain is uniformly distributed but is typically kept small in order to maximize the benefit from resource redistribution. This allows us to express the noise power requirement for a given set of parameters in Equation 3.23.

$$P_{Amplifiers} = P_{unit}\left(1+\sum_{k=1}^{N-1}\left[\prod_{i=1}^{k}\frac{1}{A_g\,\beta + A_g\,\alpha^i}\right]\right)^2 \tag{3.23}$$

$$A_{Gain} = A_{unit}\left(\sum_{k=1}^{N}\left[1 + A_g\,\beta + A_g\,\alpha^k\right]\right) \tag{3.24}$$

Here P_{unit} is simply evaluated from Equation 3.3 and leads to an area requirement that is simply expressed using Equation 3.24. Now taking some typical parameters we can evaluate a possible configuration of gains and thereby the associated allocation of resources. This is shown in Figure 3.27.

Lets take A_{unit} as some unit capacitance size that allows the deviation of gain due to mismatch to fall inside the confidence interval. In order to realize Equation 3.23, each stage has its power and input referred noise reduced by accumulated gain for the preceding stages. This

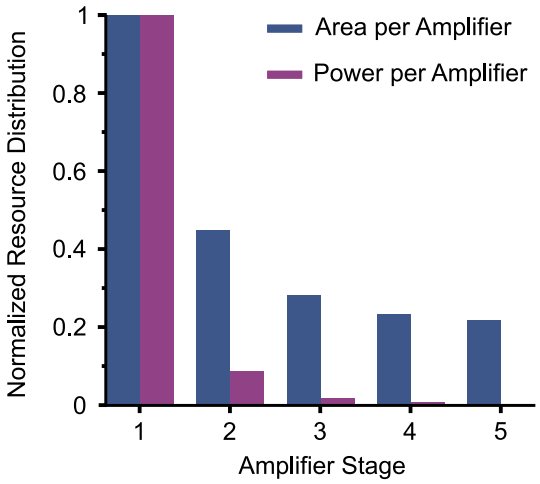


Figure 3.27: Resource allocation for analogue power and area using the parameters $G_T = 500$, $\alpha = 0.3$, and $\beta = 0.05$.

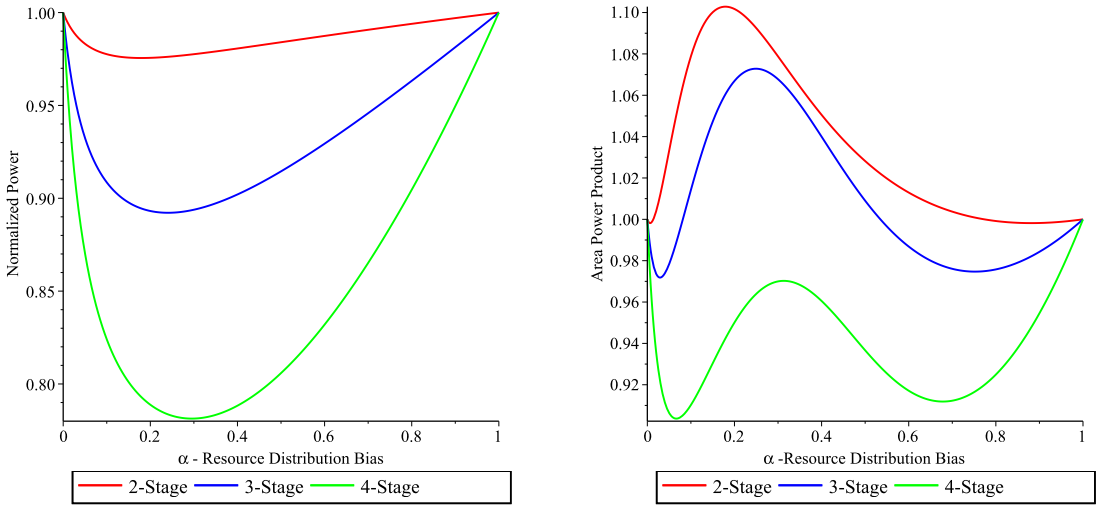
result presents us with the trend illustrated in Figure 3.28 where it appears that in many stage systems it is relatively beneficial to redistribute the resources to the front-end for a reduction in overall power. However when the number of stages is three or less we observe the increase in area can diminish this improvement for high gain system requirements.

So far we have neglected some aspects to the design consideration. The first is the multiplicative increase standard deviation as N is increases and the sensitivity to variance being inversely proportional to closed loop gain. Here we can account for the increased variance by proportionally increasing A_{unit} in order to neutralize this increase according to Equation 3.25.

$$\Delta\,\sigma^2 = \frac{A_{\mu+\sigma}}{A_{Gain}} \approx \prod_{i=1}^{k} \left(1 + \sigma\, CI \left[2 - \frac{2}{\sqrt{A_g\,\beta + A_g\,\alpha^i}}\right]\right) \tag{3.25}$$

Again σ represents the deviation for a chosen unit capacitance and CI is our confidence interval. For completeness in estimating area we will also introduce the capacitance required for performing filtering on the last K stages. Rearranging Equation 3.3 in terms of output referred noise according to Equation 3.26.

$$e_{out}^2 = \frac{kT}{C}\, NEF^2\eta \tag{3.26}$$



(a) Expected power reduction for different topologies as resources are biased to the front-end of the analogue chain

(b) Resource distribution impact on power-area product for different topologies.

Figure 3.28: Normalized resource improvements for α with respect the case when $\alpha = 1$ for each configuration.

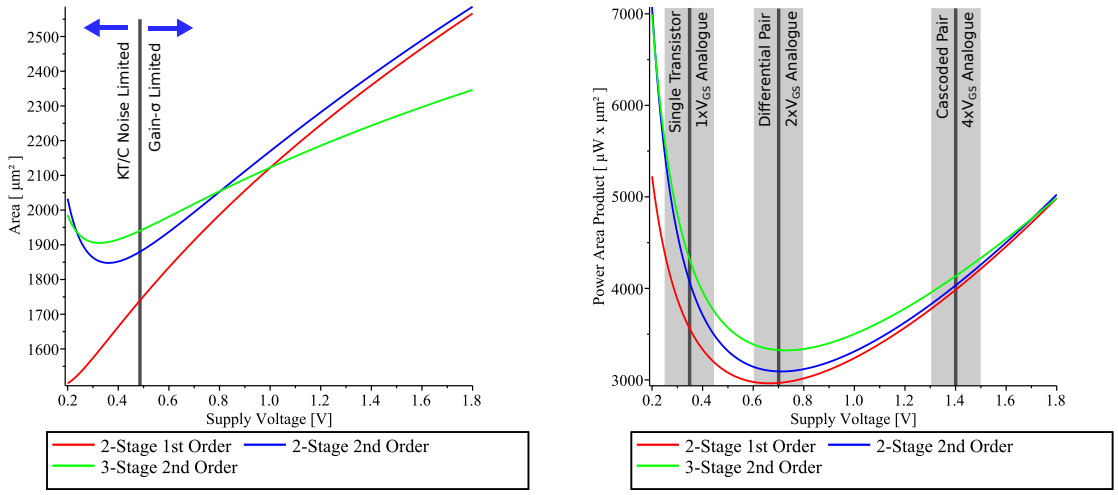
Combining these terms lets us define a more accurate area requirement that is reformulated in Equation 3.27.

$$A_{filt} = A_{unit} \cdot \frac{kT}{C_{unit}} \frac{NEF^2\, SNR^2}{Vdd^2\, \eta} \cdot \left(1 + \sum_{k=1}^{K-1} \prod_{i=1}^{k} \left[A_g\, \beta + A_g\, \alpha^{N-i}\right]\right) \tag{3.27}$$

It is important to point out that SNR here refers to the SNR of the data converter as we have fixed the input referred noise of the system for a systematic comparison and we adjust G_T to fill this dynamic range. And extending this result with the requirements for signal conversion we can estimate system level power P_{Total} and area A_{Total} requirements as a sum of individual components according to Equation 3.28.

$$A_{Total} = A_{filt} + A_{Gain} + A_{ADC} \qquad \text{and} \qquad P_{Total} = P_{Amplifiers} + P_{ADC} \tag{3.28}$$

Taking an appropriate set of parameter values, the system of relations is exemplified in Figure 3.29 with respect to the dependency on the supply voltage, Vdd. As illustrated there are two domains when considering the area requirement. For small Vdd the sampling & filtering



(a) Expected area requirement for different topologies with respect to Vdd

(b) Vdd dependency of the area-power product for different topologies

Figure 3.29: Analogue resource relations with respect to different supply voltages.

noise requirements overwhelm the design particularly in this case if α is not taken small enough and a second order roll off is needed. When there is more voltage overhead available we observe reliably matching in input dynamic range of the ADC is the dominating factor.

The area power product also tells an interesting story. When Vdd is larger than 1 V a clear proportional dependency on power is apparent that is mostly related to the total gain & noise requirements of the system because the ADC is not the limiting factor. However for small supply voltage the power dissipation requirement is more closely related to the lower noise quantization requirements presented by the SD-SAR topology. We should be careful because certain circuit topologies are simply not viable below specific supply voltages and as a result it would no be possible to achieve a NEF smaller than 2. Figure 3.29 also indicates when particular topologies are viable specific to the $0.18\,\mu m$ CMOS process where $V_{th} \approx 350\,mV$. That said it is likely a system can be designed with $0.6\,V$ supply in order to achieve significant power and area savings. The main challenge will be achieving acceptable total harmonic distortion as the supply will not easily allow cascoding transistors. Particularly sub-threshold transistors suffer from Gm nonlinearity as a function of $e^{\frac{-V_{DS}}{U_T}}$ that can only be compensated by increased loop gain and multi-stage topologies. Since it is implementation dependent, it is difficult to quantify what this increase in area an power overhead this will result in. We can assert that $60\,dB$ precision with instrumentation has very significant diminishing returns when the conventional design

approaches a $2\,V_{th}$ supply. The reader can find more details in regard to these comparisons in Section C.2.

The approach taken here can be exhaustively extended towards including more detail in the system level design in order to leverage the capability of numerical methods. Higher order Gm-C filtering structures can be accounted for as a single stage by introducing new parameters that reflect the increase in NEF and filtering capacitors. Transistor area per amplifier can arguably be assumed static if chopping techniques are employed or alternatively this can accounted for by considering the flicker noise relations for the input transistors. However these contributions have negligible effect on changing the optimal resource destitution and will be more influenced by strategic positioning of poles to reject certain noise components. The most critical parameters on the systems level is the supply voltage as well as the requirement for channel to channel gain matching. As the power area product has a inverse square dependency as either V_{DD} or gain variance tends to zero. There are only a select number of scenarios where gain matching is of significance which is primarily in the case of distributed LFP recording and multi electrode (i.e. tetrode) recordings where exact coupling of neural circuitry is in question. The supply voltage has significance with respect to the expected power dissipation of the on chip digital processing and it is understandably advantageous to aggressively dissipate more power on the analogue side if the power saving in the digital domain indicate a overall improvement.

We note another aspect to technology selection in addition enabling voltage scaling is the increase in functional capacitor density. In fact we have shown that the dominant factor for area requirement in chopper stabilized structures is capacitance through the strong dependency on gain and filtering elements. More advanced processes have an increased number of metal layers and higher transistor gate capacitance. This ultimately leads to an increased capacitor density per square millimetre. In certain scenarios this should allow us to marginally shrink amplifier configurations while keeping the same filter characteristics. The main concern would be associated with capacitor nonlinearity that requires extra consideration or correction circuits.

3.5 Conclusion

This chapter has demonstrated the capacity for conventional analogue instrumentation with state-of-the-art circuit techniques. This presents capacity for achieving very compact performance that is sufficient for the full characterization of neural recordings. The fabricated system uses $0.03\,\mathrm{mm}^2$ size silicon footprint for 4 recording channels that can characterize 5 mVpp neural signals with over 11 bits of precision. In addition proposed $\Delta\Sigma SAR$ ADC topology demonstrates how oversampling converters can achieve 10fJ/conversion efficiency with minimal circuit complexity. The techniques applied here suggests chopping and sigma-delta modulation are key components for achieving better performance particularly for size constrained systems. In association we suggest immediate digitization & coherent mixed signal processing to leverage a number of advantages. Moreover we expect modern system will allow more processing capabilities in the digital baseband for BMI systems that needs to be used effectively.

The significance of minimizing the noise efficiency factor has been revealed in terms of having profound influence to power dissipation and area. In extension we have presented a number of topologies that excel at achieving excellent power and area efficiency in the case of single stage, two stage, and ADC structures. However we are left with little surprise when methodical optimization of various configurations is limited by the fundamental bounds in terms of noise and dynamic range. In fact various idealized configuration show little benefit with respect to one another if they have been optimized and exploited appropriately with the understanding presented. It is characteristic that improving resource efficiency for full bandwidth signal quantization is difficult because we simultaniously attempt to achieve lower supply voltages.

Although digitization is crucial to most neural recording systems for extracting the signal characteristics used to train and improve signal postprocessing. It is clear that improvements at the system level will lie very much in the domain of specialized instrumentation and analogue to information converters. This notion is motivated by the desire for the system to be limited by the law of equipartition and less so by the quantization process of the data converter. The direct classification of recordings in the analogue domain has significant implications on the responsibilities of the accompanied DSP on chip and the reduction of associated processing bandwidth.

Chapter 4

Neuron Processor Interface (NPI)

A core aspect of emerging neuroscience is quintessentially performing real-time data analysis at a massive scale. However when we observe its manifestation in state-of-the-art neural interfaces we find the hardware is very limited to specific methods that can be objectively short-sighted. This chapter aims to direct our attention to a different point of view with respect to how these sensor systems can be structured. In particular we are guided by the concept where an implant is capable of performing software defined instrumentation. This is associated with a focus that lies with enabling real time & in-vivo testing of a more diverse set of signal characterization methods. More importantly we will demonstrate that this can be made feasible for large scale distributed systems.

This particular approach is motivated by a number of factors that aim to increase performance and enable research opportunities. The first is that many aspects with regard to the signal quality of an implant cannot be predicted beforehand. As a result implementing a specific algorithm for specific signal characteristics may lead to failure or an overly conservative design because the environment can potentially be excessively noisy. By introducing the capacity to dynamically execute different processing methods on neural data the implanted system to use either LFP and EAP activity in real-time. This may be a significant element to improving the success for chronic BMI implants. Moreover the prolific development in characterization methods used for decoding neural data inhibits a general consensus for DSP techniques. This prevents a

single method and corresponding architecture to be applicable in most scenarios. The second factor is that this approach conceptually enables the development of real-time resource constrained algorithms which are virtues often neglected when working with data sets. Currently most BMI development platforms have limited capabilities to allow algorithms to use external or multi-modal features to inform local operation and simultaneously provide recordings from hundreds of electrodes. This construction may be a key factor to allowing high level algorithms to directly manipulate machine learning parameters local to each implant. This hierarchical fashion should improve the efficiency of distributed BMIs for decoding information. In contrast we question the feasibility for scaling the current supervised methods that require fine measurements of each electrode's recording to approach a optimal decoding strategy. Typically the computational efficiency of this approach remains exhaustive when reconfiguring sensor parameters because it use a centralized unit that recalibrates all recording channels in an elaborate fashion.

This chapter is organized as follows; Section 4.1 motivates localized processing for increased efficiency and estimates to what extent we can perform on-chip processing. This is followed by Section 4.1.1 where typical methods used for neural signal analysis are introduced and the respective hardware complexity is demonstrated in Section 4.2. This leads into the proposed distributed processing architecture in Section 4.3 where the design is discussed with respect to the implementation. Section 4.3.3 demonstrates the realization of this platform. Finally Section 4.4 draws conclusions with respect to the digital approach to neural instrumentation.

4.1 Processing at the interface

Ideally a neural interface device is tasked with recording from a large ensemble of electrodes and transmitting information with the lowest bandwidth because the harvested power is a scarce resource for implants inside the body. However over the course of an implant's life time most signal characteristics are dynamically changing which implies that there should be an involved

learning process that similarly adapts to these changes. This can also mean that the output bandwidth is constrained by the total amount of mutual information that can be retained within the device. Such a device will predict the expected recording from one time interval to the next and differentiate any new information that needs to be transmitted. Hence we should be convinced that the processing capacity or complexity for a closed or memory limited system should reflect in its fundamental ability to store information [140].

In order to capture some high level trends with respect to processing requirements let us normalise memory capacity in terms of state variables that is independent of modality. This is particularly useful because the number of state variables in a dynamic process is a good indicator for complexity whether is a digital classifier or an analogue filter. Here we will exclusively focus on processing by assuming the signal being operated on is idealized with respect amplitude and its representation. This extends from our analysis in Section3.4 by elaborating specifically on comparing digital and analogue resource allocation associated with processing.

$$R_A = \underbrace{\frac{2\pi\, BW\, kT\, SNR^2\, U_T}{V_{DD}\, L}}_{power} \cdot \underbrace{\left(\frac{A_{min}}{L} + \frac{kT\, SNR^2}{L\, Vdd^2\, C_{dens}}\right)}_{Area} \tag{4.1}$$

If we represent the resource required as the power area product for a state variable then in the analogue domain it would be represented by Equation 4.1. Here BW, V_{DD}, C_{dense}, A_{min} reflect the signal bandwidth, supply voltage, capacitor density, and typical transconductance area overhead for a particular technology respectively. L is a normalized feature size that allows us to evaluate parameters for a particular technology and extrapolate them based on constant field scaling factors.

$$R_D = \underbrace{2BW\, \alpha\, \log_2(SNR)\, C_{gate}\, V_{dd}^2\, L^2}_{power} \cdot \underbrace{\alpha \log_2(SNR) A_{gate} L^2}_{Area} \tag{4.2}$$

Similarly Equation 4.2 represents the power area product for a digital state variable. C_{gate}, A_{gate}, α parametrise typical gate capacitance, area, and overhead for each register respectively. Generally the dependency of both parameters R_A and R_D are well understood and guide max-

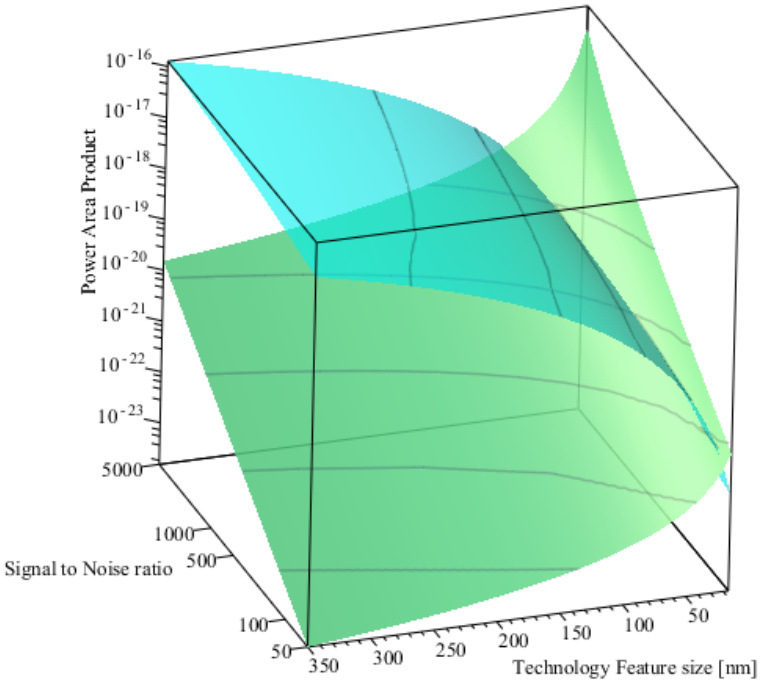


Figure 4.1: Impact of technology on R_A analogue (green) and R_D digital (blue) processing resource requirements extrapolated from a 180 nm CMOS technology under constant field scaling.

imizing system efficiency in an abstract sense [141].

For neural instrumentation however both power and area requirements must be highly constrained in order to realize a device that can accommodate a large number of recording channels and remain implantable. Figure 4.1 shows the resource requirements for processing in the analogue and digital domain with respect to signal fidelity and CMOS technology. Either approach can present an advantage over the other under specific conditions. Digital systems appear favourable beyond 65 nm CMOS where analogue will do better at lower SNR conditions given a technology with a larger feature size. The discussion in Section 3.4 suggested the analogue preconditioning requires a resource allocation of $10^{-15}\,Wm^2$ with a weak dependence on technology. Moreover if quantization is not considered then power can be entirely determined by the noise specification and the area requirements are dependent on the gain configuration. Comparing this figure with the estimate on R_D indicates that we should be able to integrate

a considerable amount of processing capabilities before the DSP uses a comparable amount of resources. This is important because improving on-chip processing capacity ideally results in requiring less supervision and a lower wireless communication bandwidth.

While we may expect other sources of over-head and extra power dissipation components, we should instead take a moment to consider the implications of this result. Particularly when considering the claim that electronic sensing of brain activity on larger scales is not viable due to the excessive data rates derived from the principle entropy relations and the associated communication bandwidth [142]. Clearly any degree of on-chip processing undermines this limitation because it enables us to achieve data rates far lower than that of the electrical signals by finding a more appropriate basis of encoding information. On the other hand it does raise an important point regarding the relationship of the generated output data rate and the recorded signal to noise ratio. In some sense we are simply faced with the challenge of best consolidating the recorded information towards high level indicators for specific objective functions. This allows us to approach data rates extracted from current BMI studies which are negligible in comparison to the Nyquist rates. In this light we argue that electronic methods for recording activity are the closest to realizing a viable neuroprosthetic solution in the near future when comparing optical, magnetic and less invasive BMI architectures.

Now we can make the assertion that there should be two approaches to solving the system level challenge of integrating wireless neural instrumentation systems. The first approach would be a mixed signal topology that extensively uses analogue processing such that technology has a weak impact on improving efficiency. Instead the critical component lies with the effectiveness of analogue dimensionality reduction. In such a case we need to adopt a well established algorithm that can accommodate analogue variability an still deliver exceptional signal characterization. The second approach is to rely on digital methods that deliver robust and reconfigurable compute resources that scale well with technology. This should grantee the capacity for a variety of fully adaptive algorithms capable of extracting multi-modal characteristics from recordings. This could much more valuable for experimental neuroscience at this point in time. Unfortunately not all forms of algorithms can use the low power characteristics of processing in the analogue domain. Moreover they are typically limited by underlying assumptions regarding

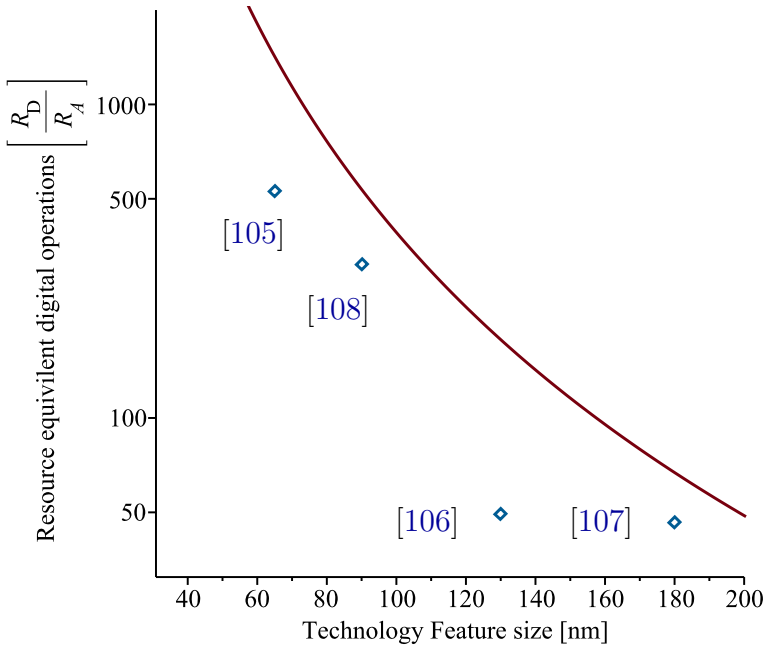


Figure 4.2: Analytic number of digital operations available with respect to different technologies (red) with references to the normalized performance of image processors (blue).

noisy perturbations. When we introduce different contexts of operation reconfiguring the analogue is not done as arbitrarily as a digital structure would. For this reason we will adopt the digital approach in order to leverage robust reconfigurable capabilities and reconsider the analogue approach in Section 5.2.

The significance here is that these trends allow us to roughly estimate the complexity of algorithms for different technologies if their resource requirements made to be equivalent to that of the instrumentation circuits. We show in Figure 4.2 that using a $0.18\,\mu m$ CMOS process should give way to approximately 100 state variables or equivalently perform about 100 operations per sample taken. In fact looking at image processors that similarly rely extensively on data intensive post processing we can see an identical dependency on technology scaling as we have predicted for various levels of digital performance at different technology nodes. It is important to note that the normalized efficiency evaluated here is in fact independent of signal bandwidth and only depends on signal to noise ratio and its relation to the supply voltage.

$$P_{system} = \underbrace{N_{channel} \cdot (P_{Algo} + P_{Transmit})}_{In\ Channel} + \underbrace{P_{Control} + P_{Comms}}_{System\ Level} \tag{4.3}$$

System level design of a embedded processing system for BMIs should be guided entirely by the optimization of compute power efficiency. As shown by Equation 4.3 we expect are two primary components in the system power breakdown. The over all objective should lie with minimizing the channel level power dissipation of the algorithm P_{algo} by increasing that of the system level control $P_{Control}$ such that the component that scales with channel count is reduced. Secondly we should keep in mind that reducing in processed output data alleviates the dissipation of the on-chip communication $P_{Transmit}$ and the external telemetry power consumption P_{Comms}.

4.1.1 Methods for Neural Signals

A key component to developing this platform is a discussion on the diverse set of signal processing methods performed on neural data and their computational requirements in order to determine our system's specifications. More importantly we want to judge what is the expected complexity for some of these operators and how many variables are allocated during each process. There are in principle four different categories for the methods that are applied to process neural signals which are listed below. In practice, a single integrated system will utilize a multitude of different techniques to achieve denoising and feature extraction.

Pre-Processing is the filtering and conditioning of each ADC output sample using FIR, IIR, or non-linear filters. Here the objective is simply to de-noise the features or signal components. The resulting signal allows for better detection or more precise evaluation of the signal characteristics and is often closely related to the characteristics of the instrumentation circuits. This output may be considered as the raw signal recording that is used to bench mark any post processing methods or other instrumentation systems.

Detection is associated particularly with capturing the intermittent spike events. There is no quantitative evaluation made with regard to the nature of the detected spike but any detection may trigger the process that records contiguous samples around the detection event that are then subsequently characterized. These events are commonly triggered upon simple threshold

crossings of the signal or its integral power over several samples. In some systems the interest lies only with the accurate detection of spike events which is sufficient to perform closed loop therapeutic treatment or control actuators external to the body. **Data reduction** can be interpreted either as representing a spike waveform in terms of defining features or with a reduced basis to allow approximate reconstruction by using spike amplitude or wavelet decomposition. Representing waveforms in terms of their quintessential components allows for more efficient post processing and reduces data rates in the case of wireless telecommunication. In its primitive sense this is simply dimensionality reduction of the recorded data and is usually followed by supervised or unsupervised signal classification.

Classification is the predominant objective for BMIs and is the main difficulty to realize inside a embedded recording system. Such a task in spike based systems primarily performs a generalization of the detected spike shape in terms the previously detected neurons. This reflects the fact that in most cases multiple neurons can be detected by a single electrode and by distinguishing these events the integrity of information can be preserved. The objective here lies with having an equivalent spike event output as simple detection to perform actuation but with better fidelity.

It should also be clear that the above operations primarily focus on reducing the recorded signal to its primitive components in terms of spike events at the rate of $100\,b/s$ instead of the $256\,Kb/s$ data stream typically generated by the ADC. The processing layer on top of this elementary function will either aim to evaluate neural connectivity or use collections spike rates to perform inference of high-level dynamics. This application specific processing of these systems will not be considered here primarily because the nature of such a problem is very different from the more generic information extraction from recordings. As a result that system architecture should revolve specifically around multichannel trained dimensionality reduction. Even when such a task can adjoin to what is presented here it will be out side the scope of this discussion which targets more generic signal instrumentation.

When we survey the various algorithms found in recent literature and estimate the expected memory/computational requirements we might observe a distribution like that shown in Fig-

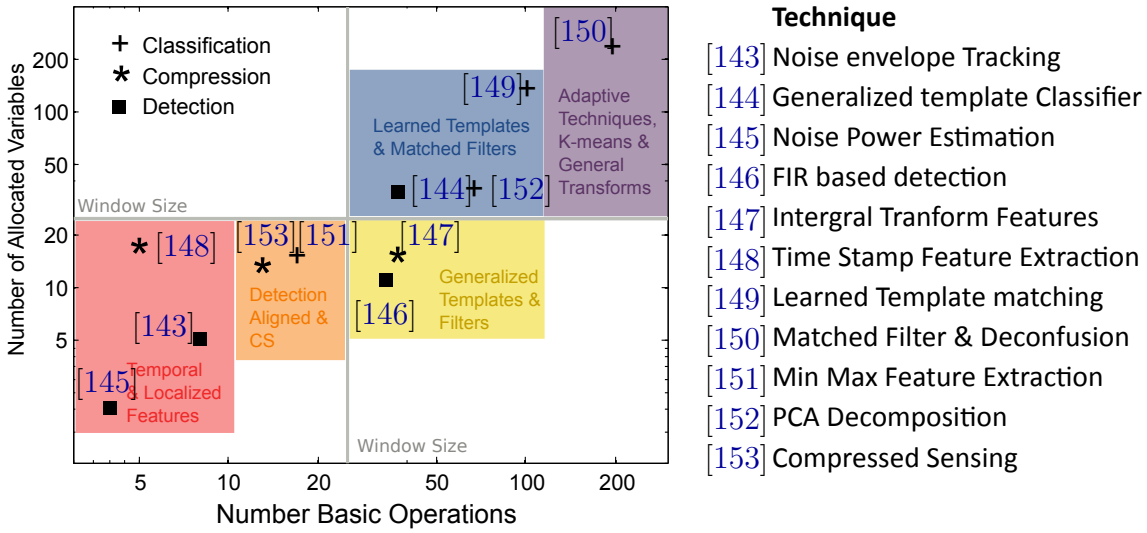


Figure 4.3: Estimated resource requirements for different classes of algorithms use for processing neural recordings found in literature.

ure 4.3. For fair comparison we have adopted these methods to operate on a window size of 32 samples with three types of dectected spike waveforms when applicable and only accounted for memory allocation that cannot be shared across channels. This should give a good normalized indication of the limiting components for each method which we could then further optimize in a more specialized manner. Notice that there is a strong correlation in the memory usage and the required number of operations for most embedded systems.

Some of the most efficient spike detection and feature extraction is associated with using temporal characteristics of the spiking waveforms. Common examples include using the time interval between minimum and maximum peaks or the duration of threshold crossing for detected spikes. The defining characteristic here is that alignment buffers are not needed which leads to using very few operations per sample. Inherently the drawback is the increased sensitivity to noise which implies a very limited capacity to distinguish and classify different spike shapes. Unless more filtering is performed. There are other methods that also operate with reduced signal buffers. For instance using compressed sensing where signals a continuously re-projected with a sensing matrix. This requires a few accumulators for each coefficient being extracted but this is strictly data compression. Ultimately This may not help directly with classification or signal characterization that must now be performed off-chip before any benefit can be realized.

Many other classes of algorithms operate on a windowed basis that exploit learned mean spike shapes that are expected in the recording. Here a convolution or distance operator will indicate which class of spike is detected. For terminology let each convolution of the signal result in a feature that is used for detection and/or classification. The adaptive component of these methods leverage a significant amount of noise shaping and separation depending on what the objective function entails when the basis for convolution is being determined. The most prevalent approach is convolving with principle components which simply maximizes the signal variance in the projected space. In contrast to using temporal feature that struggle with sample limited denoising, the windowed operators should be more robust towards noise. Instead there is some difficulty in systematic alignment of the window with the spike. This aspect often motivates increased sampling rates or interpolation in order to perform accurate alignment.

The objective function for determining the convolution kernels can be oriented towards maximising sparsity [154], signal to noise [155], cluster separation [49], or expectation maximization [156] each reflecting different signal modalities and analysis methods. Although the complexity for training can be varied extensively the local operations for classification after adaptation is almost equivalent. This operation is the F linear projections using W samples onto the feature space where F and W are the number of features used and the number of samples in the window respectively. There may be some deviation from this operation if we also consider the different confidence intervals for each class is taken into account. This can be done by evaluating centroid distances in terms of the variance for each respective centroid. Conversely to such training, generalized templates averaged over a number of recording channels may also be used for feature extraction in order to share the memory requirements at the loss of not achieving maximally separated clusters for each individual channel.

Naturally it is challenging to objectively judge feature extraction and classification methods. We could always reduce the dimensionality of the search space or by simplifying the convergence strategies to reduce memory and computational requirements. For this reason the details in Figure 4.3 should be considered in relative terms due to the generalizations made with regard to system specifications. As many methods in the literature are not performed at the sensor interface they typically will not take advantage of processing on a sample to sample basis and

opt for a batched or sequenced processing methodology.

We claim however that the segregation between methods primarily lies with whether the classification features are based on sample space characteristics or alternatively use windowed convolution operators. The former is the less rigorously justified as the features relating to amplitude and spike width have weak physiological significance but are aggressively more efficient than other methods. The characteristic requirement of the later later approach is typically related to the window size that is indirectly associated with the sampling speed in order to fit the relevant spike shape into the window. One important objective of current decoding research is related to introducing adaptive techniques that iteratively improve classification without supervision. Particularly without excessive memory requirements in order to keep track of long term statistics. To demonstrate why this can be particularly challenging, consider a simplified example of using K-means to directly cluster the sample space adaptively where we are fortunate to know the number of classes is three.

Clearly for each detected spike we would need to evaluate the distance between our data the centroids having a memory and complexity requirement of $(1 + F)\,W$ variables and $2\,F\,W$ operations respectively. Then once the class is determined an additional $2W$ operations are needed to adjust the centroid with the new data. This may include keeping track of the additional $F\,W$ truncation residues that allow using a small enough adjustment weight for convergence when our quantization is limited by an 8 bit system. Now for completion, assume our window is 32 samples and we come to the conclusion that for each recording channel we to actively need to allocate nearly $2\,K\,bits$.

Typically this primitive aspect of a generic classification algorithm is intensive enough warrant not performing it locally in contrast to the complexity of the aforementioned temporal features. It also raises the challenge of trained dimensionality reduction without supervision which exclusively relies on evaluating the covariance matrix in order to minimize the correlation of non-signal components with our new basis. The above algorithm may be representative in terms of complexity with the exception that basis pursuit relies heavily on inner products. There is some relief from the fact these optimal basis change very slowly and actively adapting

is only needed once few hours as the electrode recording changes slowly with respect to the neural activity. As will be demonstrated the feasibility of these more involved methods rely very much on the careful construction and memory allocation of the algorithm with respect to the processing architecture. Operations like k-means and PCA decomposition can be performed to a certain extent if the operations are broken down into a incremental procedures.

4.2 Resource Constrained Classification

Primarily to substantiate our expectations for performing processing at the sensor interface and evaluate where the system level requirements should lie. Ultimately such a system needs to encompass a significant variety of different application requirements. We will consider the implementation of two well known methods that process neural recording and generate classified events. This will be applied to the equivalent scenario where the proposed instrumentation front-end is used from Section 3.1 such that the digitized signal will have considerable dynamic range but lacks analogue filtering. In particular this implies the recorded signal is filtered by a first order butterworth low pass filter in addition to near-DC rejection while being sampled at $25\,KS/s$. We will address the rejection of low frequency aggressors in addition to the computational requirement of typical processing algorithms.

Many of the filter and processing considerations are guided by evaluating accuracy empirically and justified though constrained parametric optimization [157]. The particular algorithms implemented here are structured in such a way they perform specific considerations for the underlying hardware. On numerous occasions we will employ single bit accumulators as approximations to the IIR equivalent feedback structures in order to improve our effective register depth through feedback. This is a primary advantage of in-channel processing where we may exhaustively make use of recorded data without having concern for the communication of these components.

Empirical validation is demonstrated by using of synthetic data sets that are publicly available

online. This data is based on characterized extracellular recording where both background activity and spike morphologies are extracted from a human neocortex and basal ganglia. The synthesized recording was originally used to evaluate the performance of super-paramagnetic clustering with wavelet decomposition at different background noise levels in [1]. Synthetic data specifically allows the inference of the ground truth resulting in unbiased performance indicators. For fair comparison of analogue and digital techniques we additionally include low frequency content from $1 - 300\,Hz$ at $10\times$ of the largest peak to peak amplitude of the extracellular action potentials found in the recordings.

4.2.1 Spike Detection & Filtering

Arguably the most influential aspect to neural detection and classification algorithms is the signal preconditioning for systematic and accurate detection of spike events. The importance lies with the fact that detection behaviour has the significant influence on how the feature space appears when the spike is characterized with various methods. Although amplitude noise can usually be accounted for in terms of filtering. Any misalignment in the time domain due to noisy aggressors in the detection operator can up modulate low frequency components. The tendency to perform detection in the digital domain is entirely related to the instantaneous characteristic of discreet time processing which is superior to the group delay inherent to analogue implementations. Minimizing this factor will minimize additional memory for capturing any signal before the detection event.

The method proposed here tracks both the mean spike amplitude and back ground noise levels in order to assert the detection level of spike events. Motivated by using physiological characteristics to specify the underlying operation parameter k_3 is introduced to represent the relative amplitude of background activity to that the maximum spike waveform that is intended to be detected. Or in other words if we are only interested in the closest neurons to the electrode k_3 should be close to 1, otherwise if we also want to detect background activity with an amplitude at 25% of the largest spiking events k_3 should approach 4. In actuality this term should also

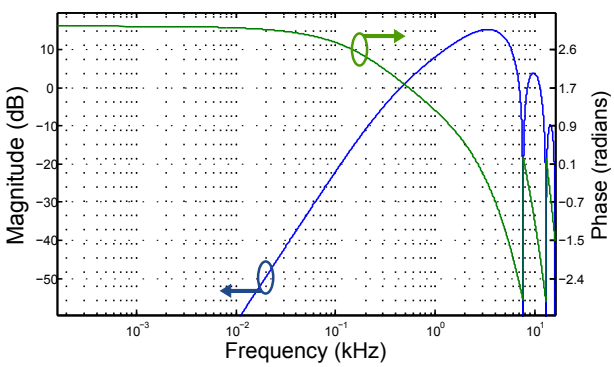
be related to how well our classification can separate noisy detection or actual spike events.

Algorithm 4.1: Spike Detection and Alignment

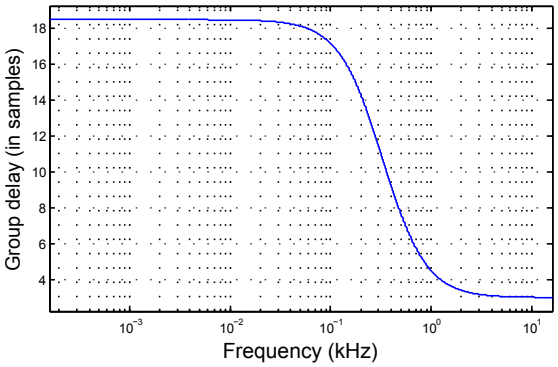
Data: Sample from ADC $X[n]$
Result: Detection events & spike window W_1

1 **begin**
2 Update V_{LFP} with $k_1 \cdot (X[n] - V_{LFP})$ // Track low frequency content
3 Set $S[n]$ with $X[n] - V_{LFP} + k_2 \cdot S[n-1]$ // IIR bandpass filter
4 Set $G[n]$ as $\sum S[n] \cdot FIR(2 * R)$ // FIR bandpass filter
5 Set $ES[n]$ as $S[n] \cdot G[n-R]$ // Energy Estimate from IIR & FIR product
6 Update V_{noise} with $k_1 \cdot (|ES[n]| - V_{noise})$ // Estimate variance on energy estimate
7 **if** $ES[n] > V_{th}$ ***and*** $ES[n] > max_{local}$ // Find threshold crossing or new local max
8 **then**
9 Update V_{th} with $k_1(ES[n-1] + 2V_{noise} - k_3 \cdot V_{th})$ // Adapt peaks and varience
10 Initiate Spike Alignment
11 Set max_{local} to $ES[n]$ // Set local maximum
12 Set *index* to 0 // Initiate data pointer
13 **else if** *Currently alligning spike (index* < 16*)* **then**
14 Set $W_1[index]$ to $G[n-10]$ // Store spike waveform with delayed samples
15 Set *index* to *index* $+ 1$ // Increment data pointer
16 **else if** *Idle state (index* > 31*)* **then**
17 Set max_{local} to 0 // Finish classification & find next local maximum
18 **else**
19 Accumulate *index* with 1 // Increment data pointer

The specifics of this operation is reflected in Alg. 4.1. Here the term Update, Set, and Accumulate represent recurrence, instantaneous, and integrated relations respectively. The state variable V_{LFP} primarily removes low frequency drift that is not associated with individual spiking events and $S[n]$ is as a result a bandpass equivalent of our sampled signal $X[n]$. The signal's instantaneous energy is represented by $ED[n]$ which is a product of $S[n]$ and the delayed derivative computed by the FIR of even order $2R$ with the coefficients $a_n = -a_{2R-n} = 1 - 2/R \cdot (n-1)$ for n from 1 to R. The factor R is in association with the ratio of sampling interval to spike polarization interval, equivalently as $R = f_{nyquist}/5KHz$. At the maximum of $ED[n]$ operation on line 5 essentially measures the product of the maximum spike intensity with the maximum derivative that proceeds it by R samples. This method primarily depends on the fact that spike detection looks for highly correlated narrow band energy which rejects a substantial amount of white noise. Moreover the operator compresses uncorrelated components in amplitude as it exhibits a square dependency in terms of $ED[n] \propto S[n]^2$ making variation in the threshold



(a) Gain and phase response of the combined IRR & FIR digital filter



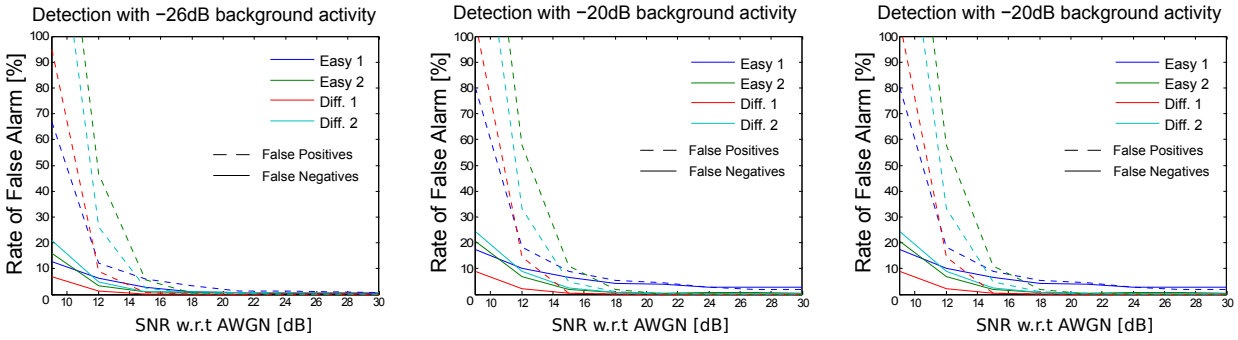
(b) Group delay introduced by the combined IRR & FIR digital filter

Figure 4.4: Extracted frequency characteristics of digital filter used in Algorithm 4.1.

less sensitive to detection. The fact that the operator is narrow band limits the detection of slower spike waveforms that do not contain large derivative components but on the other hand this grantees more systematic alignment. In this case alignment is done simply with respect to where the peak value of $ES[n]$ is detected.

The overall filtering characteristic of $S[n]$ and $G[n]$ is shown in Figure 4.4. The IIR bandpass is a result of k_2 being 0.5 such that both filters suppress components around the Nyquist frequency. The group delay should be equivalent to a single high pass pole at $250\,Hz$ but the FIR assists in further suppressing high and low frequency components. We should not expect significant contribution from group delay induced distortion as the features of interest will predominantly have $1\,KHz - 5\,KHz$ components. Besides R and k_1 can always be adjusted to reposition the high-pass poles closer to DC. Note that V_{LFP} will represent the $DC - 250\,Hz$ signal components that can be used to infer characteristics about the background activity.

The overall performance shown in Figure 4.5 reflects how spike detection is systematically accurate until the noise level approached 50% of the signal intensity irrespective of the data set with the default case where $k_3 = 3$. Note that the white noise is additive to the background activity implying $-14dB$ of white noise and $-14dB$ of background activity should evaluate to around $-8dB$ accumulated SNR. When the noise level exceeds the anticipated background activity for K_3 we observe a strong increase in the number of detected false positives. The rate of erroneously detected false negatives presents a more gradual increase but at this point classification is much more challenging. As expected background has a considerably bigger



(a) Detection with $-26\,dB$ background activity

(b) Detection with $-20\,dB$ background activity

(c) Detection with $-14\,dB$ background activity

Figure 4.5: False alarm rates normalized by true positives for data sets with different background activity.

impact on false alarm rate because spectral content and signal structure is equivalent to that of the foreground activity.

We can observe that the main component for computational complexity in this detection operator arises from the FIR & IIR high-pass filters where the order is closely related to the sampling frequency. In fact if we ignore the buffer used to capture features before the alignment event then this filter accounts for 70% of the memory utilization and 53% of the elementary operations while the rest is used for evaluating the instantaneous energy and performing overhead control. Note that the classification operator should be introduced in line 19 with a sample basis using the index as referenced pointer. This implies that we will be classifying while repolarization occurs at the electrode and our detection trigger is blanked out during this interval. This implies that we lose the capacity to detect any over lapping spikes. Such events have limited occurrences and missing such events can be acceptable because proper classification will likely fail as well.

4.2.2 Recursive Variance Decomposition

Another commonly used technique is that of principle component analysis (PCA) which extracts the largest loading vectors ν_n of the covariance matrix. This predominantly negates the systematic components of the captured signal and reduces the dimensionality of the spike

window to a sub-set of maximally varying features by linear transformation. These components are particularly useful as indication for spiking activity in the signal due to structure in ν_n but typically also suffice for providing a basis for classification in low noise conditions and reducing complexity once these vectors are found. The challenge specifically lies with the fact that determining this is basis requires both the computation of the covariance matrix that evolves over time as well as finding the transformation that diagonalizes that matrix. The implication is that in order to extract the first two principle components we need to track a total of $W(W+3)$ state variables where W is the number of samples in the spike window.

The iterative method employed here referred to as recursive variance decomposition (RVD) and is an approximation to standard PCA by recursively tracking the largest two loading vectors reducing the minimum number of state variables to $3W+3$. Similarly to PCA estimators like hebbian eigenfilters [155], every iteration incrementally updates the the learned basis without requiring extensive computation. The methodology is based on recursive extraction of the largest loading vector $|\nu_1|$ that is normalized by g_1 by checking the condition $(x - x \cdot \nu) \cdot \nu = 0$. This condition checks if there is any residue in the direction of ν_1 after removing its component to see if it is appropriately scaled. Moreover due to the strong correlation between the mean and first principle component we approximate that $sign(\mu) \approx sign(\nu_1)$ completing the extraction of ν_1. In fact these two statements allow a significant reduction in complexity as normalization is achieved through feedback. The noise shaping and orthogonality properties associated with PCA is preserved using this extraction which is the most important aspect.

Algorithm 4.2 shows the operation for estimating the first two PCA components. Here $D[n]$ is the new data point off set by the mean spike waveform $\mu[n]$ which allows the long term estimation of aggregate variance. Similarly parameter k_1 specifies how the state variables are exponentially averaged over the preceding data points. Because the projection of the first loading vector must be evaluated before the second vector these operations must be sequenced in time or with memory buffers. The evaluation of p_4 is strictly for illustrating the iterative method at which other components are evaluated while g_2 can also be adjusted to normalize the values of p_2 to prevent overflow without needing p_4.

Algorithm 4.2: Recursive variance decomposition

Data: Spike window W_1
Result: First two aggregate loading vectors ν_1 & ν_2

begin
  foreach *Sample* $\boldsymbol{n}$ *in window* // Projection phase
  do
    $D_1[n] = W_1[n] - \mu[n]$ // Get distance from mean spike
    Accumulate p_1 with $D_1[n] \cdot \nu_1 \cdot sign(\mu[n])$ // Project spike with ν_1
    Accumulate p_2 with $D_1[n] \cdot \nu_2$ // Project spike with ν_2

  foreach *Sample* $\boldsymbol{n}$ *in window* // Training phase
  do
    Update $\mu[n]$ with $k_1 \cdot sign(W_1[n] - \mu[n])$ // Track mean spike
    Accumulate $\nu_1[n]$ with $k_1 \cdot sign(|D_1[n] \cdot g_1| - \nu_1[n])$ // Move ν_1 towards $D_1[n]$
    Accumulate $\nu_2[n]$ with $k_1 \cdot sign((D_1[n] - \nu_1[n] \cdot p_1) \cdot g_2 - \nu_2[n])$
    // Move ν_2 towards $D_1[n] - p_1 \cdot \nu_1$
    Accumulate p_3 with $(|D_1[n]| - \nu_1[n] \cdot p_1) \cdot \nu_1[n]$ // Get gain error
    Accumulate p_4 with $(|D_1[n]| - \nu_2[n] \cdot p_2) \cdot \nu_2[n]$ // Get gain error
  Accumulate g_1 with $k_1 \cdot sign(p_3)$ // Adjust gain on ν_1
  Accumulate g_2 with $k_1 \cdot sign(p_4)$ // Adjust gain on ν_2

4.2.3 Template Matching using K-means

Finally we consider the implementation of template matching in channel. This can be seen as simply a K-means clustering method without dimensionality reduction on the input vector. The implication is that it is characteristically more memory intensive but requires less computationally intensive operators.

The implementation considered in Algorithm 4.3 is relatively straightforward where one section evaluates the generation of new templates and the other adjust existing templates with new data. The template approach in general has good noise performance due to the redundancy in correlated features that average out white noise. There is some usually some concern with respect to the convergence of k-means centroids. Typically due the the fact that noisy sample points may be initialized as new clusters and thereby wasting memory. The method used here is iteratively duplicating centroids after convergence. This minimizes the impact of noisy data in the feature space. As illustrated in Figure 4.6 during each iteration the centroids converge

Algorithm 4.3: Incremental K-Means classification

Data: Spike window W_1
Result: Classification with respect to aggregate clusters

```
1  begin
2      Accumulate Spike Count with 1                          // track accumulated statistics
3      foreach Sample n in window                                          // Projection Phase
4      do
5          foreach Template k in memory do
6              Accumulate p_k with W_1[n] − T_k[n]      // Get l_1 distance for each spike class
7      Find p_min = min[|p_1|, |p_2|, |p_3|, |p_4|] and Set c to index     // Find most similar
8      foreach Sample n in window                                            // Training phase
9      do
10         Accumulate K_c[n] with k_1 · sign(W_1[n] − T_c[n])      // Adjust most similar class
11         if Not all templates generated and Spike Count > k_2 then
12             Duplicate exiting templates
13             Set Spike Count to 0
```

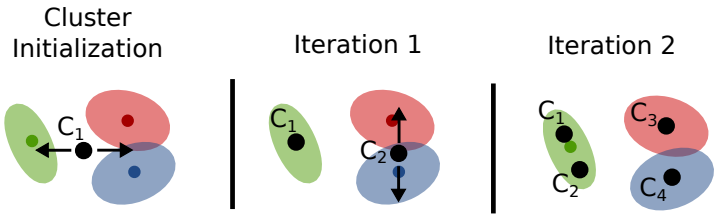
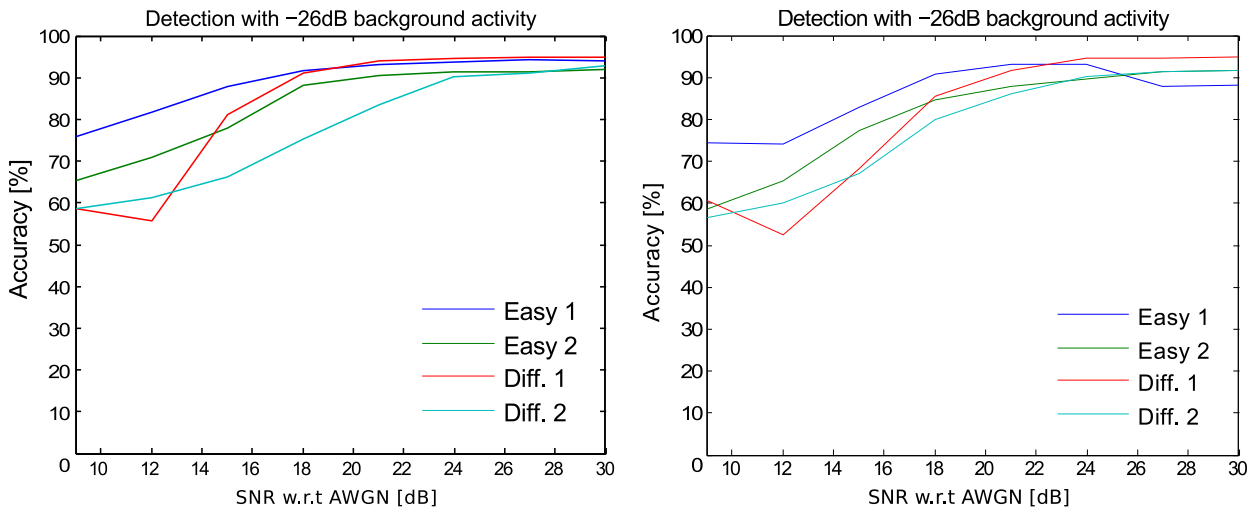


Figure 4.6: Illustration of centroid evolution over several iterations.

to mean positions. Due to the morphology that these centroids may be in we generally need more centroids than there are clusters but this approach works well when there are few spike classes. The assumption here is that we are clustering features that are characteristically Gaussian mixtures.

4.2.4 Complexity Evaluation

Generally the application of these methods should reflect a system level objective. For the configuration used here the memory and algorithmic operations are estimated in Table 4.1. Multiplications are equivalent to eight elementary operations and the memory calls are not considered as a computation but as load/store cycles. The impression made here is that tem-



(a) RVD classification

(b) Template matched classification

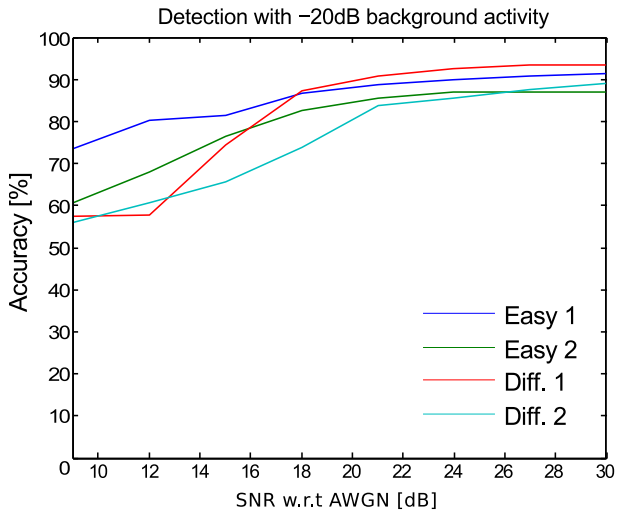
Figure 4.7: RVD and template based classification for data sets with $-26\,dB$ background activity.

plate matching is strictly very efficient if the the memory allows large allocation of active spike waveforms. Similarly RVD could show a considerable reduction in operations if a dedicated multiplier is introduced but that depends on how much we value compactness over execution speed. The disparity in memory requirement will dramatically worsen when the number of centroids is increased which is not the case for the computational complexity in RVD.

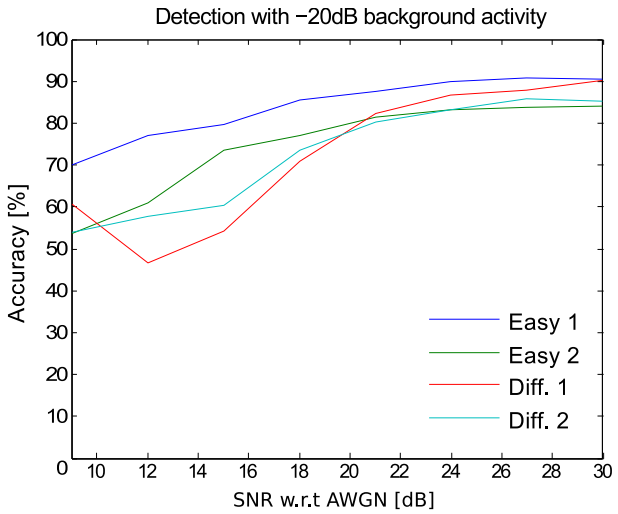
Table 4.1: Estimation on memory and computational resource requirements for each algorithm.

Algorithm	Memory	Operations	cycles per sample
NEO Peak Detection	20 Elements	30	56
RVD / training	63 Elements	29 / 88	57 / 116
Template / training	85 Elements	9 / 16	27 / 34

The empirical results in Figure 4.8 generally show that in moderate noise conditions our classification accuracy is typically better than 85% which is calculated in terms of the aggregate probability of correct classification multiplied with the probability of missing a spike event. Unsurprisingly RVD is not very effective in noisy conditions where the variance accentuates irrelevant components. The classification accuracy from template matching is also shown in Figure 4.9. These results should primarily show an improved noise rejection characteristic but more generally this approach is more resilient at dealing with false positives. In principle a new cluster will be assigned to a zero mean template representing the false positives while

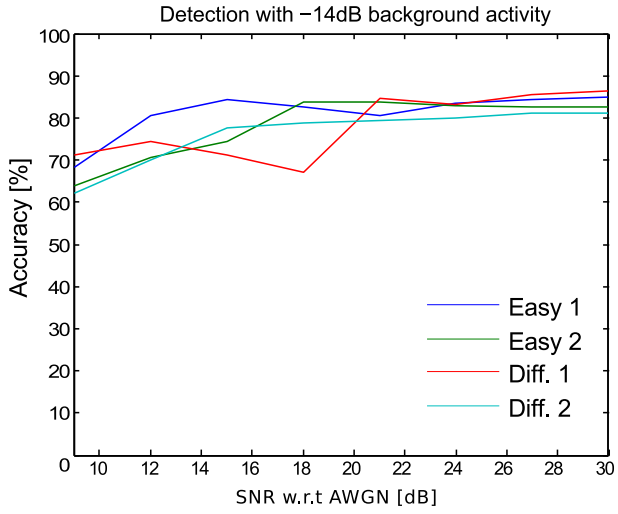


(a) RVD Classification

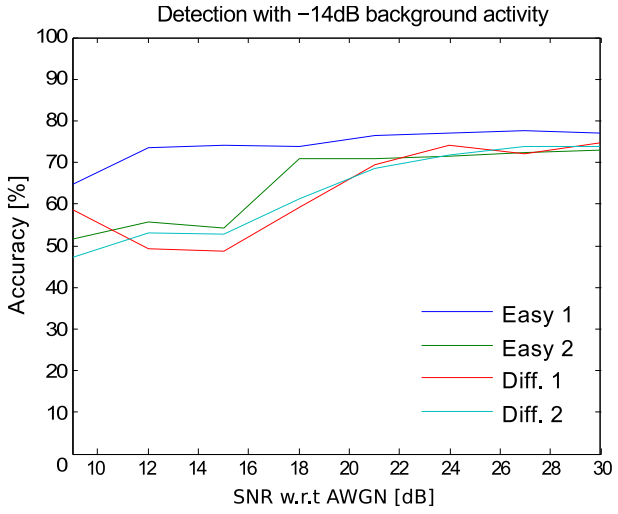


(b) Template matched classification

Figure 4.8: RVD and template based classification for data sets with $-20\,dB$ background activity.



(a) RVD classification



(b) Template matched classification

Figure 4.9: RVD and Template based classification for data sets with $-16\,dB$ background activity.

maintaining the other templates intact. Strictly the detection circuit should be readjusted to favour increased detection of false positives as long as the rate of false negatives remains low. But instead exactly the same parameters are used for every test.

We should be careful to judge the effectiveness of these implementations particularly with respect to efficiency. While we can generally increase performance by allocating more memory or introducing additional computation we need to quantitatively evaluate the objective. We suggest normalizing the resource allocation with respect to increased information extracted from the signal by classifying. That is how much more processing are we allocating for classification by proportionally increasing the signal to noise ratio of our output. In the optimistic case when the three classes neurons being detected are uncorrelated our base-line accuracy would be 33% while needing 56 cycle of operation for spike detection. In fact this leads us to believe both algorithms in this respect decrease resource efficiency by a factor around 2 accounting for an increased memory, processing requirement compensated with increased accuracy. While this claim is very sceptical with respect to the motivation for classifying spike events it also reasons the aggressive reduction in algorithmic complexity through approximations presented here. There is genuine benefit in classification that assists the convergence of further processing algorithms. In addition we simply argue that excessive dedication of resources that exceed that needed for signal conditioning may not be worth while. They key point demonstrated here is that these methods appear very much attainable in terms of on-chip processing capacity. Here we considered the case without supervision specifically in consideration for scalability. It is likely that further reductions or optimizations can be made in that regard to the structure of these methods to improve accuracy and noise tolerance.

4.3 System Architecture

The conceptual architecture of the system proposed here is foremost based on the opportunity for software defined real-time instrumentation that has not yet be exploited in chronic

implantable systems at this scale. Currently it is common place to see synthesized logic that performs all processing and data handling procedures in such a way that they have very limited high-level reconfigurability. This is strictly in order to save power and reduce complexity at the system level. It is important to note that for any recording device there are a multitude of phases during its operation where this flexibility can be highly advantageous once sensor characteristics are learned. Like discussed in Section 4.1.1 many classification algorithms benefit from training or characterizing the recording conditions.

The approach to specialized DSP in the literature reflects two problems in this field. The first is signal extraction from recordings that relates to what we have discussed in terms of spike detection to extract compressed spike train data. The other is associated with accelerating adaptive filters that map these spike trains on to estimates for cognitive dynamics or invoked limb movement. Typical examples for spike sorting are fully synthesized cores [49] [158] that can be integrated and are capable in achieving respectable processing capacity for specific algorithms. In contrast to spike train decoding that is predominantly performed by FPGAs as integration make less sense for the development high dimensional adaptive filters like [159] that do not need to be embedded within body. Interestingly the work in [40] proposes a application specific instruction set processor (ASIP) that similarly argues for high performance computation for these structures with a high degree of flexibility that reflects the different models used for spike encoding. Additionally we see the advantage of using off-chip microcontrollers like MSP430 that interface with a highly reconfigurable instrumentation front-end to leverage both adaptive and involved noise shaping to perform more intricate operations such as seizure detection or artefact removal [160]. While these works may not be viable for high channel count implantable devices it does highlight the considerations for designing fully integrated prosthetics that is in-line with this work.

Here we will consider a particular type of microcontroller topology that can support reconfigurable functionality and reflects the fact that although multichannel BMIs are highly parallel in nature the associated processing can also be algorithmically intensive. The feasibility of this notion has been estimated to an extent but many components are subject to implementation. In essence we optimistically approach this design problem with a strategy that exploits both

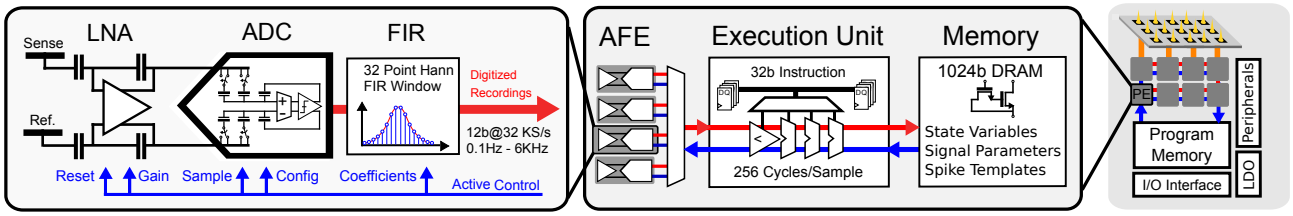


Figure 4.10: Illustration of the proposed distributed μC array for homogeneous program execution at the sensor interface.

the homogeneity in processing and the information locality in order to realize a feasible solution. This lets us focus on the in-channel operation where efficiency is maximized through the topology of the execution unit. Regardless of the end result this proposed system will be one step towards the goal for more effective chronic neural implants.

4.3.1 Distributed System

The system illustrated in Figure 4.10 represents the distributed microcontroller architecture. The primary mechanism of operation is the program memory that continuously feeds the stored instructions into the pipelined array of processors that operate locally on the recorded data. The execution of these instructions is handled with what is essentially a instruction decoder, memory module and an arithmetic unit that is interfaced with four analogue recording channels. This approach guarantees that the absolute minimum amount of energy is required for the communication of recorded data as the information is processed and consolidated to its elementary component at the quantization interface.

Inherently this implementation will sacrifice the availability of more intricate functionality found in DSPs since the data is not funnelled into one processing unit that can be very elaborate in complexity. The distributed structure is rationalized by the fact that the intensive operations such as clustering methods operate at a much lower speed due to the sporadic spiking activity that make statistical convergence slow. Furthermore these adaptations need to be performed on the order of minutes by which such functions may also be implemented through the redundancy of elementary operations. Moreover multiplexing loses effectiveness in memory intensive

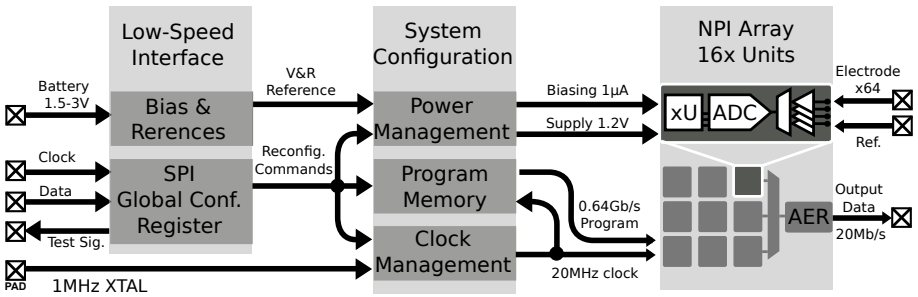
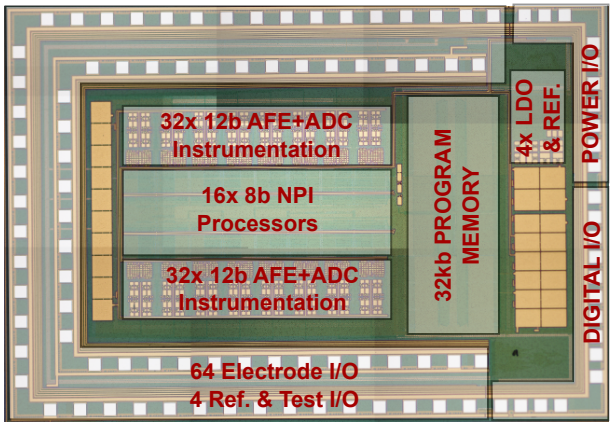


Figure 4.11: System architecture for NPI sensing platform with digital interfaces annotated.

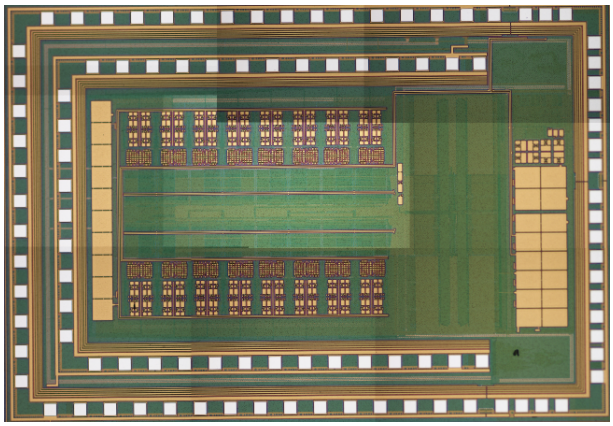
applications as it does mitigate the power & area scaling associated with memory allocation.

Also consider that the program control that gives this implementation its capability for generic computation does not scale with the number of processing units. This is an important distinction when addressing a hundreds of channels on chip that will allow this implementation to outperform any other architecture and leverage the fully integrated form factor. We also note that whether this architecture is realized by synthesized logic, FPGA fabric, or more custom logic cells is insignificant to the extent that the memory structure plays a more profound role. This claim is based on the algorithms in Section4.1.1 that allocate significantly more resources to memory than algorithmic operators. In particular memory density and efficiency is a critical component to the success of this type of large scale sensor system. Here 3-T eDRAM is employed which is more effective than alternative solutions memory solution and can still be realized on a standard CMOS process [161]. When compared to an SRAM equivalent we find it can readily achieve a factor 8 improvement in density [162].

The high-level interfaces are illustrated in Figure 4.11. There are multiple layers with respect to how internal resources are accessed for reconfiguration. This is primarily for robustness where each layer increases in complexity and chance of failure. The low-speed interface is the simplest element which acquires commands from an external device with very relaxed requirements on input timing. These commands allow us to reconfigure the high level sub-blocks like tuning the generated reference voltages provided by the power management, control reset/power of individual sub-blocks and selecting which digital test signals should be monitored. In particular the processor array and program memory layers almost operate in isolation to the peripherals. These blocks are timed by the internal PLL structure that drives significantly higher data rates



(a) Layout floor plan of the fabricated micro-controller structure.



(b) Chip micrograph showing the distribution of toplevel rounting.

Figure 4.12: Physical implementation of NPI system using a 6-metal 0.18 μm CMOS process.

that do not need to propagate to the pad level in order to save power. The back-end of the system similarly communicates data uni-directionally between two different clock domains to send data packets off-chip using a number of handshaking protocols.

The implementation of the analogue circuitry has been discussed in Section 3.2.4 where we additionally constrain all algorithms to a maximum of 1024 cycles per sample while maximally allocating 128 words of memory. With respect to our previous discussion this amount of hardware should allow a large set algorithms that are resource efficient. If not the topology will promote the construction of processing with more aggressive memory efficiency and using feedback dynamics to implement more complex operators such as division. It should be noted that these specifications have flexibility by sampling multiple times per program cycle or reducing the system clock using the configurable phase locked loop in order to reduce power.

Figure 4.12 presents the fabricated prototype device. It can be seen that integrating many peripheral blocks such as a phase locked loop, voltage supply regulators, and program memory on chip minimizes the pad count required for the digital and power domains. However even for a 64 channels system the number of analogue pads required for the sensor interface play a significant role on top level organization. In addition careful consideration has to be made with respect to how the digital signals propagate where minimizing track length not only reduces digital noise coupled to the substrate but more significantly the associated power dissipation.

The number of processing elements can in fact quite easily be scaled up by extending the instruction pipeline where the system level timing constraint for speed and fanout lies with the program memory which has an internal pipeline that needs to connect the program memory together.

4.3.2 Processing Core

In order to allow the hardware to provides generic processing capabilities in a distributed fashion a number of considerations have to be made. In particular we need to reflect the typical operations with certain modalities of operation. It is clear that although all recording channels should execute the same algorithm they will typically not share the same state of operation. This state dependency is exemplified with respect to intermittent processing during bursting neural activity and idling during quiet periods. This is an inherent limitation to sharing the program memory as the dynamic execution of the code where each core has its own program counter or a top level scheduler is not feasible for an arbitrary number of channels. The quasi-out-of-order execution makes it challenging for us to adopt scalable tile structures found in image processing [105] that excel in maximizing area and power efficiency in a scalable sense.

Lead by maximizing the locality of data execution [163] where this aspect of branch control or conditional execution is mediated by skipping a section of the incoming instructions if a condition is not met. The approach of skipping sections of code up on branching is relatively in-efficient with respect to throughput. This approach is optimal at the system level when individual cores may need to execute any section code and branching will only be limited by the dissipation related to the registers pipe-lining the instructions across the chip.

The individual components of the execution unit are shown in Figure 4.13 and details the main data buses used for exchanging data. The majority operations revolve around manipulating data in the registers R1-R16 as A operand in association with any other data sources that can be used as B operand. The operation performed by the arithmetic logic unit (ALU) will

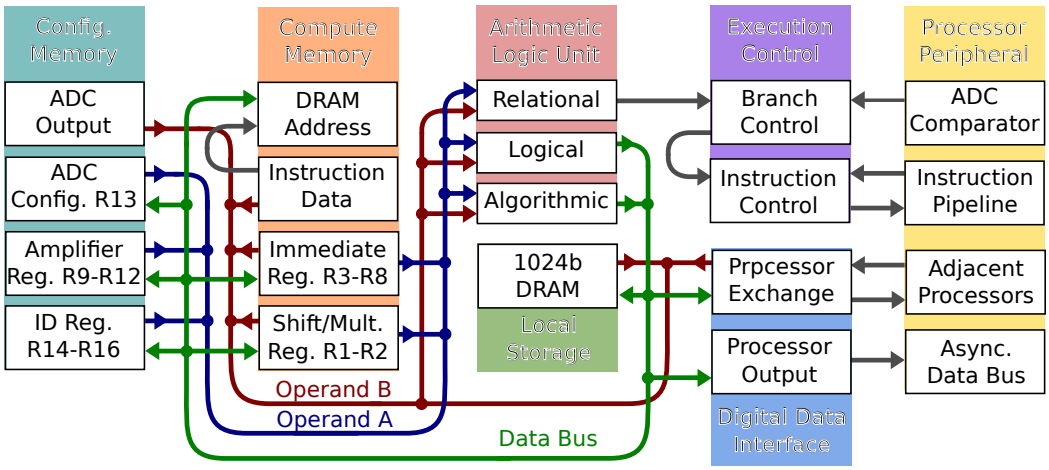


Figure 4.13: Organization of the distributed execution unit detailing components and the interconnect.

always overwrite the result to the location of the A operand but can in extension also be used to to write to other locations (i.e. memory, periphery, etc.). This implies that in terms of instructions there are always two components where the first is simply the operation executed by the ALU in addition to the two memory sources. The second component optionally extends this simple functionality by writing these intermediate values to multiple other locations or arbitrary branching operations that will take the unit out of sleep.

On that note we mention that the local execution controller consists of three registers that assist in branching operations or conditional execution. When either of these registers have logic one the instruction is gated by a null operation before execution. One of these registers will self reset allowing for if-else functionality by skipping a single instruction. The two other registers need to be cleared actively but in combination this will allow for nested conditioning of up to three levels. While in idle state no internal registers are clocked with the exception of the instruction pipeline and the branch controller saving a significant amount of power as the instruction does not need to be decoded.

The digital data interface provides the means for communicating data either off chip or to adjacent execution units. This functionality allows granular consolidation of features or signal structure and correlate measurements with system level parameters. For example consider each execution unit is listening to the most informative analogue instrumentation channel, it is conceivable that comparing its spike train with that of an adjacent units to evaluate neural

interconnect level features. The Asynchronous data bus on the other hand is a key feature that allows this system to appear as a slave at the network level that does not need to be coherent with the system or off-chip clock. This bus is in essence a large buffer distributed across many channels utilising asynchronous hand-shake protocols to funnel the data to a SPI module that is clocked either externally or internally [164]. This solves a number of coherence problems that mitigates the need of having a FPGA to drive this system as the SPI module is not timing critical. Furthermore this alleviates clock distribution as the timing constraints are always local to each execution unit and not the data bus that is distributed across the chip which may either be very restricting or power intensive.

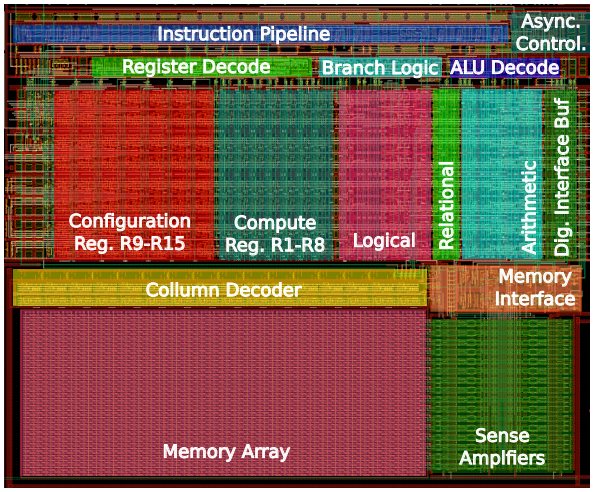
The dynamic control with respect to the analogue channel is enabled by one designated 8 bit register per analogue channel. In this particular case 4 bits are used to specify gain, 2 bits for configuring the biasing current as 0x,1x,2x,3x, and 1 bit of the reset function. In particular the reset phase will temporarily boost the transconductance on the band-limited filtering stage to allow sub-microsecond auto-zero for active noise shaping. For both the ADC and the amplifiers there is one bit that controls a multiplexer at the input that can switch in the sensor or a global differential test net for calibration or verification. Similarly the ADC has 2 bits to select which analogue channel another bit to clock at the full rate or half the rate of ad joint micro-controller. In addition there are 3 bits to control the how the chopper frequency is divided from the sampling signal which is the final control bit. Understandably the analogue configuration will remain static after the appropriately being set. The ADC configuration register is considerably more dynamic as the multiplexer needs to be reconfigured and sampling needs to toggle persistently.

There are two modes of getting quantized data from the ADC depending on the desired functionality. The first is simply reading the 8 bit quantization register that shadows the 7bits quantized by successive approximation and the LSB from the first integration result. In order to utilize the higher resolution capability the comparator output is used to integrate coefficients from the instruction onto a local register where the comparators will decrement or increment the register accordingly. If no calibration data is locally stored this operation first integrates binary weights on one register during the SAR cycles and then integrates the FIR window onto

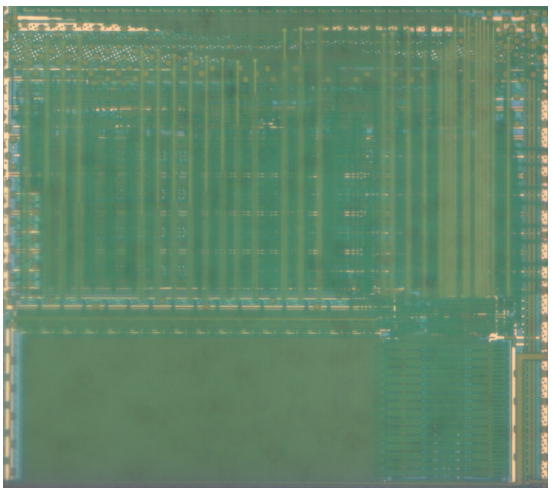
another register. This is large investment of cycles to perform high resolution quantization but this can be optimized for specific applications when it is necessary. If the calibration data is available for the 7 SAR weights then the ADC must be configured to run at half the system speed and before quantization these weights are loaded from the memory onto registers R2-R7. Followed by the usual process of SAR quantization while these weights as simultaneously also integrated on a second register. Then after the integration phase three registers will contain quantized data. The scaling of coefficients is key and should be such that the $\Sigma\Delta$ result simply copies the sign bit of the SAR operation and can concatenate the lower 7 bits with the SAR result. Then the calibration data is scaled appropriately and added to the 14bit signed double with carry logic. Clearly there a number of conventions suggested here that will best exploit the capabilities of the design.

The memory module local to each execution unit hold 128 words of data which can be shared across the analogue channels with 32 locations each. Particularly when the DSP is mainly performing filtering the recorded data can be buffered for FIR filtering or keep its high precision filter state variables for IIR structures. These filter and program coefficients are stored in the shared program memory such that the execution unit does not experience an overhead in memory requirement. However for other memory intensive algorithms such as template matching, serving the most informative of the four analogue channels will have to suffice because the memory requirement is beyond the capabilities of this configuration. The DRAM architecture has a refresh-up-on-read mechanism which implies that the used memory locations will have to systematically be read to keep the data stored valid. Fortunately this requirement is self fulfilling as the program recycles itself every $100\,\mu s$ and the DRAM retention time is on the order of $1\,ms$ implying that as long as there is a guaranteed read on the memory location it will stay valid. The physical read mechanism however does require a minimum of two cycles. The first is in the background which simply prepares the internal registers of the module while a different execution is taking place and the second is in the foreground where the location is read and the data bus is driven by the DRAM.

As the illustration in Figure 4.14 shows, keeping the 8 bit structure in terms of parallel operations maintains a very compact floor plan. This is typical of data flow intensive designs



(a) Layout floor plan of the distributed execution unit detailing components.



(b) Chip micrograph showing the distribution of toplevel rounting.

Figure 4.14: Physical implementation of execution unit using a 6-metal 0.18 μm CMOS process

where the digital logic should be placed underneath the associated data buses. This is difficult to replicate by automated synthesis tool where signal congestion is the most stringent aspect. The digital signals for the two operands and the data line span horizontally where sub-blocks extensively take advantage of the gated output buffers for each sub-block that is controlled by the decoders. The full custom approach taken here sacrifices design effort for additional performance in terms of reduced parasitics and more aggressive power gating.

$$\textbf{<C>,[<CE>],<A>,<B>,[<OE>*]} \tag{4.4}$$

The syntax for constructing instructions needs to be in the Backus Normal Form [165] as formatted in Equation4.4 with reference to Table 4.2 which summarises all possible compositions. A parser is implemented that will translate an ordered set of these instructions directly into hardware specific machine code that needs to fed into the instruction pipeline any violations or exceptions will be caught by this script automatically. Although there is no dedicated multiplication hardware there are specialized registers that allow shift add based multiplication over eight cycles. Any other primitive logical or arithmetic function can be realized with this instruction set as it is turning-complete. This assertion is made by noting that it can evaluate the operation; subtract and branch if less than or equal to zero, which is sufficient for a one

Table 4.2: Overview of instruction sub-components.

Index	Operation Subset	Summary of Possible Entries
C	Logic Operation:	Logical Shift Left/Right, Arithmetic Shift Left/Right, XOR, XNOR, AND, OR, MOVE-A, MOVE-B
C	Arith. Operation:	Compare, Add, Carry Add, Multiply, Complete Multiply
CE	Compare Option:	>, =, <, Overflow
CE	Add Option:	Subtract, Absolute Value, Increment Overflow Bit
CE	Mov Option:	Mem. Address is from Data line. Default is from Instruction
A	Operand A:	R1-R8, R9-R13, ID, Count, Memory
B	Operand B:	R1-R8, Left uC, Right uC, Instruction, ADC, Memory, Null
OE	IO Extension:	Write to Left uC, Write to Right uC, ADC Sample enable, Write Output Buffer
OE	Branch Extension:	Write to Branch Register BR1-BR3, Invert Branch Result
OE	Memory Extension:	Write Address, Write Data, Read Data

instruction set computer [166].

It should be mentioned that there a number of hardware specific details with respect to how certain instructions behave that need careful consideration towards the implementation details. For example if no comparison is made but a branch register is accessed the output of the comparator will be treated as false no matter what logic the overflow bit is. This allows us to clear or set branch registers while simultaneously performing an operation. Another example is that by default the instruction data is ready at the input of the memory address to prepare a read in the background. In most cases it is intuitive and we simply strive to maximize the cycle efficiency. At all times the execution unit is capable of dealing with the compute aspects while performing branching and memory access simultaneously.

This work also provides an elaborate set of test tools that allows compilation of instruction code and the generation of piece-wise-linear 'csv' files for test sources that can be used in the circuit simulators. This can be used in association with the transistor or verilog implementation of

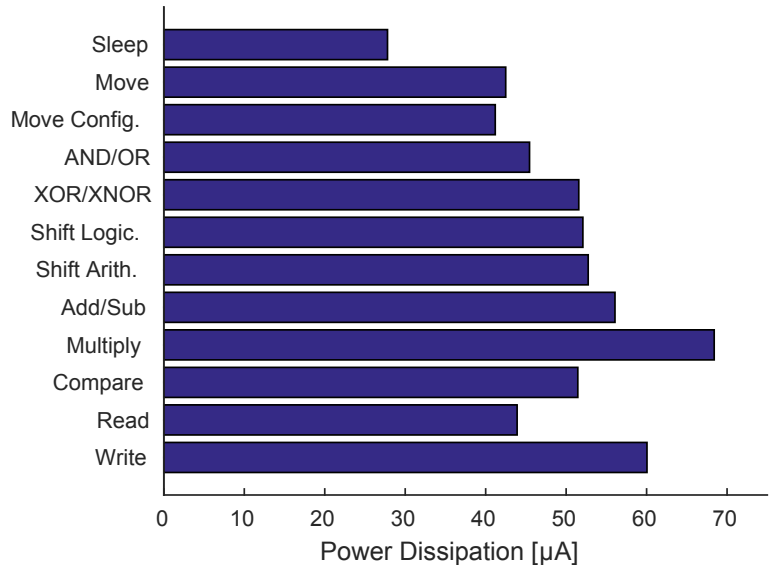


Figure 4.15: Power dissipation with respect to specific operations for the same operand A=113 & B=114 in randomized order.

the processing core. The behavioural models in particular are important for the translation of this architecture to other implementations.

The results in Figure 4.15 exemplifies the dependency of power dissipation with respect to different operators for the same operand A and B. It should be expected that the is a strong operand dependency with respect to power consumption but these results follow our expectations closely. Generally the simpler the operation the lower the current dissipation is because less complexity is involved with the switching losses. Here again we observe that when the unit is in a sleep or branching state the power dissipation is mainly associated with the instruction pipeline. As this 32bit pipeline transverses the entire execution unite it plays a significant contribution towards the baseline power consumption. The typical power consumption for full activity will lie around $45\,\mu A$ it should be noted that sporadic spiking activity will gate the majority of operations and it is likely that running at half the designed rate with 512 cycles is more than sufficient. Note the typical figure of power is $2.7\,pJ/Cycle$ or $2.7\,\mu W/MIPS$ which is several orders of magnitude better than 16-bit microcontrollers such as the MSP-430 [167].

The specifications given in Table 4.3 summarize the main features associated with this system on chip for processing neural data at the sensor interface. As the total power consumption is on the order of $1.5\,mW$ there is some concern with respect to the power density associated

Table 4.3: Summary of performance specifications for the NPI system and state-of-the-art specialized integrated processing architectures. ⋆ Reconfigurable topology.

Parameter	Unit	This Work	Markovic [168]	Arimoto [105]
Architecture		Distributed μC Array	Multi-Grain FPGA	Dedicated Tile Array
Technology	[nm]	180	40	65
Supply Voltage	[V]	1.2	1	1.2
Parallel Units		64	16⋆	2048
Instruction Size	[bits]	32	-	32
Operational Frequency	[MHz]	20	400	300
Sampling Frequency	[S/s]	$32k$	$100M$	-
Operations per Sample	[Cycles]	256	4	-
$P_{Digital}$ per Channel	[μA]	44	-	-
$P_{Analogue}$ per Channel	[μA]	16	-	-
System Power	[mA]	1.42	11.6	300
Program Memory Capacity	[kb]	32	-	-
Processor Memory Capacity	[kb]	1	36	1
Processor Array Area	[mm^2]	1.04×1.32	3.8×5.4	1.60×3.19
Power Efficiency	[GOPS/mW]	1.52	0.86	0.31
Area Efficiency	[GOPS/mm^2]	0.88	2.34	36.1

with the system in full operation that in this particular case is $26\,mW/cm^2$. In fact if the number of channels is scaled up beyond 64 channels this power density will tend to $29\,mW/cm^2$ but will not exceed it. Either figure will likely be smaller subject to the physical & software implementations but more importantly will not result in a thermal agitation or the heating of cortical tissue that exceeds $2°C$ [66]. More generally we have the advantage of tuning processing capabilities to the heat-capacity of the implanted package. In fact comparing this work to state of the art FPGA topologies [168] and highly parallel ASIC structures [105] that follow the same design methodology we find that power and area efficiency that exceeds that of stand-alone microprocessors by orders of magnitude. These figures also reflect the expectation that technology scaling should lead to even more compact configurations. In addition Gate leakage may introduces some diminishing returns with respect to power efficiency. We mention that these figures are extrapolated based on the performance of a single execution unit and we expect more overhead from other components that is not accounted for in this comparison.

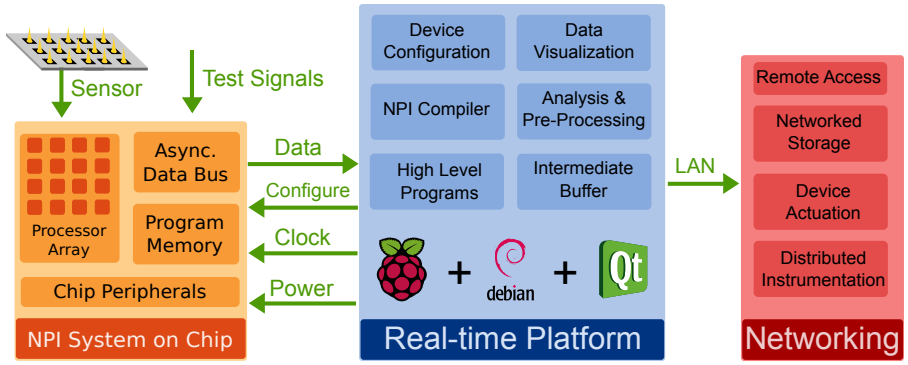


Figure 4.16: Block diagram of the instrumentation platform developed as framework for real-time applications.

$$R_D = \frac{P_{\mu C} \cdot A_{\mu C}}{N_{chan}^2 \cdot Cycles} = \frac{44\,\mu W \cdot 196 \times 158 \mu m^2}{4^2 \cdot 256} \approx 3.3 \cdot 10^{-16}\ \left[W\,mm^2\,per\,OP\right] \tag{4.5}$$

Re-evaluating our power/area figure of merit in Figure 4.1 with Equation 4.5 we observe that practically we lose a factor of ten in efficiency when compared to a dedicated ASIC implementation because resource utilization inside the execution unit can not be maximized. This was expected given that instead we attain high-level reconfigurability. However this does achieve a very good understanding with regard to where the system scales from this point both with respect to area and power requirements.

4.3.3 Testing Platform

As this system is directed at generic use for the neuroscience community where high level programming and interfaces are essential for end user adoption. The testing platform presented here is aligned in such a fashion that its fundamental components can be extended upon greatly to serve a multitude of needs. This ambitious design criteria is primarily provided by the real-time platform illustrated in Figure 4.16 that supports a standard Linux operating system. The thee components compromise of the custom NPI system on chip, the Raspberry Pi platform, and networked resources.

The software stack running on the Raspberry Pi primarily handles the high speed SPI link that

fetches data from the NPI system at 10 Mb/s and stores it to a local buffer for some of the data visualization. This data stream is then forwarded to a network routine that is connected to a server over the local area network via a UDP protocol to allow large quantities of data to be stored in a scalable fashion. The graphical user interface is built on top of this process in order to give a means to both configure the device actively and provide some form of interactive interrogation with respect to the recorded data and the algorithm being executed.

The application of a generic internet of things platform plays a important role with respect to long term development objectives. It signifies that the ASIC is there to provide a specialized interface with the sensor and a generic digital interface with the external control to allow rapid adoption of new techniques or other components as software extensions. This substantiates the modular approach where design effort is explicitly focused towards specialized hardware for the sensor and software development at the system level. This is important given the complexity of these systems where overspecialisation limits the versatility of existing designs thereby limiting the utility of other commercially available tools/devices.

The advantage here is that a multitude of procedures can be run on the real-time platform without supervision that are detailed in high-level programming code that have fast development and turn-around capabilities. In this case it significantly improved test procedures by enabling automated exhaustive characterization of logical integrity. In fact the standalone module of the microcontroller structure can run 1 MIPS of on the fly randomly generated operations. This can be seen in Figure 4.17 where the Saleae logic analyser is used to probe the internal data bus of one particular core.

This is partly shown in Table 4.4 where the internal bit-line of one such execution unit could be directly accessed. Because it is not viable for us to exhaustively simulate the hardware in various conditions we use a physical test bench in order to record the performance tolerance with respect to voltage supply and operating frequency. Moreover what the user sees is reduced to latent frames of data over several milliseconds and the corresponding instruction code executed by the platform. The physical interfacing protocols are very much transparent. By construction each core has a hard wired ID that will allow the active supervision of internal variables for

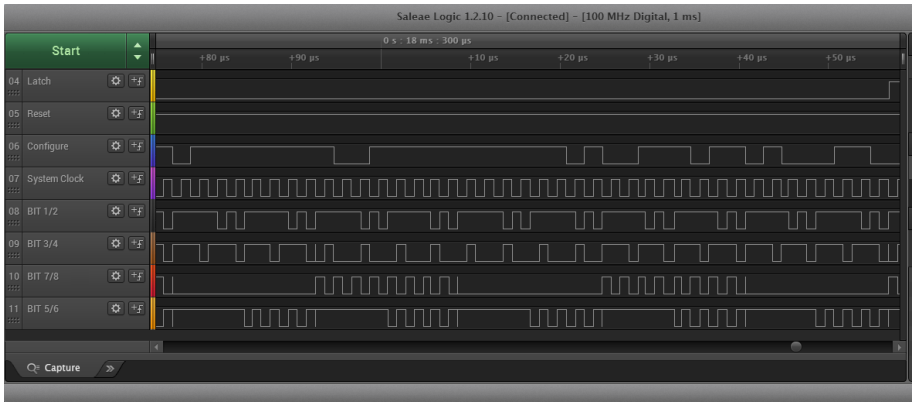


Figure 4.17: Digital waveform of the internal data bus BIT 1-8 as new instructions are being loaded into the device using the clocked Latch and Configure signals.

Table 4.4: Section of Instructions and recorded outputs from μC structure with the associated machine code.

BITLINE	**INSTRUCTION**	**Machine Code**
00011010	MOVB R5 DINST 26	00111111000000000011000000011010
11101111	MOVA R3 DINST -17	00110110000000000011010011101111
00001010	AND R5 R3	11101111000000000000000000000000
00011010	MOVB R5 DINST 26	00111111000000000011000000011010
11101111	MOVA R3 DINST -17	00110110000000000011010011101111
11111111	OR R5 R3	11101111000000000000010000000000
00011010	MOVB R5 DINST 26	00111111000000000011000000011010
11101111	MOVA R3 DINST -17	00110110000000000011010011101111
11110101	XOR R5 R3	11101111000000000000100000000000

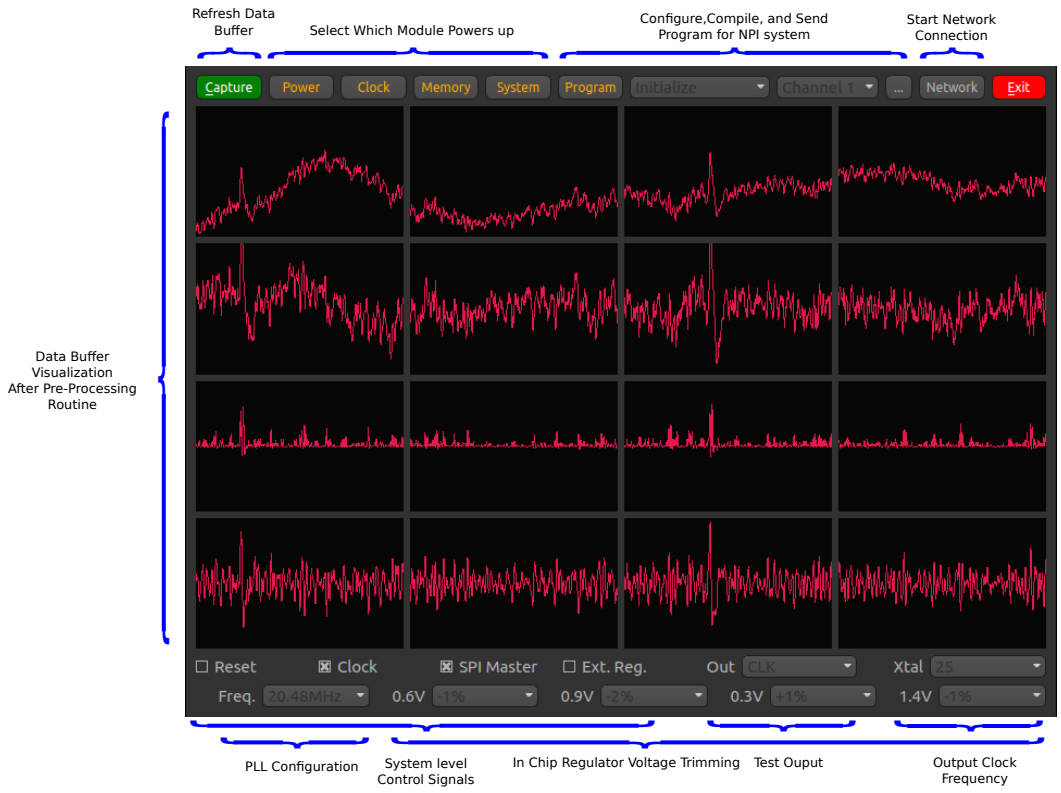


Figure 4.18: Graphical user interface used for configuring the NPI system showing test data.

development and debugging of single units. Due to the specialized hardware the instrumentation programs currently still require careful tailoring of the instruction code but this can be extended towards compiling directly from C++ code that is also used to construct the rest of the platform.

Figure 4.18 depicts the GPU accelerated graphical set-up used for testing the device where the functionality is mainly associated with reconfiguration and powering different system sub-blocks for validation. From a engineering point of view it is more of a convenience to have automated reconfiguration of the device as one interacts with the various settings. Particularly in associated with probing the supply voltages or analogue reference signals generated on chip. It would be more typical that during experimentation this functionality can be reduced to simply selecting from a set of predetermined programs.

In order to move towards fully isolated operation which will be the case for a implanted device the system on chip architecture relies on a minimum amount of off-chip components in order to bring the resource requirements of the topology into scope. This is shown in Figure 4.19. These feasibility considerations are generally with respect to reasonable assumptions associated with

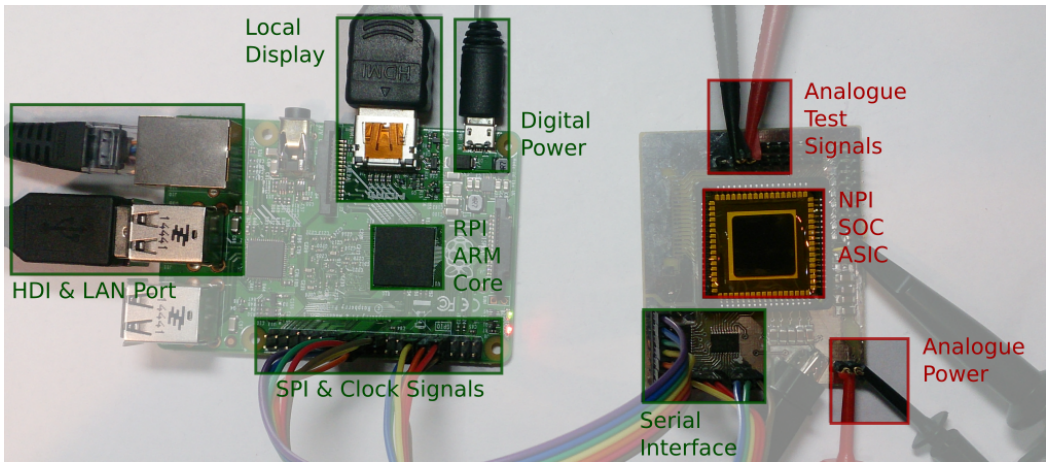


Figure 4.19: Test bed used for characterization with various components illustrated.

a wireless implant that is hermetically sealed. In this particular case we will allow a number of off-chip decoupling capacitors, a reference resistor and a reference voltage which may very well be integrated on chip in one way or another if necessary. The system also uses a $1\,MHz$ external clock reference which may be realized at the wireless power carrier frequency and is locked onto with a phase locked loop to generate the internal $20\,MHz$ system clock. Three linear LDOs were integrated to provide a $1.2\,V$ supply to the digital,analogue, and memory separately. Where the analogue supply voltage used to derive internal ADC voltages references of $1.2\,V, 0.9\,V, 0.6\,V, 0.3\,V$ from the unregulated supply using high speed buffers.

4.4 Conclusion

This chapter substantiates a scalable and long-term approach for the development of programmable neural interfaces. In particular we discuss why moving away from the fixed purpose DSP architectures seen in many conventional systems is of significance with respect to performance and reliability. In addition we provide indicators that show the majority of modern CMOS technologies using dedicated on-chip processing hardware is viable to perform local signal analysis. Furthermore we highlight the importance of efficient algorithm construction were operators should revolve around execution per sample and processing structures that im-

prove scalability for systems with many recording channels in association with the near-data-processing paradigm. PCA & template maching methods are proposed for embedded systems that require 57 operations per sample and 680 bits of memory with entirely unsupervised operation that can achieve over 80% accuracy during spike detection & classification.

A distributed micro-controller structure is proposed in effort to realize these characteristics and reveal underlying constraints. The topology reflects the nature of processing neural data in the context of achieving generic computational capacity. This discussion details both low-level and system level considerations that address the software stack. The impact of memory requirement that results from being able to execute arbitrary algorithms in isolation is evident both in-channel and chip level. In the proposed configuration the amount of resources allocated for this function is comparable to that of the signal processing but depends very much on the number of channels that are integrated together. We point out that if the number of channels is increased this component does not change and allows this topology to become more effective. The distributed processing architecture operates with an efficiency of 1.52 GOPS/mW and each core only requires a 0.02mm^2 silicon foot print with fully reconfigurable 8 bit processing capabilities.

The foregoing discussion has depicted the intricate complexity associated with these sensing systems and revealed the diversity of aspects that should be taken into consideration. Sustainable development for these systems will need long-term solutions due to the excessive design effort that prevents rapid turn around and progress. Moreover innovation needs to be contextualized at the system level to ascertain whether new techniques and methods have significant impact. This requires the abstraction and modelling of these implementations to gauge impact using empirical indicators.

Chapter 5

Time Domain Techniques

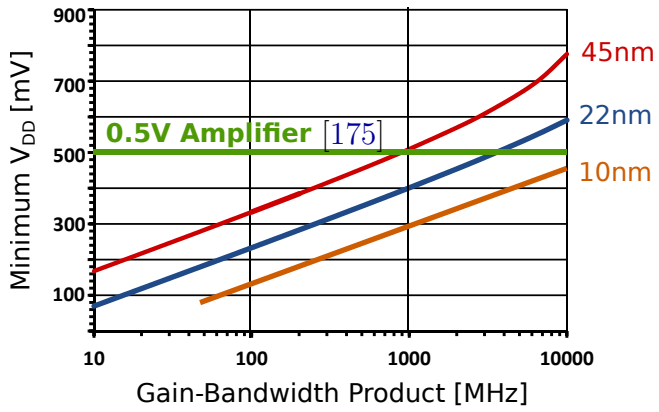
Thus far our work has detailed numerous design techniques that extend on contemporary work where the classical analogue approach with digital processing has demonstrated its capabilities. However we have also analytically shown that although we can still strive to improve area and efficiency, there are a number of factors that prevent making significant progress in terms of improving system characteristics. Moreover there is a strict need for more efficient computational processing that appears overwhelming if it is made robust and adaptive. If we keep the current processing methodology this component can only be made viable with smaller technologies and voltage scaling that can substantially diminish the performance of analogue operations. Here we will attempt to address the two factors that have the most significant impact on improving sensing electronics based on the observations made in the foregoing discussions to consolidate this work. The first is introducing all-digital instrumentation that is not diminished by technology related scaling and the characteristics of nano-meter transistors. The second objective is developing a mixed signal topology for analogue to information conversion where feature extraction is performed adaptively in the analogue domain.

This chapter will focus on exploring the emerging time domain processing modality in order leverage increased digital performance associated with modern CMOS processes. In fact this motivation is carefully addressed in the literature [169] where logic-gate based topologies demonstrate better scalability with respect to linearity and bandwidth. Here we will demonstrate how

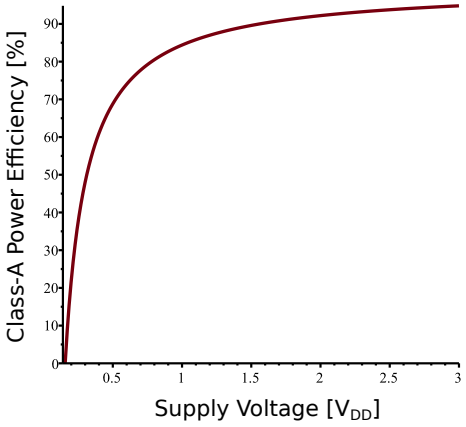
the fundamental limits of noise efficiency can be approached by proposing several topologies and design techniques. Further we will elaborate on the characteristic relations between analogue performance and resource requirements that enable these structures. The organization of this chapter is as follows. Section 5.1 will introduce the essential design considerations for continuous time-domain circuits by considering the phase characteristics of oscillators in relation to driving transistors that are used for analogue feedback. This structure will be used to implement both amplifying and filtering structures to compose the instrumentation front end. This is followed by Section 5.2 where we propose a mixed signal topology for analogue domain classification.

5.1 Principles for time domain processing

There are two driving factors to approaching time domain concepts where signals are represented in terms of delays between pulse edges or phase components in oscillators. The first benifit is the inherent digital operation where continuous valued signals are represented by digital events with respect to a global or local reference [170]. This implies that the typical analogue processing has the same power scaling and advantages as the digital processing in terms of technology parameters. This allows oscillator structures to approach very efficient operation irrespective of the oscillation frequency or supply voltage [171]. The second is that many operations are not restrained by non-linearity from individual transistors giving way to ideal integrators and other operators [172]. The overall result is that even with limited power budgets the topologies have an overwhelming excess in bandwidth where performance can scale with digital gate delay or its switching energy. The abundance of digital operations for such systems allows these topologies have the potential for digital synthesis using standard cells and a digital design flow to directly process analogue signals [173]. Moreover event based representation of continuous valued signals allows for often a surprisingly efficient implementation with reduced complexity for a variety of elementary operations. For example [174] presents a clockless PVT invariant true random number generator based on the collapse of a ring oscillator



(a) Minimum V_{DD} For CMOS amplifiers from [112].



(b) Linearity constrained power efficiency for Class-A Amplifiers.

Figure 5.1: Voltage supply relationship with respect to the bandwidth and linearity requirements with respect to different technologies

structure.

Let us elaborate on the notion of scaling analogue with digital characteristics quantitatively. Figure 5.1 illustrates the drawback of conventional analogue techniques from first principles by looking more closely at the voltage scaling characteristics. Here transitioning to nanometre technologies gives us the capability of reducing our voltage supply because the desired bandwidth can be achieved with a smaller inversion coefficient or equivalent gate voltage. However the transconductance and consequently linearity and noise efficiency can degrade as the drain voltage is reduced. This dependency is because transistor gain requires a large channel resistance which is a function of $(1-e^{-V_{DS}/U_T})$ in addition to any DIBL which introduce a asymptotic limit where the former is in fact not process dependent [176]. This limits the output swing V_{max} with a overhead that is $5\,U_T$ [119]. Figure 5.1 b) demonstrates the resulting the class-A power efficiency measured as V_{max}/V_{DD} to reflect how efficiently we can use the provided voltage supply. This is evaluated in terms of P_{out}/P_{vdd} where P_{out} and P_{vdd} are the output signal power and the power dissipated by the voltage supply respectively. We can conclude that conventional low noise amplification structures can no longer benefit from technology scaling unless we adopt topologies that do not rely on amplification in the voltage domain. Because the input referred noise of a circuit relies only on its current dissipation scaling supply voltages remains a viable means to reduce power if time-domain structures can mitigate the need for voltage gain.

5.1.1 Sub-threshold Ring Oscillators

The understanding and the interpretation of principle elements for a given modality has the most influential impact on how well it can be utilised. Moreover encoding signals in the time domain will influence how flexibly certain objectives are approached either using digital operations or analogue feedback. Here we will review some basic understanding for ring oscillator structures that are biased in weak inversion. This component will provide a fundamental basis for the topologies proposed here because it connects analogue signals at the input to phase and time domain signals at its output. The interest here specifically lies with using current controlled oscillators that have a well defined linear relationship associated with any injected charge and the resulting shift in output phase. This will lead to the small signal transfer function that relates to the biasing current of the oscillator. Moreover we must be able to evaluate the different components of phase noise and refer it back to the input of transconductive element because we will apply this structure to instrumentation.

$$V_{out}(t) = A(t) \cdot f\left[\,\omega_0\, t + \phi(t)\,\right] \tag{5.1}$$

A generalized time dependent model for an oscillator is represented by Equation 5.1 where A and ϕ represent the amplitude and phase state variables of the system. f describes the limit cycle of the oscillator over time that maps the steady state output voltage V_{out} as function of phase. The challenge for sub-threshold current biased ring oscillators is that the non-linearity in f is difficult to analytically predict without well informed priori. This is of significance as it will determine how noise sources perturb coupled to the output phase state.

Many principle aspects of phase dependencies in oscillators have been well described in a generalized form using numerical methods [178] and by approximation [177]. The underlying characteristics however are illustrated in Figure 5.2 where charge perturbations integrate on to the phase of state the oscillator with respect to the impulse sensitivity function (ISF) $\Gamma(x)$. This factor is a cyclo stationary function that describes how the coupling changes as a function of the phase state ϕ subject to the source of perturbation. Moreover this allows us to predict

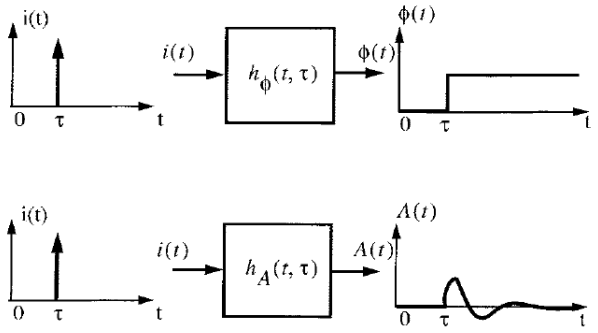


Figure 5.2: Model impulse functions for $\phi(t)$ and $A(t)$ from [177].

the accumulated phase noise due to a time varying process according to Equation 5.2.

$$\phi\,(t) = \int_{-\infty}^{\infty} h_\phi(t,\tau)\, i(\tau)\, d\tau = \int_{-\infty}^{t} \Gamma(\omega_0\,\tau)\, i(\tau)\, d\tau \tag{5.2}$$

The integral dependency on accumulated phase is what leads to the infinite open loop gain for oscillator based amplifiers. This also implies that any white noise source that is incoherent with the oscillator fundamental frequency will translate to the output phase as Γ_{rms}. This depends on the assertion that incoherence implies uncorrelated which is subject to the beat frequencies of the two sources. In practical cases this is a fair approximation not only because the oscillator frequency drifts freely but also because we explicitly consider closed loop implementations that aggressively shape in band perturbations. The utility of Equation 5.2 lies with its ability to predict the single-side band output noise spectrum due to a white noise current source with spectral density $i_n^2/\Delta f$ with carrier off-set frequency f_{off} according to Equation 5.3 [179].

$$L(f_{off}) = \left(\frac{\Gamma_{rms}}{2\pi\, f_{off}}\right)^2 \cdot \frac{i_n^2/\Delta f}{2} \tag{5.3}$$

The N stage ring oscillator structure of interest is illustrated in Figure 5.3. Opposed to voltage controlled structures this configuration is current biased and the oscillating ring is isolated from the supplies. Here the oscillating frequency in sub-threshold operation can be approximated as $f_0 = I_B/(N\, C_{gate}\, V_{RS})$ where I_B is the biasing current and C_{gate} is the input capacitance of the delay element. V_{RS} is the voltage across the oscillating structure and is evaluated using

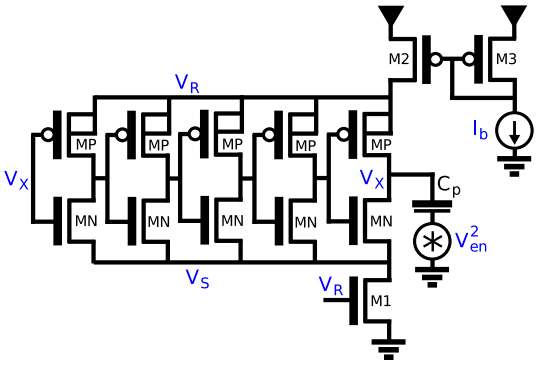


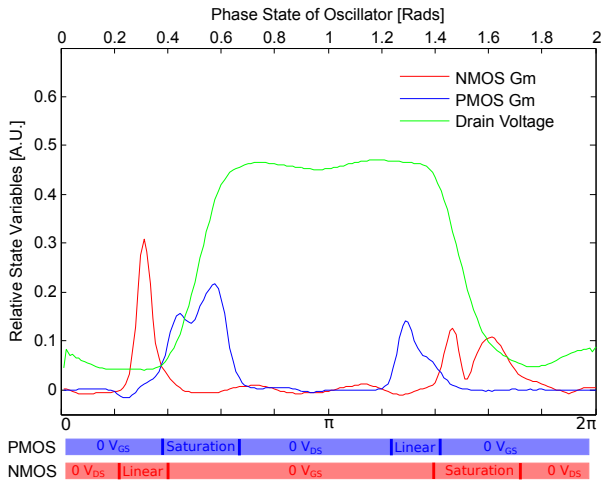
Figure 5.3: Schematic of current regulated ring oscillator with capacitively couple noise source.

Equation 5.4 where V_{th} is the transistor threshold voltage. In this case M_2 provides a biasing current from the PMOS side and M_1 if designed appropriately will allow isolation from the ground supply. This is particularly useful in differential configurations where capacitance on a common V_R or V_S can minimize high frequency noise from coupling directly to the differential phase component through the common mode feedback.

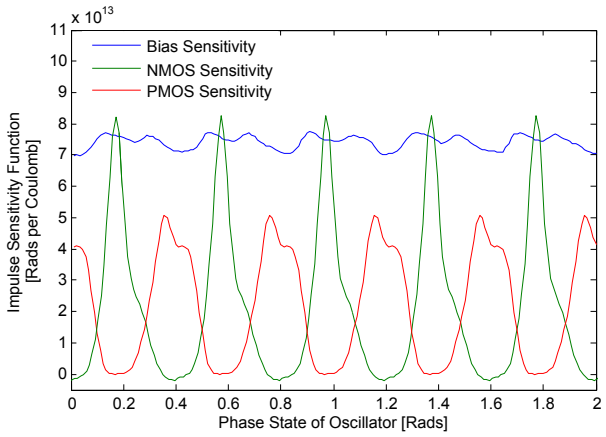
$$V_{RS} = V_{th} + \eta U_T \ln\left(\frac{2I_B}{2\eta\, U_T^2\, \mu\, C_{ox}}\,\frac{L}{W}\right) \tag{5.4}$$

The defining characteristic of the current biased oscillator is that the conduction of the NMOS and PMOS devices in each delay element is strictly non-overlapping. This is different when compared to oscillators biased in strong inversion and implies maximum current efficiency in a large signal sense. In addition it leads to the respective NMOS and PMOS ISF being non-negative. Thus the focus should lie with optimizing its rms value by balancing pull-up pull-down conductance. In fact we can empirically demonstrate that despite the intricacies of non-linear phenomena a current starved ring oscillator presents a significantly superior noise excess factor when compared to that of a transistor biased with the same weak inversion conditions due to change being retained in the high impedance nodes.

Figure 5.4 exemplifies the challenge of being able to predict internal parameter dependency analytically. Specifically in **a)** where the NMOS and PMOS of a single delay slice is evaluated both the saturation and linear conduction phases contribute towards accumulated phase noise. It is indicative to note that the bias transistor has a near uniform ISF equal to $2\pi/q_{max}$ inde-



(a) Simulation results illustrating the decomposition of transconductance as a function of phase for a single delay cell



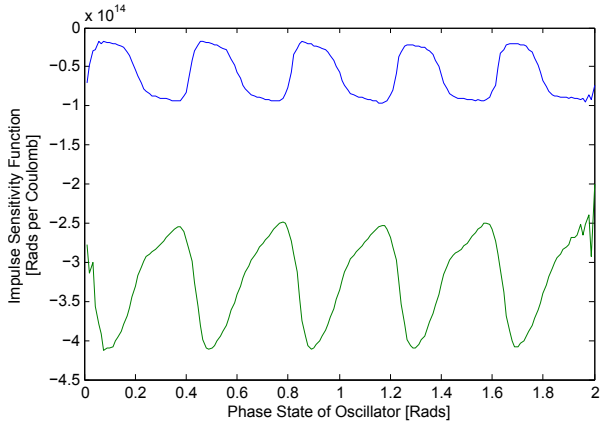
(b) Simulation results contracting the ISF of M2 to the aggregate ISF of PMOS and NMOS delay components

Figure 5.4: Simulation results outlining the dependency of parameter dynamics as a function of oscillator phase

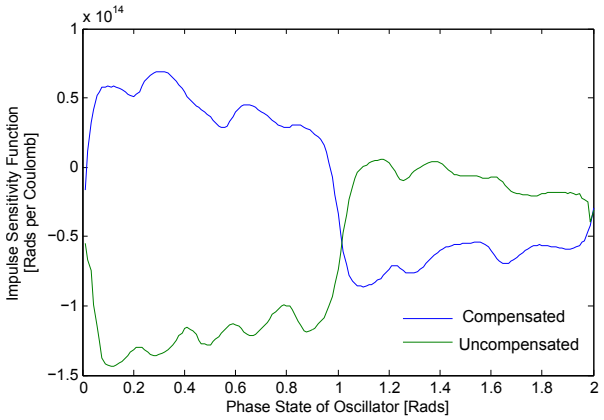
pendent of phase state as expected from the linear phase to charge relation. Here q_{max} simply represents the total charge dissipated by the ring oscilator each cycle which is $2N\,V_{RS}\,C_{gate}$. In particular this phase independent sensitivity is surprisingly independent of oscillator configuration in terms of number of stages and delay cell input capacitance. Instead the characteristic relies on the capacitance and channel resistance seen at the drain of M2 such that increasing impedance improves linearity.

When the aggregate contribution of all delay elements is taken into account as well as the increased noise excess factor in the linear region the ISF in **b)** appears predictable when normalized to that of M2. One may expect that the aggregate ISF of the ring oscillator to exceed the sensitivity to that of M2 as its contribution should have a similar profile and more noisy elements are involved. However the soft-switching of each delay element filters out a significant component of injected noise in addition to the fact that the nodes V_R and V_S retain accumulated current noise that feedback on the following stage.

Insight to optimizing the oscillator consideration is drawn from considering the lossy integration phases on V_X. Specifically as the transistors M1 and M2 present high impedance when considering the injection of charge or integration of a noisy current. We can infer that resulting voltage fluctuations are either one of two cases; coupled to V_R or V_S through a transistor in



(a) Simulation result of the ISF when a noise source is capacitively coupled equally to all switching nodes



(b) Simulation result of the ISF when a noise source is capacitively coupled only to node V_X

Figure 5.5: The compensation effect of M1 on the ISF for capacitively couple noise sources with reference to Figure 5.3

the linear region, or coupled to the switching capacitance during a transition. Rejection of the former will rely on increasing q_{max} and minimizing coupling factors as the ISF is equivalent to that of the bias current.

It is well known that the dominant factor of noise in ring oscillators comes from supply variations that are capacitively coupled as illustrated in Figure 5.3. This represents the coupling expected from substrate noise and supply noise that is not generated by the transistors them selves. The impact of introducing M1 opposed to grounding V_T is shown by Figure 5.5 with a dramatic improvement in ISF characteristics. Moreover large drain resistance of M1 allows the peak to peak ISF to be adjusted by exploiting the dynamics previously discussed. On that note it is important to realize that unlike Gm-C differential implementations the rejection of common mode signals is not present due to the coupling dependency of on phase. The matching/minimization of these factors can still allow a considerable improvement towards performance in practice but the process of optimization is challenging due to the fact that these components can not be well predicted as a priori. More generally incoherent perturbations in differential implementations will scale with $(\Gamma_{rms} - \Gamma_{dc})^2$.

It may be obvious that there no high impedance analogue nodes in this configuration that could introduce undesirable poles. But more importantly we do not need to provide extra voltage headroom or a second gain stage to let our output signal vary with maximum amplitude. In

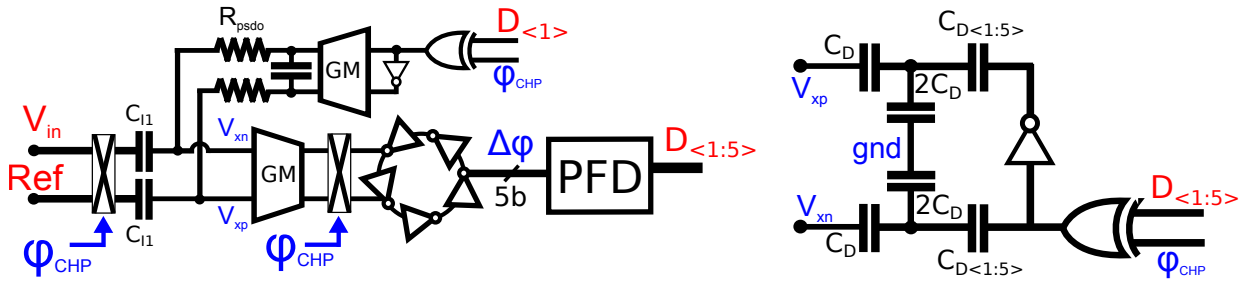
this case the oscillator mostly reuses the V_{SR} voltage headroom. This raises an interesting question; what limits the required voltage headroom for this circuit? Typically the complementary structure necessitates that the source drain voltage of the current bias transistors and differential pairs is sufficient to provide good channel resistance. However there is another component with regard to the noise generated by the oscillator that should be considered in terms of the oscillator voltage overhead V_{RS}. This leads us to evaluate the dependency on sampling noise with respect to the loading capacitor of each delay cell. Considering that the scaling the technology can result in a higher oscillator frequencies and equivalently using a small loading capacitance for the same power budget. It is important to realize that it is a charge induced as sampling noise on each capacitor before each up/down transition as residue from the previous cycle. This sampling noise can be referred to the input of M2 which leads to the expression in Equation 5.5.

$$v_{smp}^2 = \frac{1}{Gm_{M2}^2} \cdot \underbrace{2N\, f_{osc}^2\, kT\, C_{gate}}_{Noise\,power} \tag{5.5}$$

As Equation 5.5 suggests this noisy charge injection occurs for every transition in a delay element which is $2N$ times per period. When we expand this expression in terms of the oscillator power dissipation we can show its underlying dependency in Equation 5.6.

$$v_{smp}^2 = \frac{4kT}{P_{osc}} \cdot (\eta\, U_T)^2 \tag{5.6}$$

Now it should be clear from Equation 5.6 that this contribution only depends on the total power dissipation of the oscillator P_{osc}. This profound result confirms that without considering band-limiting factors all transistor generated noise densities are in fact independent of the frequency or total capacitance when referred to the gate of the biasing transistor. Following our expectation is that the dominant factor for noise is the total biasing current of the structure which is fundamentally identical to that of an conventional amplifier.



(a) Open-loop structure of the chopper stabilized time domain instrumentation circuit.

(b) Capacitive DAC feedback structure used for linearisation.

Figure 5.6: Time domain instrumentation topology for low noise voltage to time-domain conversion.

5.1.2 Time Domain Sensor interface

A principle element to these systems is associated with achieving effective conversion from continuous analogue signals to time encoded binary signals without distortion or excess signal corruption. It is typical to see the removal of VCO non-linearity though LMS post-processing [180] however this level of in-channel DSP can also be avoided through feedback utilizing the linearity of passive components. Our endeavour here lies with applying the discussion and topology selection in Section 3.2.4 to VCO based structures that follow closely to our optimization methodology. We suggest thinking of the oscillator's phase as an analogue memory that represents the state variable of the system which we can freely adjust by injecting charge.

This approach is different from that currently seen in the literature for time-domain based instrumentation of low frequency signals. The time domain encoding concept is predominantly used in asynchronous ADCs that aim to avoid quantization noise from being introduced [181] [182]. There is some motivation here to approach a neuromorphic amplifier topology that generates tokens with time-domain events that encode the input signal intensity [183]. Many of these structures leverage signal dependent power dissipation that reduces as the input signal varies more slowly. However they are typically open-loop topologies to avoid a complicated feedback DAC where events are generated upon asynchronous level crossings that reset internal integration nodes or toggle the reference voltages. Linearity and dynamic range can become difficult to achieve while maintaining aggressive power efficiency because resetting integrators or changing references are large signal discontinuities.

The proposed implementation shown in Figure 5.6. This structure opts for a direct conversion of analogue to phase domain signals by relying on the integration to filter out oscillator harmonics present in the feedback signal. Abstractly the topology is seen as a ideal integrator with integration factor $\frac{Gm}{q_{max}}$ proceeded with a non-linear element that introduces spurs around N times the oscillator frequency when feeding back. Here N is the number of taps in the ring oscillator used to simultaneously evaluate the phase difference of the differential structure. This allows us to freely adjust N for improving Γ through increasing q_{max} without sacrificing the ability to suppress the harmonics. Since the signals at the output of the phase frequency detector represent the phase difference between the two oscillators is full scale. The capacitive network need to scale down by a relatively large factor to assure V_x does not exceed the linear range of the transconductor and is implemented using a capacitance area reduction technique [184]. When the closed loop gain is large however this concern can be dismissed since the quantization levels scale with $\frac{V_{DD}}{A_{cl}\,N}$ which will typically be the same order of magnitude as the input signal.

While we are free to adjust the transconductance for noise requirements there is a limitation to the increase the complexity resulting from the capacitive feedback DAC and parallel digital phase processing. Because digital power dissipation scales with $N f_{osc}$ which is bounded by I_B it is independent of N for a fixed capacitive load in the oscillator delay cell. In fact increasing N reduces the total power of the oscillator harmonics as we effectively increase the number of quantization levels. This can be seen at the output of the capacitive DAC but this aspect will not be evident with respect to the processing performed in the time domain.

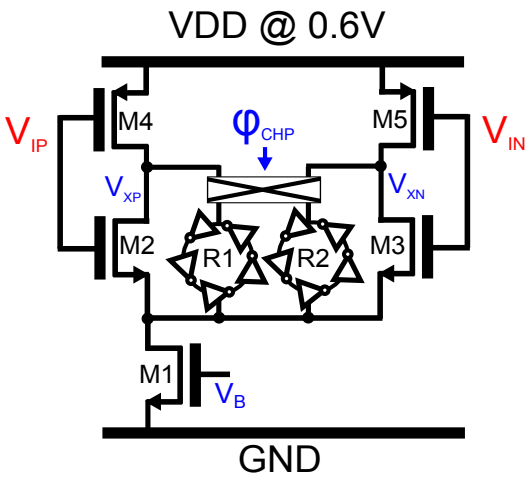
Note that when using ring oscillators with large number of stages in order to reduce leakage and non-linearities in the limit cycle to some extent we can retain a small factor of N by sub-sampling the output taps of the structure. This does require an integer ratio between the total number of stages and N in order to position the harmonics beyond $N\,f_{osc}$. Also consider that relation between the phases of the oscillator will imply a specific frequency shaping and harmonic modulation at high frequencies [185].

The primary design criteria for the phase detector structure and it respective time domain

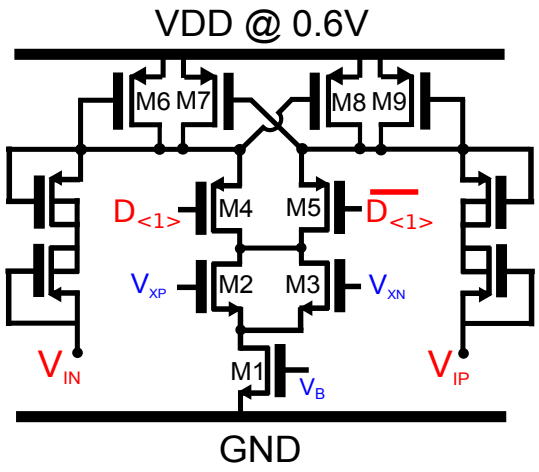
encoding should be related to maximizing power-bandwidth efficiency of digital cells. This is because the time-domain characteristics of the detector could introduce a inverse relation with regard to signal level and required logic gate bandwidth. Using conventional $1.5\,b$ encoding with up/down signals for example would give rise to this unwanted discontinuity. This is because the encoding scheme will generate narrower pulses for smaller signals that require exceedingly more bandwidth to process and feed into the time domain memory. It is conceivable that if this bandwidth is insufficient a dead-zone is introduced that is characteristically similar to class-B amplifiers.

Using a single bit representation that results from a XOR phase detector inverts this problem such that for small error the minimum bandwidth is required that successively increases as the loop error increases. In extension any asymmetric switching & delays in driving the capacitive feedback that is expected from process variation exacerbates any capacitive mismatch in the different phases of the feedback additively. These components primarily excite distortion on the output depending on the ratio of gate delay to oscillation period. Here Chopping the input will remove off-set and mismatch related components to a certain extent by up modulating them.

The motivation for using the single stage structure or allocating all the gain to the first stage is also associated with how the supply noise couples to the signal. In this respect we suggest that this structure should be thought of equivalent to that of a ADC. Particularly with respect to the digital feedback where providing asymmetric feedback implies that supply noise coupling can not be cancelled out. In addition capacitive mismatch between the positive and negative branches will also contribute to supply noise coupling. As since supply noise sources couple to the output of the amplifier while providing the maximum closed loop gain minimizes the input referred component. It should be noted that this type of supply sensitivity and capacitive mismatch is equivalent to that found in analogue to digital converters hence this drawback is only with reference to an all analogue solution. Further more once our signal has been encoded in the time domain which we expect to exhibit improved resilience to supply noise because its influence is proportional to the gate delay of the technology used.



(a) VCO element that feeds two oscillators for integration in the phase domain with current recycling. Transistor sizing is detailed in Figure C.10.



(b) Current DAC based psuedo-resistive feedback for chopper stabilization. Transistor sizing is detailed in Figure C.11.

Figure 5.7: Transistor level implementation of the phase domain integrator structure with phase detector feedback.

The schematic implementation of the VCO is show in Figure 5.7 which is derived from the complementary amplifier structure used in prior work. The fact that both ring oscillators are isolated from the supplies and floating in the middle of the rails presents an improved ISF as well as assuring the buffer that amplifies the clock phases to the full scale is guaranteed to be centred around the switching point of a balanced inverter. The most crucial component for effective operation however lies with the sizing of the input NMOS M2 with respect to loading ring oscillator. The DC operating point M2 and R1 will present an load equivalent to that of a diode connected transistor. If the delay element is balanced the current bias of the oscillator is evaluated with K_{M2} and K_N representing the W/L ratio of transistors M2 and the NMOS in the delay cell respectively.

Figure 5.8 clarifies the principle operation of this topology. We can see that as two currents are being integrated on the differential oscillator a phase shift will start to emerge when the two waveforms are compared. This phase difference on node $\Delta\phi$ represents our system output where the signal is encoded in the pulse width of the digital signal. This signal is applied to the capacitor array for feedback.

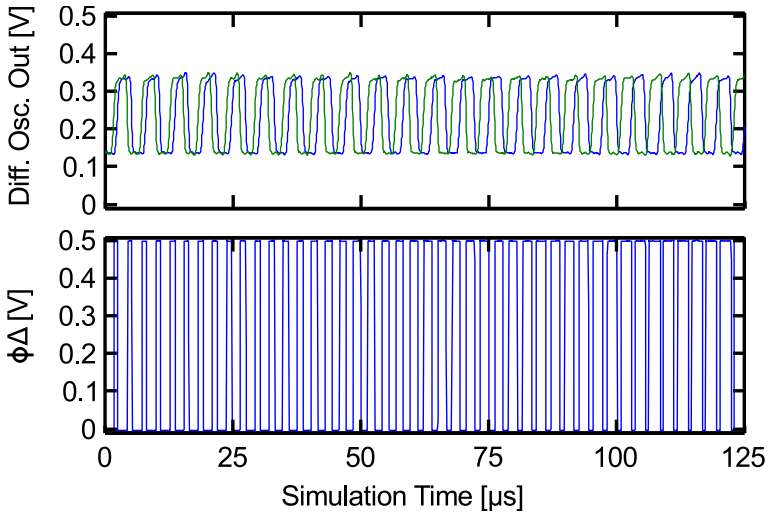


Figure 5.8: Simulated transient behaviour of the differential oscillator and the generated digital output.

$$f_{osc} \approx \frac{\alpha\, I_{M1}}{N\, C_{gate}\, V_{th}} \qquad \text{where} \qquad \alpha = \frac{K_{M2}}{2\, K_N} \tag{5.7}$$

The factor α in Equation 5.7 dominates the noise performance when referred to the input which would ideally approach the NEF of that without the oscillator. Similarly the corner frequency of the oscillator flicker noise which is not rejected by the chopper scales with this factor. It follows that the transistor length of the oscillator has a strong relation with respect to $f_{cor} \propto 1/L^2$. Fortunately it is easy to diminish this contribution as only a small bias is needed to result in a oscillation frequency several orders outside the signal bandwidth.

$$H_{sys}(s) \approx \frac{\eta\, f_{osc}}{s\, U_T} \cdot \frac{2-\alpha}{\alpha}\, N \quad \textit{and} \quad f \approx \frac{C_I}{C_D} \cdot (N+2) \tag{5.8}$$

The overall open loop system characteristics H_{sys} evaluated in Equation 5.8. This reflects the single pole nature of the topology that scales with the oscillator frequency and the number of phases taped out as one may expect. Notably the capacitive feedback structure used can represent a very small feedback factor $1/f$ without excess input capacitance that accommodates a large number of oscillator taps [186]. Evaluating the low pass 3dB point of the system which reveals a dependency as shown in Equation 5.9.

$$f_{3dB} = \frac{\eta\, f_{osc}}{\alpha\, U_T} \cdot \frac{N}{N+2} \cdot \frac{C_D}{C_I} \tag{5.9}$$

This expression is primarily dominated by the oscillator frequency which even for a small bias current can result in a considerable bandwidth. Although this is partially expected due to the fact there is no explicit load capacitance it also illustrates the benefits in FOM that can be achieved with this configuration of current-mode time domain architecture. There is a instinctive concern for the stability of the system as a result of the excessive bandwidth driven by maximizing efficiency. The non dominant poles introduced in the voltage domain is due to the parasitic capacitance on node V_Q typicall will not compromise stability due to coupling to the input of the transconductance at higher frequencies. The non-dominant pole on the time domain is introduced by any delay t_d from the VCO to the output buffers of the PFD as $e^{-j\omega t_d}$. This component can be more restrictive for small loop gain as it does not scale with the power of the input stage but with the supply voltage.

The voltage requirement of this structure is improved by biasing the NWell of the PMOS peudo differential pair to V_{XN} & V_{XP} in a cross coupled fashion to reject the differential lading component. The forward biasing reduces the threshold voltage of the devices allowing a supply voltage down to $0.6\,V$ without any considerable impact from leakage currents. This configuration also implies that the common mode at V_X is well regulated by the body transconductance of M4 & M5 rejecting common mode input fluctuation. The main voltage requirement actually comes from the switches of the chopper that feeds the ring oscillator that need good on-resistance to prevent noise injection which implies a minimum voltage of approximately $2V_{th}+V_{ov}$. Back-gate biasing will allow us to reduce the impact of V_{th}.

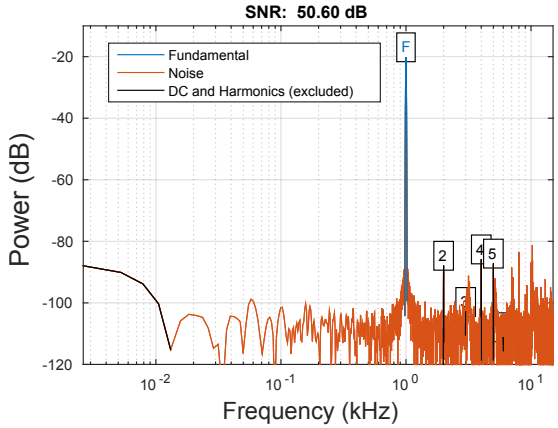
The psuedo-resistive feedback structure in Figure 5.7 b) extracts the signal component from up modulated aggressors using a current DAC which is resistively coupled to the input to close the loop. This allows us to feed back the full swing digital signals to cancel a DC off-set and sets the input common mode by matching the cross coupled transistor with the input pair. This primarily prevents having to use a cascaded resistor structure in order to deal with the large voltage swing on the output that can significantly degrade performance. While stability

is trivialized by the capacitive feed forward signal that grantees stability [187], it is important to note the design choice associated with the two poles in this feedback loop. One pole lies at the input of the complementary pair associated with C_{fb} and the other is at the gate of the cross-coupled pair C_x.

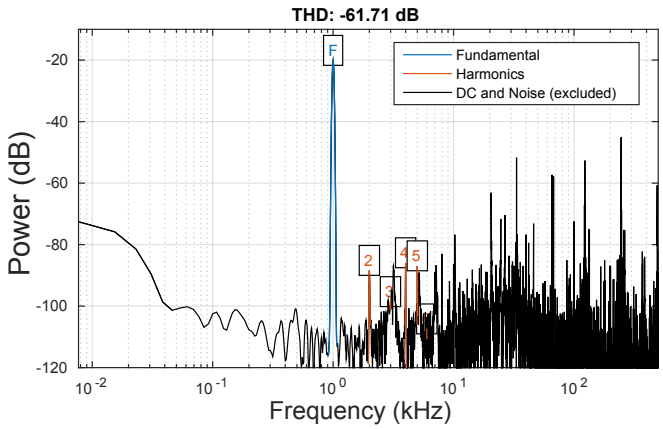
$$\tau_{hp1}(s) \approx C_{fb} \cdot R_{psudo} \qquad \text{and} \qquad \tau_{hp2}(s) \approx C_x \cdot \frac{\eta\, U_T}{I_{M1}} \cdot \frac{W_{M6} + W_{M7}}{W_{M6} - W_{M7}} \tag{5.10}$$

Equation 5.10 described the dependency of the two time constants in addition to the capacitive feedback. The reduction in capacitance of the feedback network implies that the high pass filter needs careful design in terms of the resulting pole location as the noise expected from the psuedo resistor will appear increasingly wider band as we try to reduce the total capacitance. Here we allow the second pole of C_x to approach DC by having $W_{M6} \approx W_{M7}$ resulting in a integration node. This means that the noise around the chopper frequency is strongly related to amount of capacitance we can allocate to C_x and the 1/f agressors are now shaped by the VCO integrator and this capacitance. The bias of I_{M1} in the current DAC should be adjusted to set the pole location close to but smaller than the chopper fundamental similar to the conventional design approach.

When we compare this structure to the conventional topology we realize a number of significant advantages. Primarily the inversion coefficient of the transistors is not bound like in a complementary input stage where the V_{GS} voltage for both the NMOS and PMOS has to be sufficiently large to allow the drain voltage to fluctuate by several $100\,mV$. This is particularly significant because the minimum feature size is inverse proportional to the optimal inversion coefficient confirming again that conventional means to not work at nano meter technologies. Here the threshold voltage can be arbitrarily small and we still retain a topology that is independent of supply voltage in the sense that it is strictly current biased. This will lead to improving the tolerance towards wafer level variations of the threshold voltage and carrier mobility which many sub $1\,V$ structures do not have. Similarly this implies class-A type power dissipation that minimizes switching current seen at the analogue supplies.



(a) Spectral power characteristic showing the in band unfiltered white noise floor and suppressed flicker noise component.



(b) Spectral power density of output band oscilator harmonics and modulation components where $f_{osc} = 250\,kHz$ and $f_{chop} = 66\,kHz$.

Figure 5.9: Transient noise simulation result of the 180nm CMOS time domain instrumentation topology with a $6\,mV$ peak to peak sine input at $1\,kHz$.

The excess in bandwidth from the VCO despite operating with a very small inversion coefficient has enabled us to achieve both $40-50\,dB$ closed loop gain while still retaining excess loop gain that easily exceeds $30\,dB$. This excess loop gain in the signal band is facilitated by the near ideal VCO integration of this topology that shapes a number of external noise sources and nonidealities. In particular technology scaling allows us to minimizes the noise gain due to the input capacitance C_g according to the expression $1 + C_g/C_{in} + N/A_{cl}$ [63]. Hence the VCO topology can allow a reduction for the input capacitance by a very significant factor relating to an impedance enhancement that scales with technology.

A Transient noise simulation performance is shown in Figure 5.9. This demonstrates that the dependency on nonlinearity is mainly due DAC mismatch components which are modulated the oscillator frequency spurs. Secondly the noise-floor and corner frequency characteristics follow closely to analytic predictions. In addition for the same current bias as a conventional implementation the structure can achieve an equivalent noise floor but at a reduced voltage overhead. Noticeably in the full spectrum there is a considerable amount of harmonics out side of the band induced by the chopped and oscillator aggressors. These components will need to be filtered out in order to approach a $60\,dB$ signal to noise ratio. Interestingly there is an observable gain in noise floor as we approach the point where there is no excess loop gain. Note that the spurious free dynamic range of this structure almost exceeds that of the structure

used in Section 3.2.4 by a factor of 10 for the same power budget due to the increase input range.

5.1.3 Time Domain analogue filter

Now that we have addressed the aspects of achieving low noise and linear instrumentation we must proceed to address the mechanisms for filtering to implement a band limiting characteristic necessary for the processing algorithms. There is some diversity in the number of approaches used to filter time domain signals. Most notably the continuous-time FIR based structure that represents a number low power characteristics that scales well with technology without sampling or clocking [188] [189]. However similarly to conventional FIR structures it is limited to signals where the frequency dynamic range is small in order to keep the filter order small. Other examples are found in PLL structures that lock using coherent phase domain signals which is inherently second order due to the analogue integration node which result in a loss of noise efficiency at low frequencies [190]. We do mention that VCO-based ADCs have been very successful in achieving efficient high-order noise shaping [191] [192].

It is important to realize that our proposed instrumentation topology converges on incoherent phase domain signals and neglects the modulation products through construction similar to that of asynchronous $\Delta\Sigma$ modulators [193]. Here we will take a similar approach to construct a first order phase domain integrator where the time-domain signals are also incoherent. Through simplicity the structure achieves a significantly better dynamic range and voltage scaling capability than its analogue domain counterpart. The premise will lie with our assumption that the intermodulation products of the incoherent frequencies are sufficiently out of band to allow construction of higher order filter structures.

The topology used for analogue filtering of time-domain signals is illustrated in Figure 5.10. This is based similarly on the phase difference of two ring oscillators that integrate a switched current which is generated by evaluating the difference in duty cycle with respect to the input and feedback. The logic simply advances or recedes the phase difference of the oscillator when there is an excess or lack pulse width when comparing the two inputs respectively. This

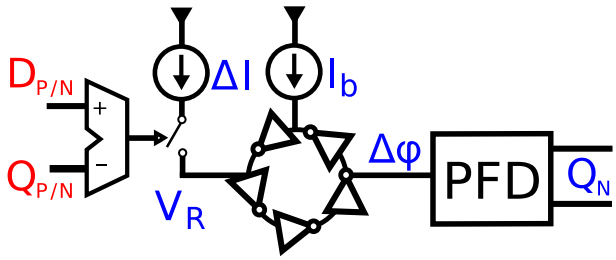


Figure 5.10: Closed loop time-domain analogue filtering structure

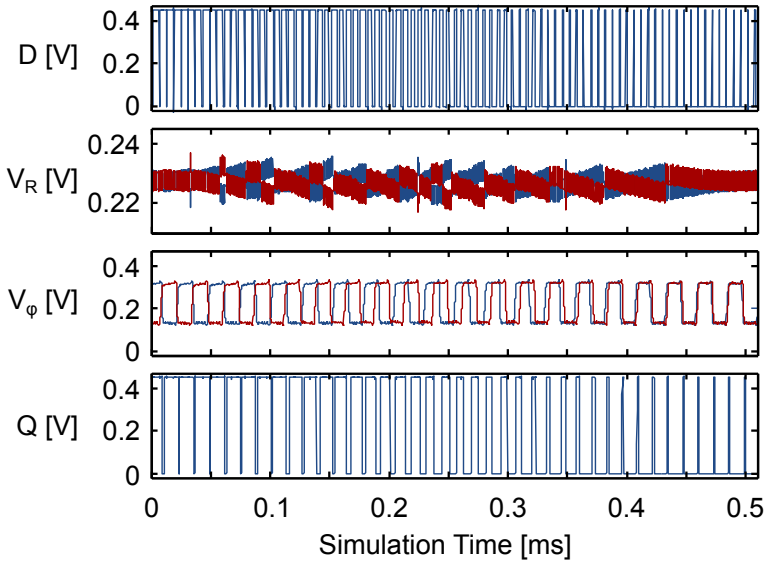
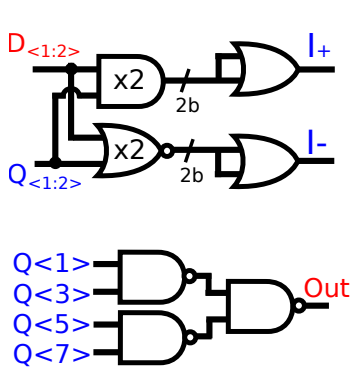


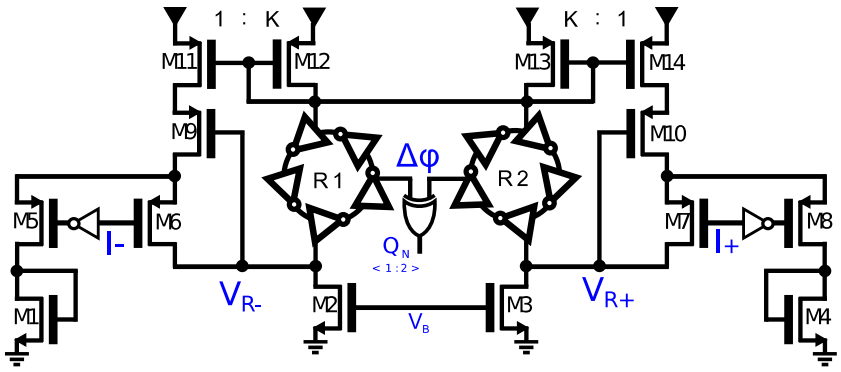
Figure 5.11: Simulated transient behaviour of the differential oscillator and the time encoded digital signals internal to the feedback loop.

behaviour is shown in Figure 5.11. The use of logic gates over avoids any drawback that arise from limited linearity and mismatch in the case of approaching this design with current mixing techniques. More over the efficiency of these operations allows miniaturized reconfiguration in the digital domain with a minimal analogue structure.

Most design considerations here are similar to that of conventional filters. We expect the analogue in-band noise components will scale with the biasing current I_b which will determine its input referred noise relative to the transconductance element ΔI. This however does represent a fundamental drawback since the charge pump transconductance element does not benefit from the sub-threshold slope gain factor. ΔI will typically be larger for the same bandwidth requirements by a factor of $1/U_T$. Although this factor is essential for achieving smaller cut off frequencies while maintaining large oscillation frequencies. The decreased noise efficiency is the fundamental drawback of using a digital logic instead the capacitive feedback network.

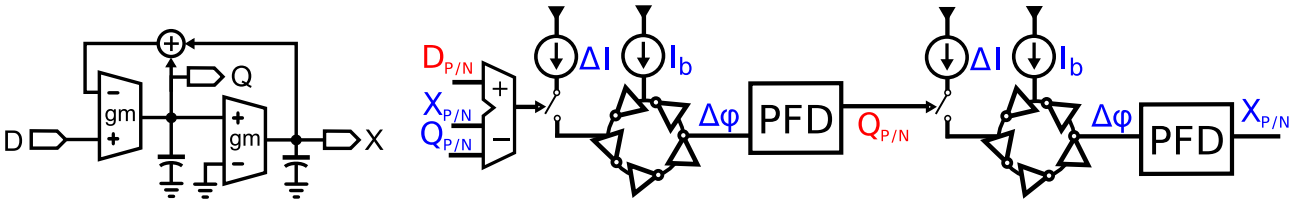


(a) Logic for subtracting and adding time domain signals.



(b) Differential oscillator structure with high-speed current injection DAC. Transistor sizing is detailed in Figure C.12.

Figure 5.12: Schematic sub-blocks for first-order time domain analogue filter



(a) Gm-C Equivilent.

(b) Time domain implementation.

Figure 5.13: Bandpass time-domain analogue filter which is cascaded to realize a 4^{th} order TD-BPF.

However in this scenario the processed signal will already be at full dynamic range with reduced noise requirements.

The gate level implementation is elaborated in Figure 5.12. The charge pump structure here uses a cascaded current source and dummy load for bandwidth improvement. This configuration is important because this operator precedes the integration and consequently has a substantial influence on off-set or distortion near the cut-off. The self referenced bias of the charge pump through M11-M14 should allow to good matching independent of the configuration in biasing transistors M2-M3. Similarly the noise figure is improved by sharing the drain voltage of M12-M13 as its noise is coupled to the common mode.

The two digital components in Figure 5.12(a) represent the subtraction for feedback and a gain factor G when the phases of Q are mapped to the output. The subtraction logic is determined by considering the XOR-PWM waveform as a two state input of ± 1 and similarly the DAC input states which in this case is $+1/0/-1$. The configuration of both components should

$$F = D - (Q + X)$$

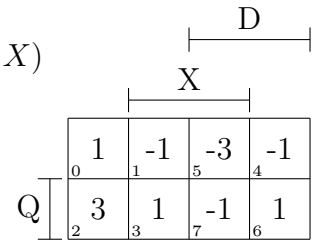


Figure 5.14: Karnaugh map associated with the subtracting XOR type PWM signals Q & X from D

compliment each other. This is exemplified when we consider the summing node for another case when two integrators are cascaded with both outputs fed back to achieve a bandpass response. This configuration is shown in Figure 5.13. The boolean operation required is shown in Figure 5.14 where four levels are needed to include carry signals. Here we compromise with two DAC structures with input states ± 1 and $+2/0/-2$.

The logic that sums the different phases relies on the coherence of it input. Because we know the different taps of the ring oscillator will not over lap with respect to certain signal range at the output we may isolate the components with the AND operation of two different phases to isolate small variations in pulse width and combine them. Analogous to a variable gain amplifier, if the signal variations exceed the section where the two phases overlap the output will saturate. The significance here is that gain achieved with this operation has arbitrary gain bandwidth product with negligible power dissipation. Once blocking or interfering signals have been removed we may give the signal reconfigurable gain by only using a handful of gates. It simply relies on increasing the number of oscillator taps from the previous stage while maintaining its feedback configuration which is independent of noise & linearity performance. In fact for a gain G on the PWM signal D we require $2G$ taps that are summed according to Equation 5.11.

$$Q = \bigcup_{k=0}^{G-1} \{D_{Rk} \cap D_{Rk+N/2-R/2}\} \qquad Where \qquad R = \left\lfloor \frac{N}{G} \right\rfloor \tag{5.11}$$

While AND & OR will perform equivalent operations that retain the oscillator phase difference Φ but subtract a signal independent component which the gate delay between the two phases

being operated on from the pulse width. Here Φ is normalized such that it represents 1 and 0 when the oscillator phase difference is π and 0 respectively. This implies that for a positive gate delay number of ΔT we will have an output off set by $Q = \Phi - \frac{\Delta T}{N}$. If an XOR gate is used we extract a signal independent component as $Q = 2 \cdot \frac{\Delta T}{N}$. Both statements will hold true as long as $\Phi < 1 - 2\Delta T$ which implies that the pulse section used for computation is signal independent. Using this rather simple construction of logic one may sum phases that are one radian apart with a $XNOR$ gate to realize the absolute value operator which exemplifies the rich utility of this time domain processing.

$$H_{sys}(s) = \frac{G}{1 + s/p_1} \quad \text{where} \quad p_1 = \frac{N}{q_{tot}} \cdot \Delta I = \frac{N\,\omega_{osc}}{K} \tag{5.12}$$

When the primitive topology in Figure 5.10 is analysed in the Laplace domain we can derive the Equation 5.12. This demonstrates a first order characteristic similar to the amplifier and has is a close relationship with the oscillation frequency and the filter bandwidth with the addition of the gain factor G.

This filter configuration here specifically designed for a $0.18\,\mu m$ process. Considering that digital filters will become more viable as the technology node decreases it should be acknowledged that the proposed time-domain filter structure will only be advantageous when frequency dynamic range is large and memory is limited. This is primarily because the kT/C relations inhibit very aggressive sizing in the oscillator structure particularly if no excess loop gain is available. We still require a large amount of energy storage q_{max} to prevent external noise sources from perturbing the output. We point out the proposed topology discussed here provides the means by which instrumentation can successfully scale with technology characteristics. Particularly as it is robust towards transistor non-linearity and imperfections. A large component for performance enhancement will rely on calibration components that improve the resilience of the capacitive feedback structure and filter parameters to allow miniaturization. While this specific time-domain topology will not allow the absolute minimum supply voltage this configuration does take advantage of transistor sub-threshold slope which implies a fundamentally superior noise performance that can also circumvent supply noise.

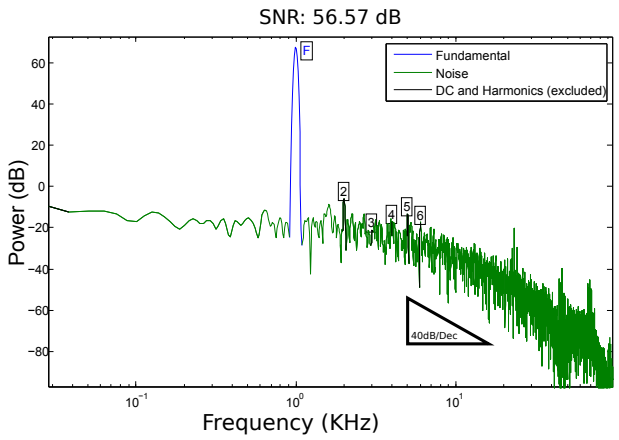


Figure 5.15: Transient noise simulation of proposed instrumentation amplifier with time-domain filter structure with a $6\,mV$ peak to peak sine input at $1\,kHz$

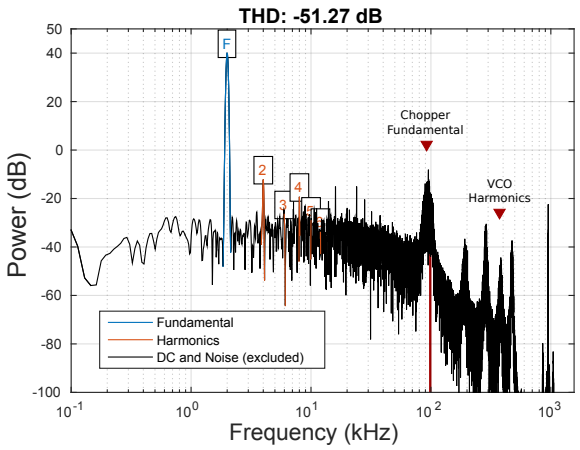


Figure 5.16: Transient noise simulation result of the proposed instrumentation topology in 65nm CMOS with $2\,mV$ peak to peak sine input at $2\,kHz$

As shown in Figure 5.15 both a low noise floor and band limiting behaviour is achieved. In particular we see a $40\,dB$ roll off in the noise floor at the $6\,KHz$ cut-off frequency from the 4^{th} order bandpass filter. While Table 5.1 reveals some similar performance characteristics as the conventional implementation in Section 3.2.4 which is the result of using the same optimization strategy. However this work is the first to consider NEF maximization for the design of time-domain circuits. As a result we are much more confident about the power efficiency for this implementation. As a reference this topology was also implemented using a 65nm CMOS process without filtering structure to confirm the scalability of this structure with the noise transient simulation shown in Figure5.16. The compact capacitive feedback may not allow linearity beyond 60dB but given the scalability and efficiency of this design there is a significant advantage over current state-of-the-art.

This brings us finally to finding a satisfactory answer to how the area-power product figure of merit limited or bounded in some sense. Section 3.4 argued that linearity and quantization was crucial constraint in the conventional structures which is not the case for the proposed structures. Instead we observe from Equation 5.5 we must dissipate a certain power level in the oscillator which we know is biased by a fractional current related to the input referred noise through α. As a result the current of the oscillator is fixed and thus when the voltage V_T is scaled down this sampling contribution will progressively become larger to the point that is

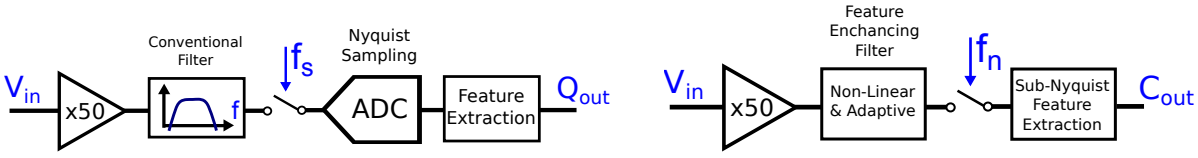
Table 5.1: Performance specifications of the proposed time domain instrumentation topology and those found in the literature. † based on transconductor size ⋆ Estimated with %65 utility.

Parameter	Units	This work		Chandrakasan [194]	Tsividis [195]	Markovic [180]
Modality		Time	Time	Voltage	Time	Time
Technology	[nm]	180	65	180	28	40
Supply Voltage	[V]	0.6	0.5	0.2 & 0.8	0.65	1.2
Total Current	[μA]	0.8	1.5	1.8	36.92	5.8
Bandwidth	[Hz]	375-6k	1-6k	1-1k	40M	$3-1.5k$
Filter Order		IIR 4^{th}	-	IIR 2^{nd}	FIR 8^{th}	-
Noise Floor	[$nV/\sqrt{Hz}$]	69.4	57.5	36	514	427
Noise Corner	[Hz]	< 10	< 10	0.5	-	< 1
SFDR	[dB]	58	54	50.4	30	78
Area	[mm^2]	$115 \times 100^{\star}$	$64 \times 69^{\star}$	800×775	72×45	280×360
NEF		1.18	1.22	2.1	8.13	12.9
Chopped Input Capacitance	[pF]	0.04	0.03	21.5	$0.01^{\dagger}$	-

it comparable to that of the thermal noise. This equality reveals the oscillator voltage should not approach $\alpha \cdot 2U_T$ or the NEF efficiency factor will degrade. We can assert that although at first it appeared that the sampling noise limited the minimum size of the instrumentation circuit here is limits the minimum power of the circuit in a much more explicit manner. The area requirement is more simply proportional to K_F/α reflecting the location of the flicker noise corner and the target oscillator frequency. There are some details remain with regard to choosing α which represents one degree of freedom for trading off oscillator area for minimum voltage but is also strongly depedent on the system bandwidth.

5.2 Analogue Signal Classification

Now our interests will be redirected towards the methods and hardware implementation of neural spike analogue classification in order to faithfully demonstrate that continuous time instrumentation can provide a substantial improvement over the conventional approach at the system level. In particular we demonstrate an unsupervised method that will allow the clas-



(a) Conventional Sampling Strategy (b) Sub-Nyquist Sampling Strategy

Figure 5.17: Illustration of Nyquist rate feature extraction and using feature enhancement in order to operate at sub-Nyquist rates.

sification of spikes without requiring signal quantization at any stage of the adaptive process with empirical results.

The abstract motivation here is illustrated in Figure 5.17. Utilizing digitized recordings as basis for feature extraction implies the necessity of operating with excessive data rates in order to capture the full bandwidth of features in the signal such that $f_s > 2\, f_{BW}$. By using mechanisms that enhance & extract prominent features directly in the analogue domain this sampling constraint is eliminated [141]. Instead we may sample at the rate of spikes present in the recording. Even by approximation we can assert that $f_n << f_s$. In some sense this motivation inspired by that of adaptive compressed sensing or sparse representation methods [196]. Here we will introduce a perspective based on realizing a less generalized method that can be integrated effortlessly which has not yet been attempted in the literature. The challenge for this approach is finding efficient analogue operators that allow direct feature extraction and more importantly feedback mechanisms that adaptively improve the feature extraction process without substantial resource requirements or supervision.

The notion that motivated this specific classification structure is that in order to improve alignment and thereby reducing how noise couples to features relatively high sampling rates are needed. This implies high temporal resolution for these spread-spectrum signals is desirable such an approach but has unavoidable implications on larger memory and power requirements. As we shall see analysing K features for M centroids with mixed signal methods will require $KM+K$ registers and $max(K, M)$ integrators where register depth has logarithmic dependency on temporal resolution. In contrast to PCA, template matching, and many all-digital methods where temporal resolution or window size is linearly proportional to register count. While it is still vital to align the analogue operators with the spike waveform, increased clock rate does

not influence the analogue power dissipation since signal quantization is not performed.

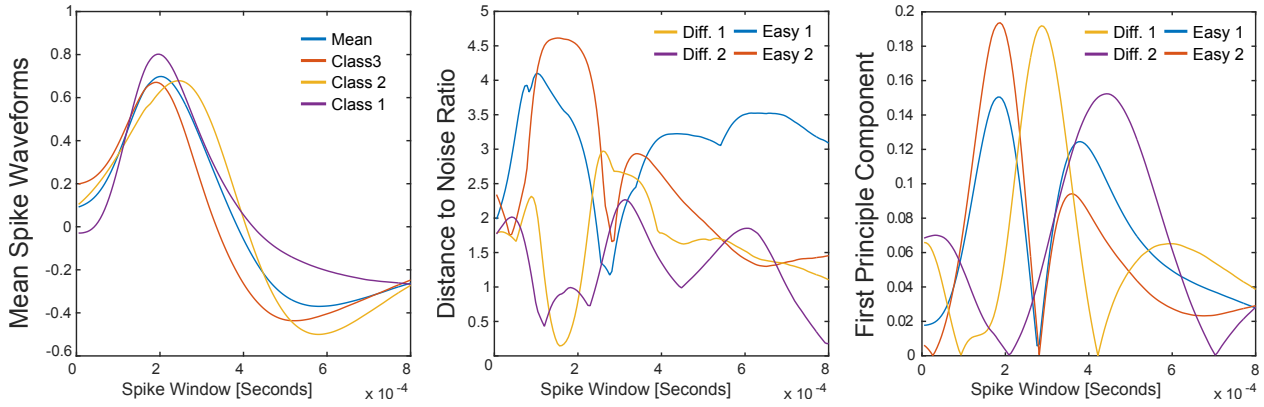
It is relatively rare to see analogue or mixed signal implementations for machine learning classifiers due to the convoluted impact circuit imperfections has of learning dynamics [197]. This makes it difficult to successfully realize more complex architectures but the methodology can significantly increase the information storage density at very low power budgets. This is exemplified by the system in [198] that not only achieve 1.04 GOPS/mW but also an area efficiency of 0.03 GOPS/mm^2 using a 130 nm technology. Here the continuous valued charge on floating gates was used to preserve learned features without quantization of errors.

5.2.1 Feature Selection

It should be expected that when a spike event it sampled at a relatively high rate we typically only need a select few samples once it is aligned in side a window in order to tell different classes apart. Only when noise becomes a considerable component must we consider multiple samples before we can make an accurate distinction. In such a scenario we would like to use the samples that maximally distinguish classes. That is, we would like to maximize the quality of our feature Q_F by maximizing for instance a simple sum of N maximally separating samples c in the window W with the linear distance operator D.

$$Q_F = \frac{\sum_{i=1}^{N} D(W[c_i])}{N} + \frac{en_{rms}}{\sqrt{N}} + \frac{BG(t)}{N} \tag{5.13}$$

The expression in Equation 5.13 primarily tells us that in the presence of white noise en_{rms} and background activity BG. Increasing the number of samples reduces the contribution of white noise by $\sqrt{N}$. However if the new samples have negligible signal quality our aggregate distance will decrease fractionally by $1/(N+1)$. This suggests that we should avoid using on a large ensemble of samples because it avoids complexity and may very well improve classification accuracy. Implementing an effective analogue solution requires the evaluation of optimal section in the spike window. To analyse this problem let us define a distance to noise ratio (DNR)



(a) Mean spike waveforms of difficult-2 aligned and detected using a narrow bandpass filter.

(b) Evaluation of sample quality in the spike window through separation to noise ratio.

(c) Magnitude of first principle component for each sample.

Figure 5.18: Illustration of feature dependency for windowed neural spikes for data sets from [, roman] with $-16/, dB$ background noise and $-20\, dB$ added white Gaussian noise.

with respect to the mean spikes of M classes and their standard deviation for each sample in the spike window as;

$$DNR[n] = \frac{\sum_{i=2}^{M} |\mu_0[n] - \mu_i[n]|}{\sqrt{\sum_{k=1}^{M} \sigma_k^2[n]}} \tag{5.14}$$

Where μ_0, μ_k,σ_k are the mean of all spike waveforms, mean of spike waveforms in class k and the aggregate standard deviation of class k respectively. In a more general form of choosing samples that maximize Equation 5.14 can be seen as classification by expectation maximization [199]. In Figure 5.18 we exemplify this quality factor for a number of training data sets along side the first principle component of each set respectively.

A number of observations can be made. In particular for a variety of spike shapes show PCA peaks in the first moment which are convex with respect the depolarizing and repolarization sections. In contrast to the DNR plot shows multiple local minima and thus optimization in this space is not considered trivial. What should be pointed out is that the peaks in the PCA curve will typically correspond to good DNR points as they will relate to sections of high variance due to maximum class separation on top of the noise variance. It should also be clear that our interest does not lie in the refractory period as the slow-wave component is small in magnitude and typically corrupted by high-pass filters that reject environmental interference.

Hence we inherently expect poor DNR in the latent refractory period of the recorded action potential. One of the more significant implications of not quantizing the signal is related to the fact that features can generally not be extracted before the detection event. This implies that the operator used for detection should avoid introducing group delay that could result in completely missing the most energetic features in the spiking event.

In actuality this hints at the contradictory advantage of analogue filter detection over an FIR filter equivalent. That is this group delay exclusive to IIR filters is highly dependent on the frequency content of the spike waveform. This can to some extent be observed in Figure 5.18 where class 2 is deliberately delayed with respect to the alignment point since it has a smaller derivative or equivalently less high frequency content. Here the alignment is a achieved by conditioning the signal with a narrow bandpass filter and looking for a peak above the threshold. While a relatively simple in implementation, if designed to maximize output signal to noise power it can be quite effective.

The approach to self mixing spike classes to with the sampling or alignment strategy is effective at improving features of already very similar spike shapes. In fact this alignment in some sense captures the features existent before the detection event and is mixed with latent features. As a result although analogue techniques are less effective at demonstrating resilience towards background noise they can efficiently mix different features to improve discrimination between spike shapes.

5.2.2 Mixed Signal Implementation

The adaptive method employed here is to section the spike waveform depending on the redundancy in the number of features required. Noting that features in the same polarizing phase will correlate strongly since they emerge from the same phenomena. Each section after the detection event is bounded as priori to grantee non-overlapping features (i.e. $0 - 150\,\mu s, 170 - 310\,\mu s$) and the maximum variance is found in each section respectively. These are assumed as optimal DNR features referred to as Ωs. Our classification will integrate around these points and

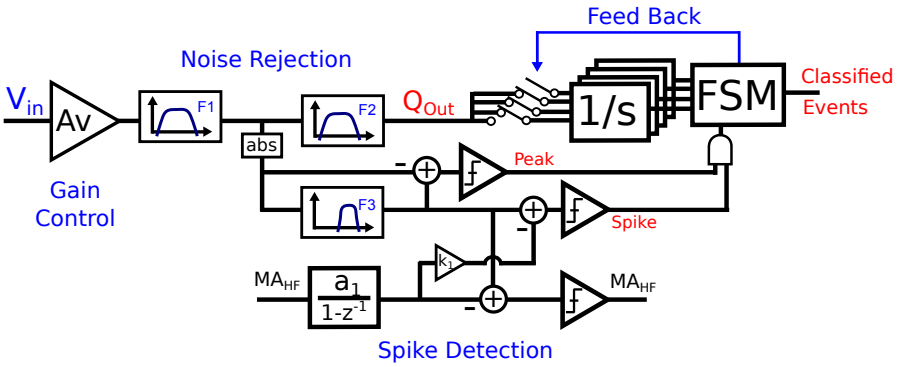


Figure 5.19: System abstraction showing the configuration for pre-filtering detection and reconfigurable integrators.

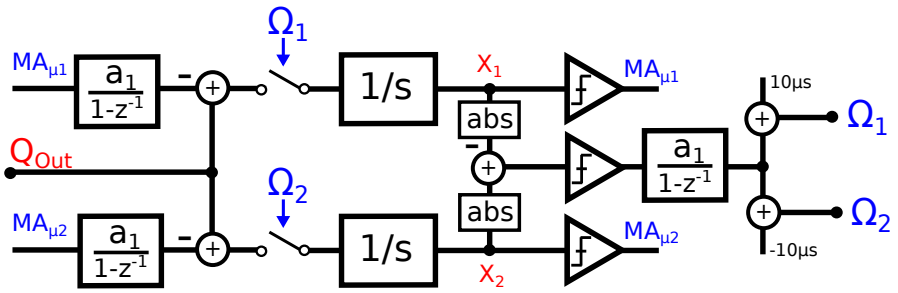


Figure 5.20: Illustration of closed loop control for finding point of maximum variance.

perform k-means clustering in the resulting space.

The topology is summarized in Figure 5.19 where the predominant active component lies with the state machine reconfiguring the integrators to optimize the classification. The noise rejection performed by three bandpass filters primarily removes out of band aggressors that prevent accurate classification. Notice the feedback loop for detection relies on a long term average of a narrow band component in addition to a indication of a localized peak above the threshold. Filters $F1$ & $F2$ are band pass filters with first order roll-off and a $0.4-7\,kHz$ bandwidth where as $F3$ is a narrow band of $2.5-4.5\,kHz$ to maximize group delay sensitivity. There is an additional advantage of pre-sectioning the waveform which is that these optimal points can be found independently in sequence. During initialization we seed the starting point in the middle of the section.

The control loop for implementing this method is shown in Figure 5.20. Here $\Omega_{1/2}$ represents the temporal off-set separated by $20\,\mu s$ with respect to the detection event around which the signal is integrated in the analogue domain. The digital integrators primarily average the aggregate

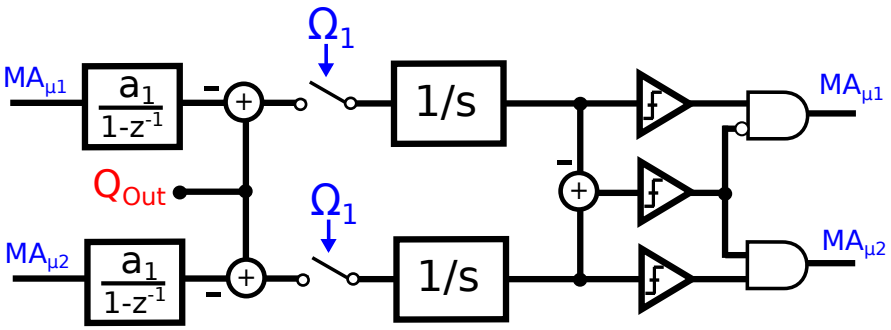


Figure 5.21: Illustration of closed loop control for tracking two-centroids with the Ω_1 feature.

statistics to assist long term convergence by tracking the mean value of both integration results. This allows us to compare the deviation from the mean for each spike event and evaluate which has a larger variance. Note a key characteristic is that signal being analysed is band limited to such an extent that we do not expect local minima within each separate section. Since the digital integrators are accumulating boolean results to reject the uncorrelated noise the factor a_1 represents the register depth of the counter. To be more precise the Ωs are taken as a first order difference between the detection event and an off-set.

$$X(t) = \underbrace{\int_{t0+\Omega}^{t0+\Omega+\Delta+MA} Q_{out}(t)\,dt}_{Integrated\ Signal} - \underbrace{2 \cdot \int_{t0}^{t0+\Delta} Q_{out}(t)\,dt}_{Off-set} \tag{5.15}$$

This first order difference operation is shown in Equation 5.15 where t_0, Δ, and Q_{out} are the time where the spike even is aligned, static integration time of $20\,\mu s$, and analogue signal containing the spike. This clarifies that the moving average MA is mixed with the signal when it evaluates the mean of the Ω_1 section. This approach primarily helps with a self referenced gain that allows smaller register depth.

After several seconds of training data or equivalent spike count we can assume that the sections that maximize variance have been approached. At this point the Ωs are fixed and the centroids need to be generated to complete the adaptive process for classification. As illustrated in Figure 5.21 a similar feedback mechanism is used to adjust the mean centroids based on boolean results. This particular configuration is the adjustment of two centroids based on one feature in the Ω_1 section. By adjusting the centroids MA_μ when it is the closest to the new data point

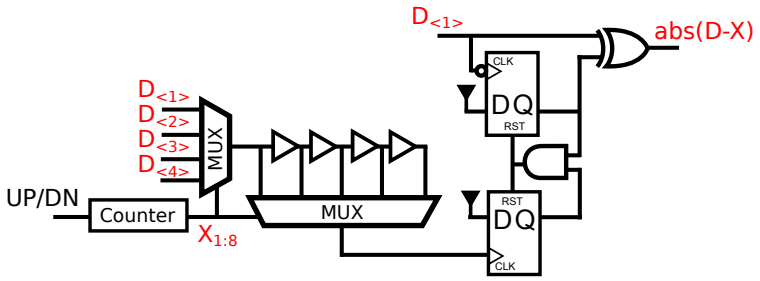
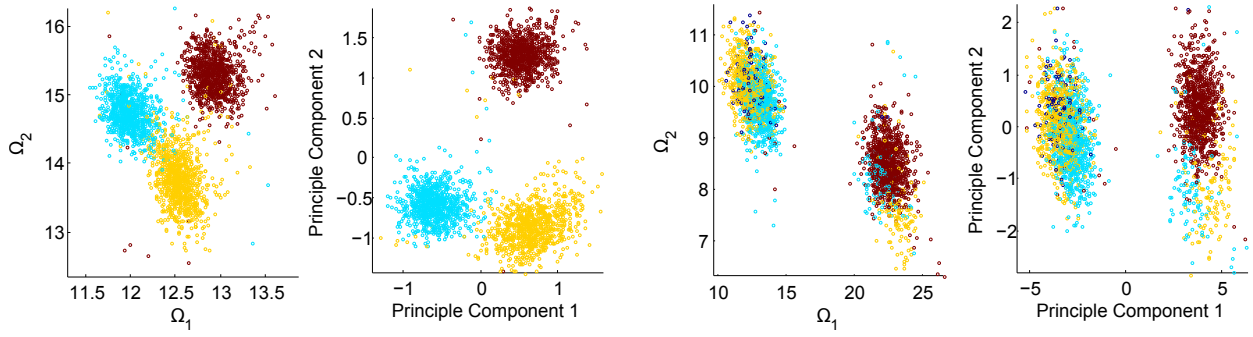


Figure 5.22: Delay line configuration for evaluating the absolute difference between the asynchronous time domain signal D and the registered centroid position X.

we realize a k-means clustering method with a l_0 norm distance operator.

Since we bound the Ω sections to be strictly non-overlapping the same analogue integrator can be used to evaluate the accumulated error of all features to each centroid. Moreover for a small a_1 the centroid adjustment can be time multiplexed leaving a reduced requirement on the total number of integrators required. This implies that K integrators are needed to iterative adjust all the centroids and M integrators are needed to evaluate the distance from each centroid. Because this adaptive process is performed in isolation we may perform the training in phases that updates clusters and features separately we will only need M or K integrators concurrently which ever is more demanding. However KM registers are needed to specify the location of each centroid which should be converted to a time-domain signal by using reconfigurable delay lines.

Such a configuration is illustrated in Figure 5.22 where there is course control by selecting different phases of D and fine control with a conventional multiplexed delay structure. Again the reduction in complexity and rejection of quantization noise by performing time domain computation opposed to the equivalent $8b$ full adder is typical of this processing modality. Finally the question should remain is that how are centroids initialized without requiring quantization. This requirement is avoided by using an iterative method with respect to centroid generation. After having a single centroid converge to the mean of the feature space we iteratively split centroids in two while training similar to that discussed in Section4.2.3. We presume redundant clusters will be generated that are removed if supervision is allowed to intervene or in the case that high level control is used to analyse which clusters are significant after several iterations.



(a) Feature space with $-26\,dB$ AWGN and background activity

(b) Feature space with $-14\,dB$ AWGN and background activity

Figure 5.23: comparing PCA and Ω feature distribution for the Difficult2 data set. Ground truth for spike classes annotated as cyan, maroon, yellow and blue for false positives.

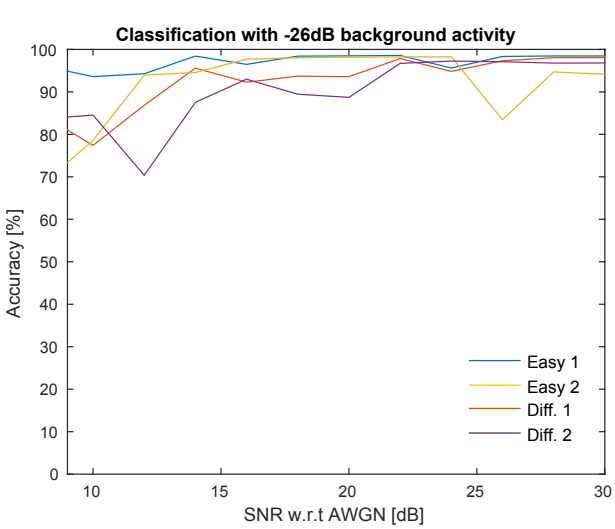
The results presented here however do not consider this supervision.

5.2.3 Validation

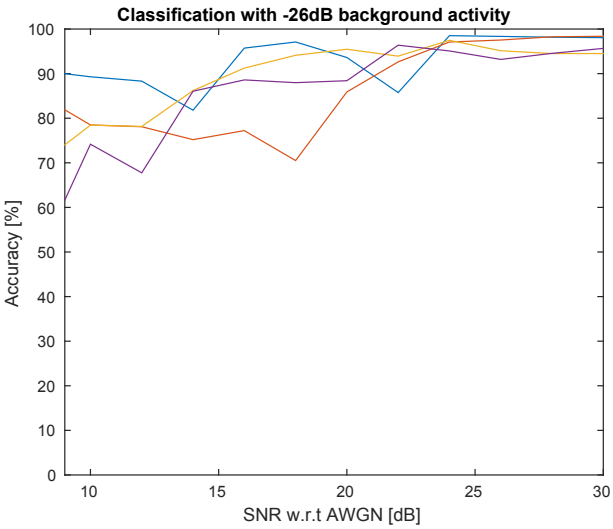
In order to demonstrate the viability of this approach we will simulate a linearised model that is constructed using Matlab. Here we aim show to what extent unsupervised methods are constrained with respect to classification performance. Original data sets used in Section 4.2 have been up sampled from $24\,kS/s$ to $240\,kS/s$ after the band limiting filters to emulate the continuous time logic that will operate at a high clock rate.

The feature space resulting from this method is exemplified in Figure 5.23. Here we compare it to that of the two component PCA feature space since the Ω represents its approximation. It is typical to see multiple additional clusters form either due to the detection of false positives or miss alignment of a spike class in the presence of noise. Initializing extra clusters can typically retain classification accuracy in noisy environment but degrade precision in pristine conditions.

The results in Figure 5.25 demonstrate classification accuracy in terms of the percentage of all correctly classified events with respect to the ground truth including false positives and false negatives. This indicates that when a signal to noise ratio exceeds $20dB$ the conditions are quite forgiving towards the simplicity of the algorithm. The two features used here require

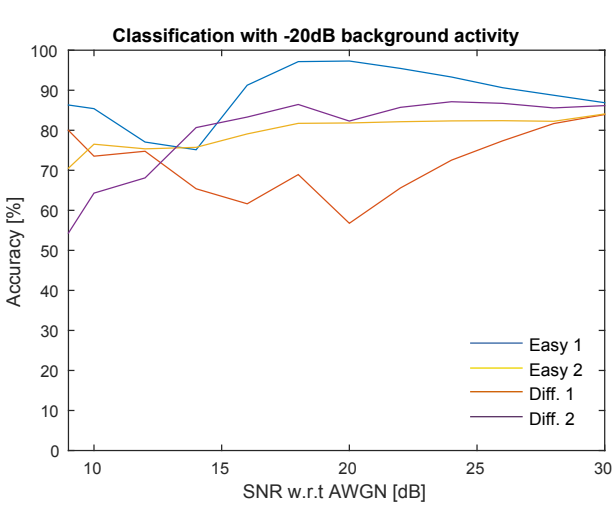


(a) Classification with two integral features Ω_2

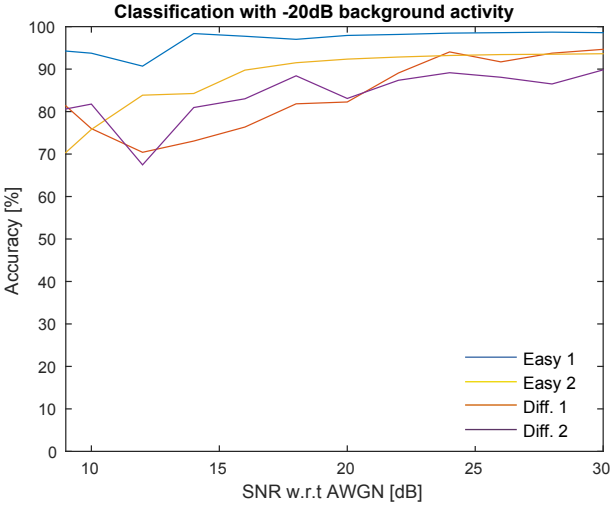


(b) Classification with three integral features Ω_3

Figure 5.24: Ω_2 and Ω_3 classification for data sets with $-26\,dB$ background activity.

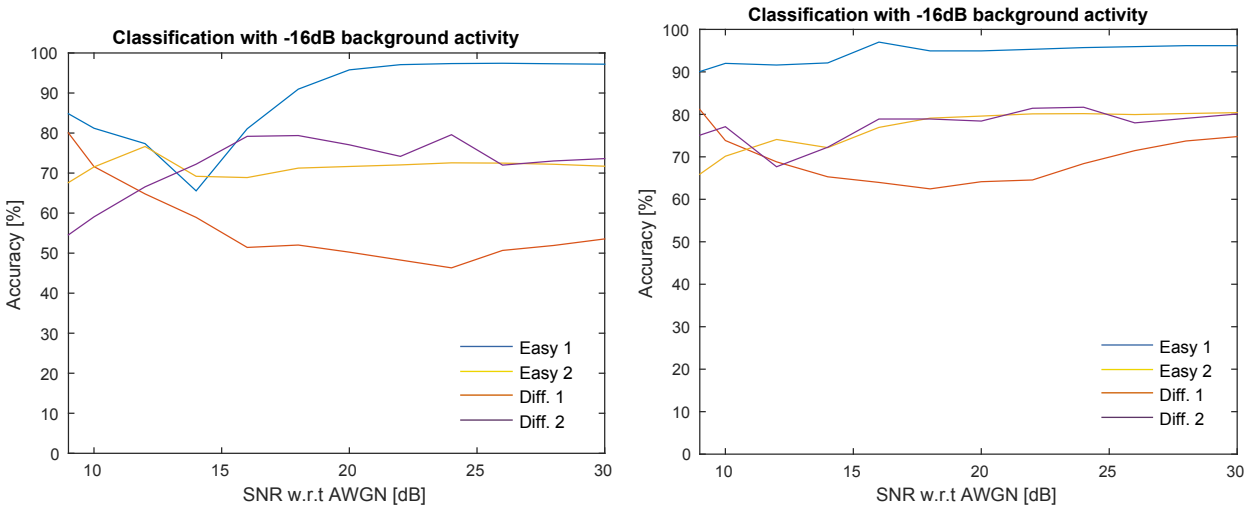


(a) Classification with two integral features Ω_2



(b) Classification with three integral features Ω_3

Figure 5.25: Ω_2 and Ω_3 classification for data sets with $-20\,dB$ background activity.



(a) Classification with two integral features Ω_2 (b) Classification with three integral features Ω_3

Figure 5.26: Ω_2 and Ω_3 classification for data sets with $-16\,dB$ background activity.

very little effort to adapt and classify activity. It is important to mention that a fixed filtering configuration is maintained for all test points in order to demonstrate adaptive characteristics.

The results in Figure 5.26 shows improvement in noisy conditions if the number of sections is increased to 3 implying a three dimensional feature space to improve centroid distance. Using the same algorithm for feature selection the configuration can deal with twice the amount of background noise without supervision. In some sense the fact that the signal is not quantized does not have a significant impact on classification accuracy. This highlights the importance of closed loop algorithms whether resources are constrained or not. As such constructing a convex search space or extracting well reasoned features from underlying phenomena is crucial to reducing in complexity.

Figure 5.27 shows that these noise levels detection is relatively consistent but not as adaptive as the digital approach. The threshold for detection will favour generating false positives over false negatives. The main point of failure for noise levels beyond that point lies with the inability to perform feature selection based on localized variance maximization. This is partially expected as PCA will similarly perform poorly when noise levels become comparable to the signal.

In a practical case it may be difficult to ascertain if signal to noise level is adequate to trust classification unless there is confidence to do so in the sense that there may be redundancy

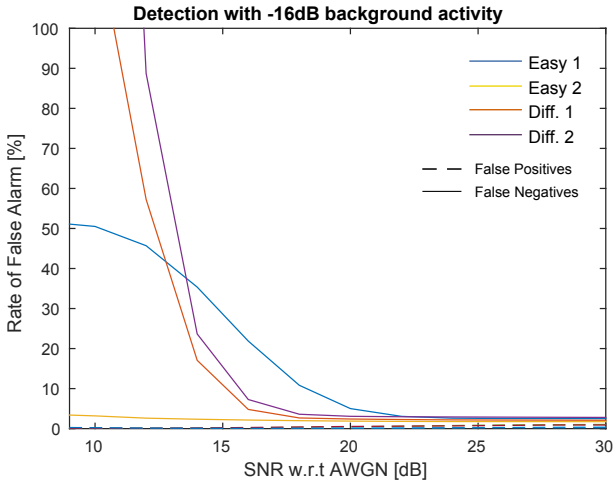


Figure 5.27: False alarm rates normalized by true positives for the analogue detection.

in the recording taken. However it is significantly more viable for realizing sub $1\,\mu W$ neural spike classification for large scale recordings considering the resource requirements for adaptive classification. Given that each integrator consumes less that $50\,nW$ in $0.18\,\mu m$ CMOS and each structure needs minimal supervision.

A number of methods are shown in Table 5.2 where we see classification accuracy and the corresponding hardware requirements in both the analogue and digital domain. The RVD and template methods presented in Section 4.2 represent the digital approach where little analogue components are needed beside the quantizer. Allocating more processing power or memory resources would imply choosing one over the other. As expected supervised intervention allows methods like WDF [200] to leverage a substantial improvement with respect to resource efficient classification. In this perspective we see using Ω_3 features as distributing our resources in the analogue domain while still maintaining comparable classification accuracy but require less reliance on digital scaling factors. We provide more comparison details in Section D for the proposed digital and analogue methods proposed by this work as well as the equivalent Matlab implementation used for evaluation.

Table 5.2: Overview of Detection & Classification performance in green for data sets from [1] for different methods. ⋆ Requires off-chip Supervision. † White noise is also added at -20dB of the signal power.

Method	Analogue	Digital Registers	Cycles / Sample	Data Set	Background @ -16dB †	Background @ -20dB †
RVD	1×ADC	83	172	Easy 2 Diff. 1 Diff. 2	73.4% 72.9% 74.8%	84.2% 87.1% 84.8%
Template	1×ADC	105	90	Easy 2 Diff.1 Diff. 2	82.0% 86.0% 80.3%	87.6% 83.5% 87.5%
WDF [200]	2×BP-Filter 1×ADC	41	104⋆	Easy 2 Diff. 1 Diff. 2	95.1% 85.0% 84.6%	99.1% 92.9% 91.6%
Ω_3 Features	3×BP-Filter 4×Integrator 4×DAC	16	1	Easy 2 Diff. 1 Diff. 2	80.0% 72.3% 79.8%	94.6% 93.1% 88.6%

5.3 Conclusion

This chapter has proposed a number of time-domain constructs that encourage mixed signal design for instrumentation. Where we derived underlying concepts from the phase state of a ring oscillator in order to represent continuous valued time domain memory as the equivalent of a clocked filp-flop or sampled capacitor. In addition we have discussed the means to analytically evaluate and optimize the characteristics of these topologies. Overall we present are clear benefits over conventional implementations such as instrumentation and the functional manipulation of continuous valued signals. Moreover these structures will scale performance with technology due to the extensive use of digital gates. The instrumentation structure in particular gives way to fully synthesized platforms. Performing processing and filtering in the digital domain remains to be critical for robust sensing of LFPs and EAPs in very poor signal to noise conditions. A 0.6 V 58 dB SNDR time domain instrumentation architecture is demonstrated with a NEF of 1.18 that generates multiphase PWM encoded digital signals using sub 0.01 mm^2 footprint and employing bandpass filtering with 40 dB/Dec roll off.

In extension we demonstrated the capacity for mixed signal analogue to information conversion

with respect to unsupervised classification that uses adaptive techniques to converge towards specific signal characteristics. Using reconfigurable integration of selected temporal sections in the spike shape lets us effectively focus resources on feature and cluster evaluation without open loop quantization. This mitigates the trade off associated with resolution and digital complexity. The main challenge as pointed out is establishing what dynamics will allow convergence to optimal feature extraction with reduced hardware requirements. Here we exploit certain phenomena in the principle components of spike shapes and the sensitivity of group delay of analogue detection to frequency content in spike waveforms to achieve direct classification.

It is typical that techniques behind instrumentation and signal acquisition are more mature in development and direction when compared to different processing modalities. Especially when realizing mixed signal methods for machine learning where a multitude of convoluted factors impact performance. There is much still to addressed when adaptive techniques are evaluated with respect to their resource efficiency and this will likely be a important aspect that will emerge in many intelligent sensor systems.

Chapter 6

Conclusion

As brain machine interfaces continue to progress in establishing a better communication link between neural activity and our models for cognitive or behavioural dynamics. We can expect with certainty the design considerations of these systems will remain a very involved process. Using abstractions and resource models will play a pivotal role in realizing these complex systems. Particularly as these systems require operational efficiency far beyond current state-of-the-art to enable chronic and implantable neuro-prothetics for full limb replacement. The models discussed here not only guide the design process towards more optimal systems they also provide excellent insight to inherent design limitations at different stages of design though clarity in system level objectives. A considerable amount of effort was put into realizing highly efficient and ultra-low power amplification of signals which can naturally be extended towards improving on-chip voltage regulators or RF-telemetry circuits to advance the capabilities of these systems while operating with several hundred millivolts of supply head-room.

There still remain many opportunities as emergent processing modalities that find them selves adapting to the current paradigm or platform of operation. Currently that paradigm is nanometre CMOS which suggests nothing other than "there is plenty of room at the bottom" [201]. While the results presented here advocate for transitioning to smaller feature sizes to miniaturize recording and increase the in channel processing capabilities that accommodate the algorithms for extracting information from electrode recordings. We must also give way to

software abstraction that deals with the increased complexity of these systems to retain long term utility and higher level adaptive function. Moreover we signify the resulting flexibility empowers electronic sensing systems to meet the broad set of challenges found in fabrication, implantation, and realization of chronic solutions. If we consider the maximal density of the proposed structures, by estimation for two-dimensional sensing arrays it is viable simultaneously record at a density of 400 locations different simultaneously per square millilitre of CMOS chip. That figure for some areas in the visual cortex implies a ratio on the order of 100 neurons per electrode [202]. This encourages us that this direction is an important step forward towards high fidelity neural decoding systems and will allow precise studies of neural circuits in rodents for disease models.

Putting our findings in perspective of the SAR limit to inducing power inside the body [203] and state-of-the-art results in midfield powering of millilitre sized implants as coupling scenario for implants deeper than 5cm [14]. This work suggests the power budget for a $2\,\mathrm{mm}^2$ CMOS chip and coil is around $200\,\mu\mathrm{W}$ at 11% of the maximum safety threshold. This may be improved with near-field coupling of multi coil configurations but either solution is challenged in powering multiple devices without more active electronics. We can be confident that our micro-controller structure in 65 nm CMOS is compact enough that the main limitation is power related. Extrapolating the power requirements from the current 180 nm CMOS design gives us the impression that we can integrate at least 136 reconfigurable instrumentation channels dissipating $1.47\,\mu\mathrm{W}$ by both analogue and digital domains. In fact integrating 64 channels can already be realized by our 180 nm CMOS implementation with power induction at 71% of the maximum safety threshold. While this seems unattainable by many systems presented in the literature today we argue that this approach will even allow active telemetry with UWB data transmission from the chip because of the highly compressed data-rates [204].

6.1 Original Contributions

The technical contributions presented by this thesis are detailed in the list below of which some has already been published. These publications are outlined in Section A.

- **Chapter 3**
  - A 1.26 μW instrumentation amplifier with 6 kHz bandwidth that provides 52 dB gain and exhibits a input referred noise figure of 5.6 μV_{rms} with a flicker noise corer of 20 Hz.
  - A 14.4 μW 11.4 ENOB $\Delta\Sigma$SAR ADC that can perform 200 kS/s with a 2.4 pF sampling capacitance that achieves a 10 fJ/conversion figure of merit.
  - Fabricated fully integrated signal acquisition front end that uses 0.01 mm^2 of silicon area to perform 5 mVpp signal quantization with a 58 dB dynamic range.

- **Chapter 4**
  - PCA & template matching spike sorting method for embedded systems that require 57 operations per sample and 680 bits of memory to perform fully unsupervised classification with >80% accuracy.
  - A 1.52 GOPS/mW distributed processing architecture for neural recording applications that has a 0.02mm^2 silicon foot print with fully reconfigurable 8 bit processing capabilities.
  - Linux based development platform for developing decoding techniques with support for C/C++/Python and other functional languages.

- **Chapter 5**
  - A 0.6 V 58 dB SNDR time domain instrumentation architecture with a NEF of 1.18 that generates multiphase PWM encoded digital signals with a sub 0.01 mm^2 footprint.
  - A 4^{th} order oscillator based bandpass filter for conditioning the time domain signals with reconfigurable gain that dissipates 120 nW of power.
  - Unsupervised analogue classification technique that achieves over 72% accuracy without necessitating signal quantization or a large number of analogue components.

6.2 Future Work

We have presented some of the most compact high performance instrumentation structures that can be integrated in nano-meter CMOS for wide-band recording of neural activity. This shifts the focus towards careful system integration of electrodes, data processing, wireless communication, and power management in order to realize a truly implantable solution. The first core objective that extends on the work here is stream lining the design with industry standard synthesis for large scale integrated systems. This implies formal silicon verification for proof of concept from our latest developments and involve more conservative design choices associated with improving device lifetime and risk-factors. This is important in providing a more universally translatable platform to allow other components of the system to be integrated arbitrary to the technology node of fabrication.

However we should also consider the long term goal for these systems. Our micro-controller approach dedicates a significant amount of resources just to perform spike sorting or signal characterization. While this is the focus in current systems to allow wireless solutions it is not the primary challenge for a neuro-prosthetic device which instead lies with decoding spike-train data or decomposing LFP activity. This implies that we need to adopt more resource efficient structures to extract features or spikes in neural activity equivalent to that proposed in Section 5.2. More concise signal acquisition allows us to focus on decoding the collection neural

activity rather than single recordings and possibly increase the prevalence where we reconfigure what signal features are extracted based on what is most informative at the system level. Specifically we picture a hierarchical approach where without intervention the system tunes to relevant information to guarantee some level of performance which is then tuned incrementally with off-chip supervision to maximize performance with higher order statistics and better optimization methods to find the best configuration that selects electrodes, LFP/EAP features, denoising bandwidth, etc. Moreover some aspects in regard to the design of a medical grade device are not yet fully investigated. For instance can we make the analysis of spiking activity suitable for long term reliability? This is specifically the endeavour of current research where experimental platforms need to reflect the informed capabilities.

We find it highly probable that integrating processing structures and other auxiliary circuit components will be one of the simpler aspects of realizing BMIs with better decoding performance and capacity. Primarily because the design tools allow directed and precise development following methods already presented in the literature. Innovation in electronics will likely allow the required performance improvement by scaling when it is needed. The principle challenge lies with compact system integration of these devices. Compatible fabrication of penetrating electrode structures with the right impedance characteristics for low noise and sustain unimpaired neural tissue interface for coupling activity may raise significantly more unexpected challenges. In addition we note importance of introducing miniaturized antenna structures that perform harvesting power as well as back-scattering recorded data. Both of these aspects are crucial to scaling integrated recording systems in a distributed fashion. That said, the effort towards developing computational modalities for high dimensional data analysis that specifically use analogue or time domain techniques still hold uncovered potential. Drawing from structures in the brain, bio-inspired ultra-low-power processing architectures for machine learning is an emergent challenge that can also find utility in brain machine interfaces [205].

Appendix A

Publications

The relevant conference and journals publications that have emerged from this thesis are summarized in the list below with reference to the corresponding technical chapter.

- **Chapter 3**
  - L. B. Leene and T. G. Constandinou, "Ultra-low power design strategy for two-stage amplifier topologies," in Electronics Letters, vol. 50, no. 8, pp. 583-585, April 2014, [Online] doi: 10.1049/el.2013.4196.
  - L. B. Leene, Y. Liu and T. G. Constandinou, "A compact recording array for neural interfaces," IEEE Proceedings of Biomedical Circuits and Systems Conference, pp. 97-100, May 2013, [Online] doi: 10.1109/BioCAS.2013.6679648.
  - L. Zheng, L. B. Leene, Y. Liu and T. G. Constandinou, "An adaptive 16/64 kHz, 9-bit SAR ADC with peak-aligned sampling for neural spike recording," IEEE Proceedings of International Symposium on Circuits and Systems, pp. 2385-2388, May 2014, [Online] doi: 10.1109/ISCAS.2014.6865652.
- **Chapter 4**
  - L. B. Leene and T. G. Constandinou, "A 2.7 μW/Mips, 0.88 GOPSmm^2 Distributed Processor for Implantable Brain Machine Interfaces," IEEE Proceedings of Biomedical Circuits and Systems Conference, October 2016.

- **Chapter 5**

  - L. B. Leene and T. G. Constandinou, "A 0.45V Continuous Time-Domain Filter using Asynchronous Oscillator Structures," IEEE Proceedings of Electronics Circuits and Systems Conference, December 2016.
  - M. Elia, L. B. Leene and T. G. Constandinou, "Continuous-time micropower interface for neural recording applications," IEEE Proceedings of International Symposium on Circuits and Systems, May 2016,[Online] doi: 10.1109/ISCAS.2016.7527295.
  - K. Faliagkas, L. B. Leene and T. G. Constandinou, "A novel neural recording system utilising continuous time energy based compression," IEEE Proceedings of International Symposium on Circuits and Systems, pp. 3000-3003, May 2015, [Online] doi: 10.1109/ISCAS.2015.7169318.

Appendix B

Nomenclature

B.1 List of Symbols

q	-	Unit charge of an electron
μ	-	Charge carrier mobility
η	-	Transistor slope factor (EKV model)
U_T	-	Semiconductor thermodynamic voltage
ζ	-	Amplifier noise excess factor.
f_T	-	Transistor transition frequency
f_c	-	Clock frequency
f_{lp}	-	3dB frequency for a low pass filter
f_{hp}	-	3dB frequency for a high pass filter
f_{UGF}	-	Unity gain frequency
f_{osc}	-	Oscillator frequency
BW	-	Bandwidth in Hz
V_{DD}	-	Circuit's supply voltage
V_{ref}	-	ADC's reference voltage
V_{os}	-	Amplifier offset voltage

A_{cl} - Closed loop gain

A_{ol} - Open loop gain

k - Boltzmann's constant

T - Operating temperature

IC - Transistor inversion coefficient

gm - Transistor transconductance

I_D - Transistor drain current

P_x - Power Dissipation of block X

en_{th}^2 - Thermal noise power

en_{fl}^2 - Flicker noise power

Γ - Impulse sensitivity function

B.2 List of Abbreviations

ADC - Analogue to digital converter

ASIP - Application specific integrated circuit

ASIP - Application specific instruction set processor

CMOS - Complementary metal oxide semiconductor

DARPA - Defence advance research project agency

DSP - Digital signal processing

DR - Dynamic range

DRAM - Dynamic random access memory

FET - Field effect transistor

FIR - Finite impulse response filter

FPGA - Field programmable array

FOM - Figure of merit

GPU - Graphical processing unit

EAP - Extra-cellular action potential

IIR - Infinite impulse response filter

LFP - Local field potential

MIPS - Million instructions per second

NEF - Noise efficiency factor

NESD - Neural engineering system design

NEO - Non-linear energy operator

NMOS - N-channel metal oxide semiconductor

NPI - Neuron processor interface

PCA - Principle component analysis

PMOS - P-channel metal oxide semiconductor

PVT - Process voltage and temperature

RVD - Recursive variance decomposition

SNR - Signal to noise ratio

THD - Total harmonic distortion

UDP - User datagram protocol

Appendix C

Supplementary Results for the Resource Models

This section will elaborate on some of the analytic findings in Chapter 3 by detailing the derivations and illustrating additional results that emerge from these relations. In particular we will demonstrate the impact of technology related factors to a larger extent that include quantitative measures for some of the more intuitive relations.

C.1 Topology Selection

We first introduced Eq. 3.15 as a base line for the figure of merit calculations. This reference SAR ADC only dissipates power though the capacitor array and the driving control logic without considering the comparator. We identified the power dissipation in the register logic as;

$$P_{Register} = 2N\,(N+2)\,f_s\,E_{gate} \tag{C.1}$$

Eq. C.1 simply reflects that a typical SAR operation that resolves N bits requires $(N+2)$ per

sample inferring the required clock frequency and the standard control logic uses $2N$ gates to perform the binary search operation.

$$P_{Capacitor\ Array} = E_{search} \cdot f_s C_{unit} V_{ref}^2 (2^{N-2} + 2^2) \tag{C.2}$$

Eq. C.2 demonstrates the power dissipation for capacitive switching has three terms. The total amount of capacitance in the DAC is represented as $C_{unit}\,(2^{N-2} + 2^2)$, the maximum input dynamic range as V_{ref}^2, and the average search efficiency for quantization as E_{search}. While this dependency is not surprising the term E_{search} needs to be evaluated under the proper assumptions. While we may be inclined to use sinusoidal test signals to extract this parameter if the sensor signals have similar structure but note that this will not reflect an even distribution with respect to the tested input ranges. In this respect triangular ramp test signals are representative to capture even-distribution of output quantization codes.

To realize a model where an analogue pre-amplifier is employed we estimate the power requirements for achieving the required bandwidth for settling. The capacitance for which we evaluate this settling speed is found be considering the output referred noise the results from the capacitor value. These two relations are summarized as follows;

$$(N+2)\, f_s\, A_{op} \cdot (N+2)\, \ln(2) = \frac{I_{amp}}{2\pi C_{out} \cdot \eta\, U_T} \tag{C.3}$$

The result in Eq. C.3 is simply derived from the unity-gain frequency that needs to exceed the clock rate in excess related to the settling requirement that scales by $(N+2)\,\ln(2)$.

$$\frac{V_{ref}}{2^{N+2}} = \sqrt{\frac{\zeta}{\eta} \cdot \frac{kT}{C_{amp}}} \tag{C.4}$$

Here we again refer to ζ as the amplifier noise excess factor introduced in Eq. 3.4. After some simplification by taking the most significant terms we arrive at Eq. 3.16 but evaluated here in full as;

$$P_{amp} = 32\pi\ \ln(2) \cdot \frac{\left(U_T\,(N+2)\,2^N\right)^2}{V_{ref}\,q} \cdot A_{ol}\,f_s\,NEF^2 \tag{C.5}$$

Similarly we derive the power for the dynamic latch type comparator following closely the analysis of [134]. The derivation for estimating the additional components for the hybrid data converter follows similarly to the pre-amp case where the bandwidth of the integrators is now also related to the order of oversampling but not that of settling. In particular we estimate the output referred noise of the first and most power intensive integrator as;

$$(1+2^{N-K})\,FIR\,f_s = \frac{I_{int}}{2\pi C_{int}\cdot\eta\,U_T} \tag{C.6}$$

Eq. C.6 simply evaluates the integration factor required to achieve a consistent output swing irrespective of SAR resolution. This shows us that the input signal level heavily dictates the corresponding biasing current I_{int}.

$$\frac{V_{ref}}{2^{N+2}} = \sqrt{\frac{\zeta}{\eta}\cdot\frac{kT}{C_{int}}} \tag{C.7}$$

Similarly Eq. C.7 relates the integration capacitance to the expected output referred noise. Summarizing these assertions gives us a direct relation between ADC resolution and power dissipation;

$$P_{amp} = 32\pi\cdot\frac{\left(U_T\,2^N, NEF\right)^2}{q\,V_{ref}}\cdot FIR\,f_s\,(1+2^{N-K}) \tag{C.8}$$

For the sake of clarity in the discussion we estimate the FIR order based on the expected variance from a rectangular window operator. Because this digital filter is estimating a DC component such that any up-modulated white noise power will scale as $\sqrt{FIR}$ for such a case and the dependency of a specific window operator will follow this upper bound. This result follows from the maximal dynamic range of real signals for window functions and assert that if we require another K bits to be resolved by the oversampling loop the will need a filter

order of $(K+1)^2$ to grantee that when using a more effective noise shaping window operator we will achieve the desired noise requirement. Applying a similar estimation on digital power dissipation as before such that;

$$P_{fir} = [(N-K) + K + \log_2(K-1)] \cdot (FIR + N - K) \cdot f_s E_{gate} \tag{C.9}$$

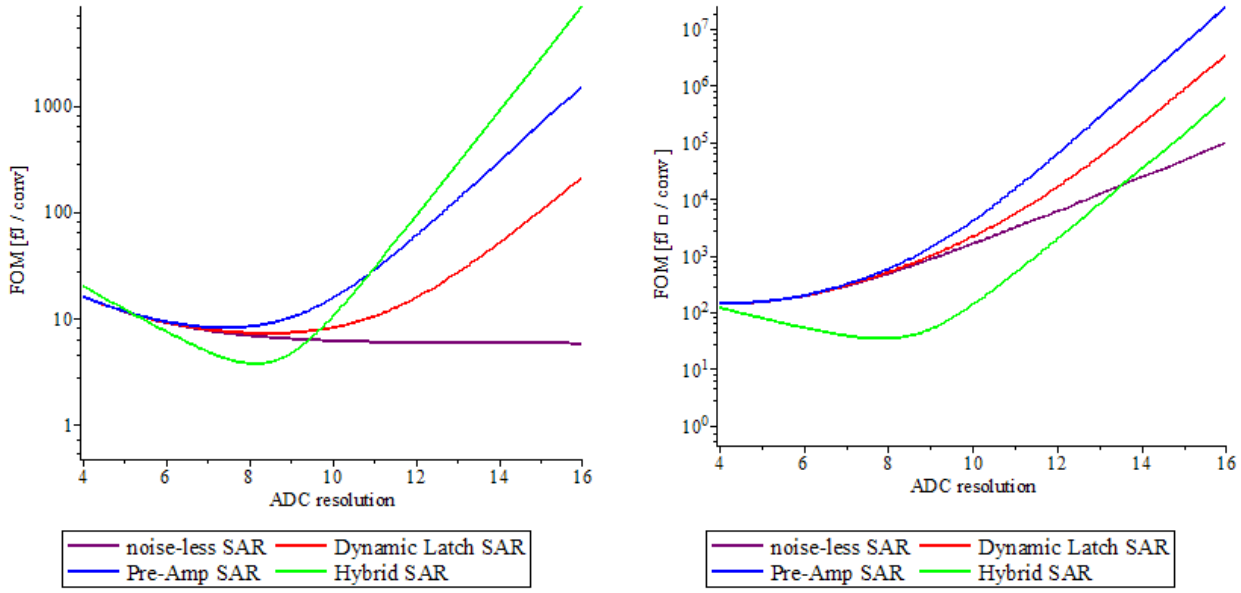
Eq. C.9 states that the required clock frequency is $FIR+N-K$ and the total number of gates being clocked is $(N-K)$ for the SAR operation and $K+\log_2(K)$ for the $\Sigma\Delta$ decimation. For completeness we also note the dissipation of the capacitive DAC is estimated as;

$$P_{\Sigma\Delta\ Array} = E_{search} \cdot f_s C_{unit} V_{ref}^2 (2^{N-K-2} + 2^2 + FIR) \tag{C.10}$$

While we have extracted many of the empirical parameters that are associated with the implementation presented in Sec 3.3.3. Let us consider the impact of parameter variation and reflect on the expectations. This will allow us to evaluate the sensitivity of these parameters with regard to performance measures.

First we introduce Fig. C.1 which is the normalized case that we presented earlier however here we apply a different scaling with regard to increasing resolution. Here the resolution of the SAR and $\Sigma\Delta$ loop are increased simultaneously which again demonstrates that this topology can provide compact quantization for from 8-10 bits. However because we are using vertical Metal-Insulator-Metal Capacitors provided by the technology that exhibit minimal variance per unit area the unit capacitance C_{unit} is $32\,fF$. This is relatively big when compared to very aggressive SAR designs presented in the literature that use unit capacitors of several femtofarad.

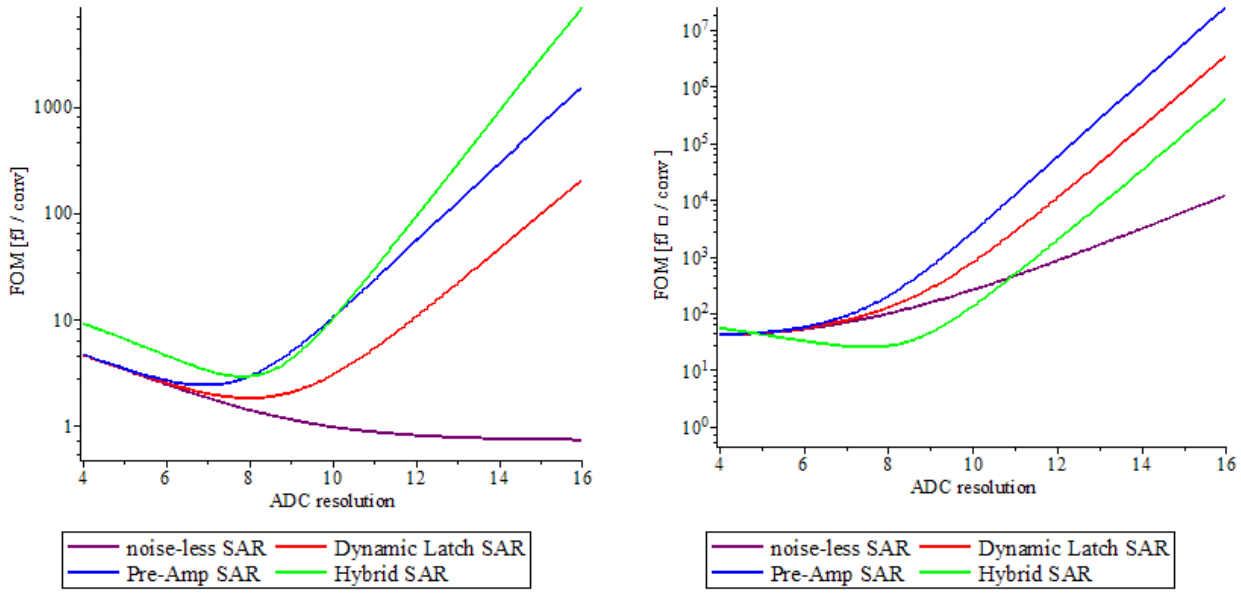
This leads into results presented by Fig. C.2 where the unit capacitance is reduced to $4\,fF$. This illustrates that the hybrid topology has difficulty scaling its power budget because there



(a) Figure of merit for different SAR topologies.

(b) Efficiency area product for different SAR topologies.

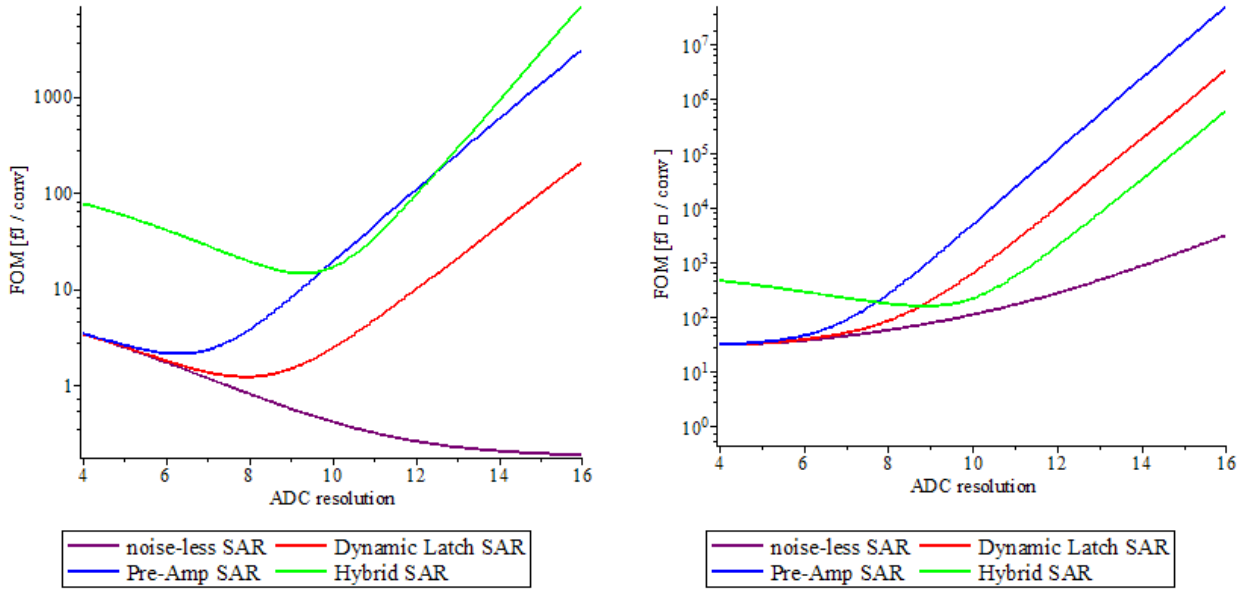
Figure C.1: Summary of the performance merit for each topology.



(a) Figure of merit for different SAR topologies.

(b) Efficiency area product for different SAR topologies.

Figure C.2: Summary of the performance merit for each topology with a 4× reduction in unit capacitance.



(a) Figure of merit for different SAR topologies.

(b) Efficiency area product for different SAR topologies.

Figure C.3: Summary of the performance merit for each topology with a 2× reduction in reference voltage and a 4× reduction in unit Capacitance.

are significantly more analogue components that the simple structures. In fact if we reduce the power supply from $1.2\,V$ to $0.6\,V$ the inhibitory analogue scaling that the integrators have exacerbates this effect. Particularly the dynamic comparator structure which has good digital scaling characteristics will allow significantly smaller power dissipation. However in all cases we note that for precision that exceeds $60\,dB$ the hybrid data converter topology will allow for a more compact designs.

C.2 System Level Abstraction

As most of the relations presented in Sec. 3.4 follow naturally from Eq. 3.22. Here we will provide some additional details regarding our assertions and observations. To show that Eq. 3.25 represents the expected increase in variance, let us consider the typical gain sensitivity to the variance of our normalized unit capacitor.

$$\frac{\sigma_A^2}{A_{cl}^2} = \frac{\Delta_A^2 - A_{cl}^2}{A_{cl}^2} \quad \text{where} \quad \Delta_A = \frac{A + \sqrt{A_{cl}} CI\,\sigma}{1 + CI\,\sigma} \tag{C.11}$$

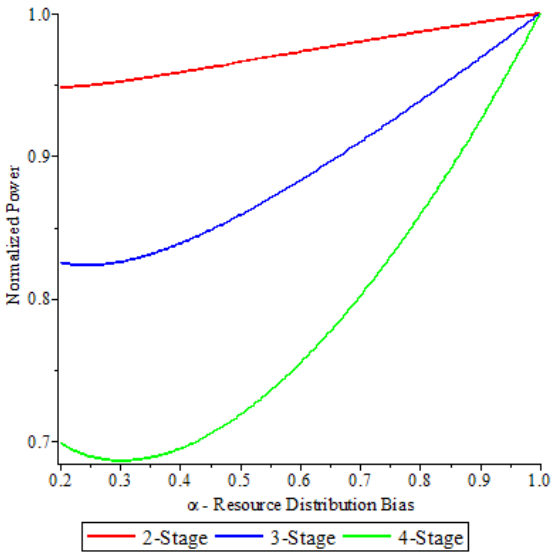
This expression is similar to that used in estimating the variance of a capacitive DAC in Eq. 3.13. Δ_A represents the expected gain deviation from the mean that relates to our confidence interval CI. Strictly there are only two components that attribute to gain variance which are the input and the feedback capacitors. These contributions have exactly a ratio of $\sqrt{A_{cl}}$ for their respective variances. Expanding this result will reveal;

$$\frac{\sigma_A^2}{A_{cl}^2} = \sigma\,CI \cdot \frac{2\left(A_{cl} - \sqrt{A_{cl}}\right) + \sigma\,CI(1 - A_{cl})}{A_{cl}\left(\sigma\,CI + 1\right)^2} \approx \sigma\,CI \cdot \left(2 - \frac{2}{\sqrt{A_{cl}}}\right) \tag{C.12}$$

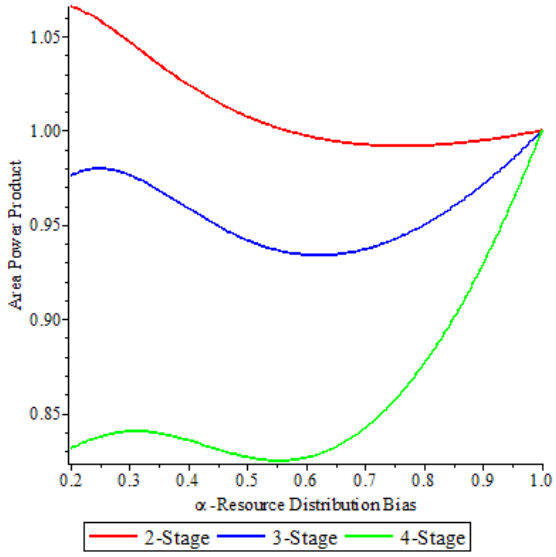
For exceedingly larger A_{cl} it makes sense to take the higher order terms however instead we chose include the $\sqrt{A_{cl}}$ term to considering the design of low-gain multi-stage configurations. Since the typical deviation of σCI is on the order of several percent it is conceivable that these higher order terms have vanishing contribution irrespective of gain. In order to compensate the increased variance the capacitor sizes need to be adjusted proportionally to that of the change standard deviation. This results in the penalty factor accumulated from every stage as Eq. 3.25.

The results in Fig. C.4 show that particularly in low gain settings that relate to smaller V_{DD} configurations the resource distribution becomes more important. This is because the individual gain of each stage is low resulting in very poor noise figures for minimal area configurations. With conventional structures we see that multi-state topologies can benefit greatly from this factor primarily because the baseline performance is much worse than single or two-stage structures.

Fig. C.5 reveals some of the more intuitive details associated with gain matching and ADC SNR performance with quantitative measures. In a) we observe the general inverse proportionality that gain matching exhibits with regard to capacitor are requirement. We point out that it

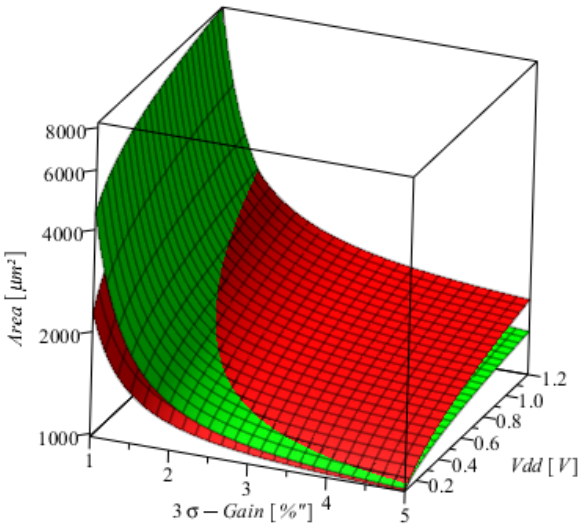


(a) Normalized power reduction for different amplifier configurations with respect to resource redistribution.

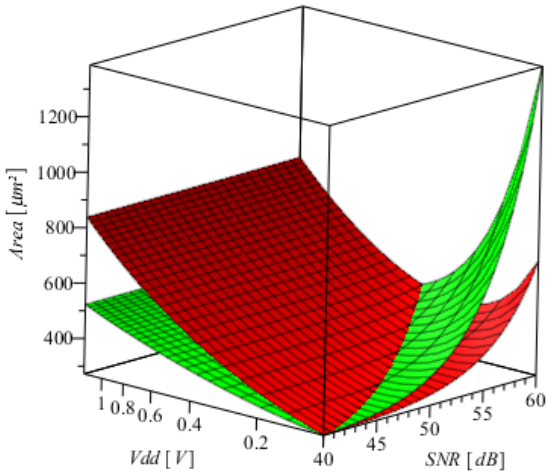


(b) Normalized power area product for different amplifier configurations with respect to resource redistribution.

Figure C.4: Normalized resource redistribution dependency in low gain settings $G_T = 200$.



(a) Supply voltage and target gain variance dependency.



(b) Supply voltage and data converter SNR dependency.

Figure C.5: System area requirement for two-Stage/single-Pole (red) and three-stage/double-pole (green) amplifier/filter structures.

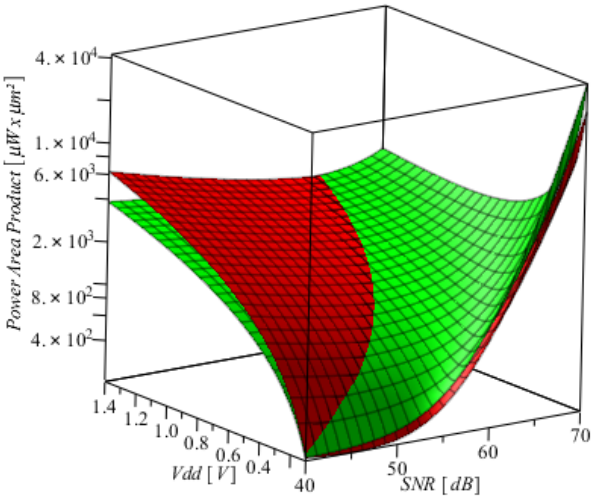


Figure C.6: Detailed dependency with regard to system supply voltage and the precision of the two-stage/single-pole (red) and three-stage/double-Pole (green) amplifier/filter structures.

can be marginally reduced by reducing G_T or equivalently the supply voltage. On the other hand in b) we see that reducing G_T by increasing the ADC dynamic range is more effective of a solution when we do not consider the associated increase in power dissipation.

Fig. C.6 demonstrates a more generic result when the amplifier power budget and gain configuration is actively matching the dynamic range of the ADC. Neglecting linearity we observe that there is an interesting strategy for selecting the appropriate supply voltage in accordance to the targeted ADC resolution. Similarly we make the assertion that there is gain-varience, quantization limited, and linearity limited regimes in each extrema of this graph that will bound viable implementations.

C.3 Transistor level implementation

For clarity the transistor level implementation is detailed here in terms of the transistor sizing.

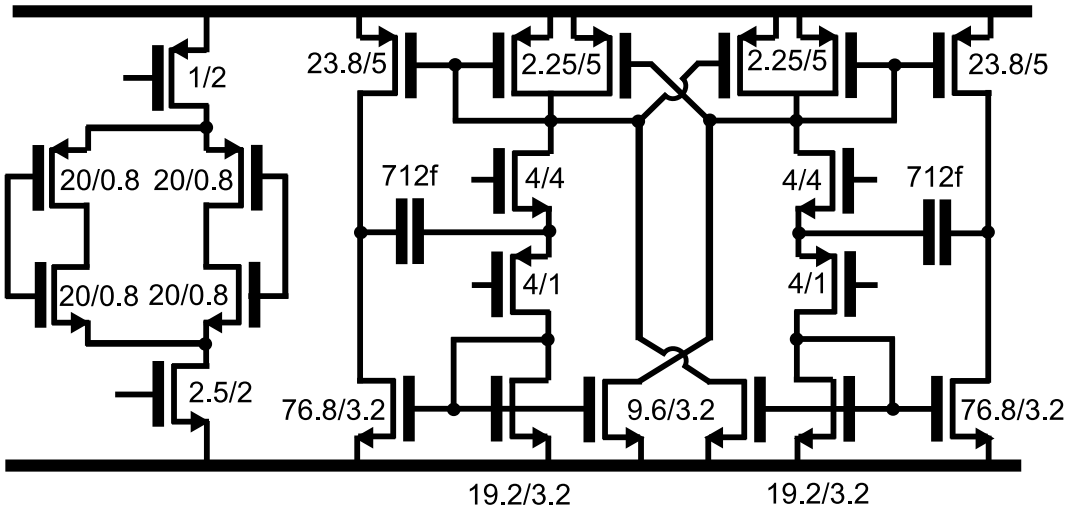


Figure C.7: Transistor sizing for the two stage operational amplifier with bandwidth boosting implemented in 0.18μm CMOS. Sizing units shown are in microns.

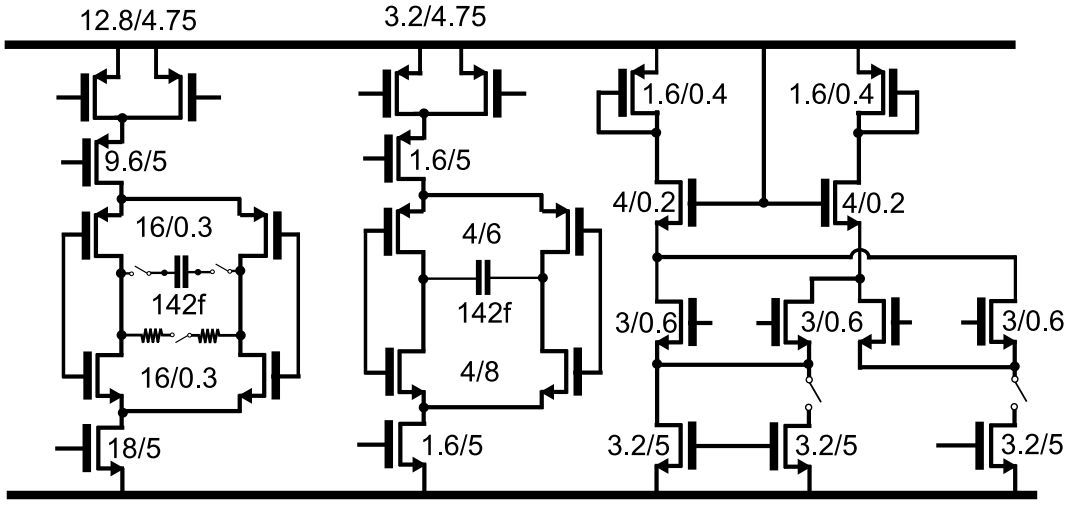


Figure C.8: Transistor sizing for the fully differential $\Delta\Sigma SAR$ switched loop filter implemented in 0.18μm CMOS. Sizing units shown are in microns.

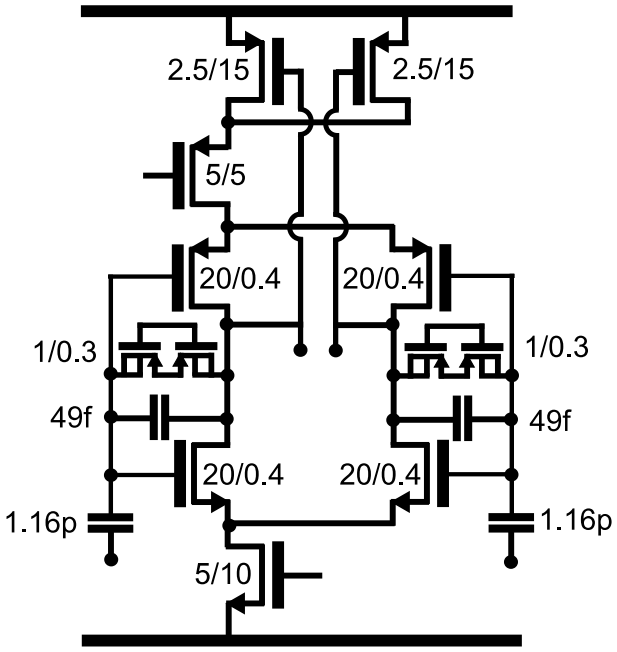


Figure C.9: Transistor sizing for the differential low noise complementary instrumentation amplifier implemented in 0.18μm CMOS. Sizing units shown are in microns.

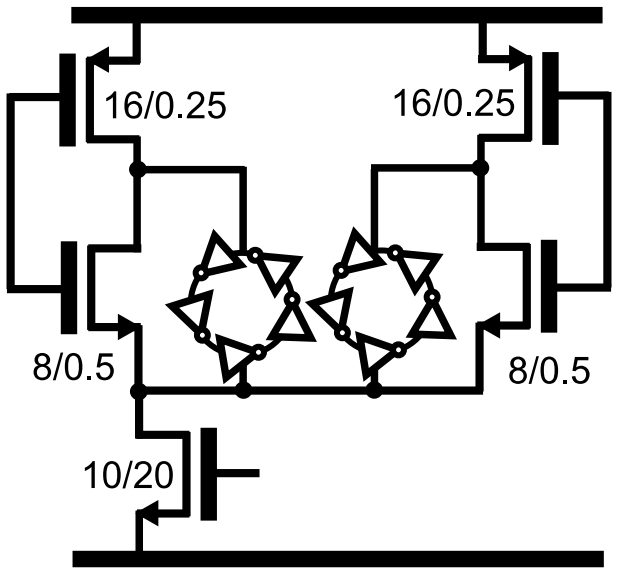


Figure C.10: Transistor sizing for the fully differential time domain instrumentation amplifier implemented in 0.18μm CMOS. Sizing units shown are in microns.

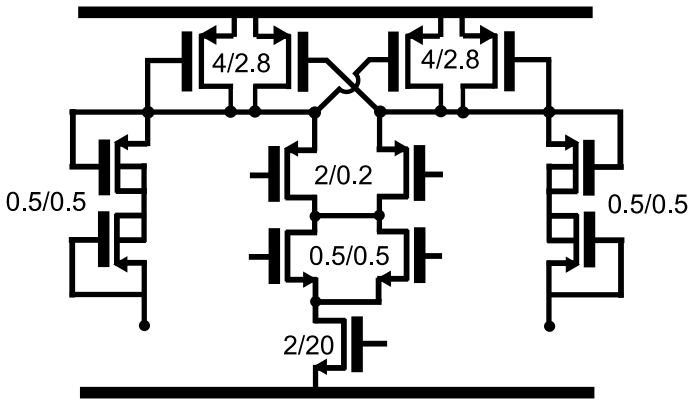


Figure C.11: Transistor sizing for the time domain instrumentation high pass filter structure implemented in 0.18μm CMOS. Sizing units shown are in microns.

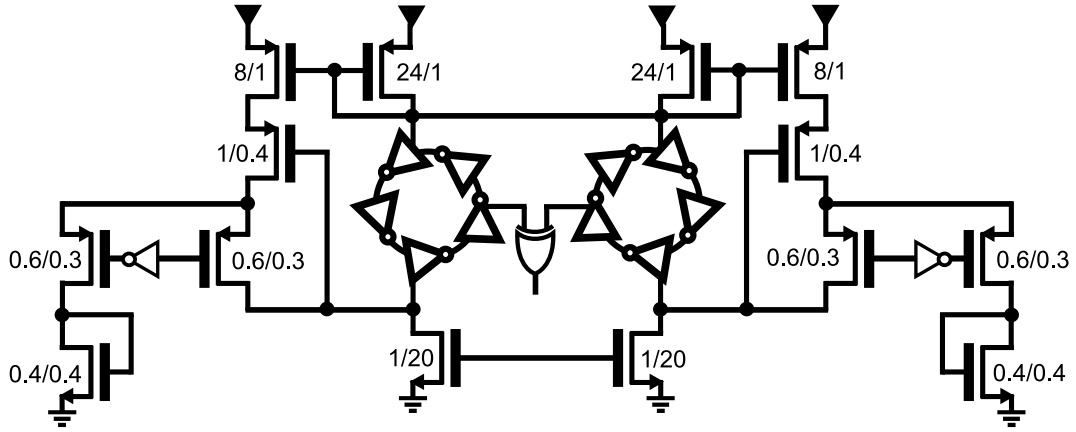


Figure C.12: Transistor sizing for the differential time domain filter implemented in 0.18μm CMOS. Sizing units shown are in microns.

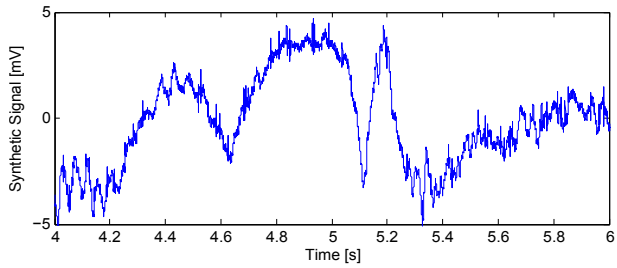
Appendix D

Supplementary Results for Classification Methods

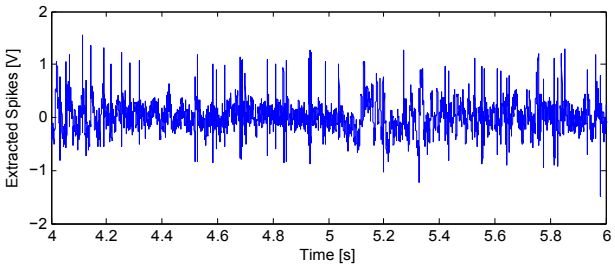
This section will detail the signal characteristics found in the data sets used extensively for validating the methods proposed here. We will exemplify some of the signal features and demonstrate the implementation of the proposed methods with Matlab code.

D.1 Signal Characteristics

As was mentioned used evaluation recordings contains different sets of synthetically generated spike trains for three classes of neurons each with different spike morphologies. The spike shapes are based on recordings with background noise of similar spikes randomly distributed in time at lower amplitudes. While the these datasets provide time stamp data for when the individual spike wave forms can be found in the recording they do not introduce the low-frequency content that is expected from typical recording. Here we have extracted these low frequency variations from real recordings with a high order FIR filter and added them to the synthetic data. While this will not greatly effect the classification results it assures us that the proposed methods are not inadvertently sensitive to low-frequency components.

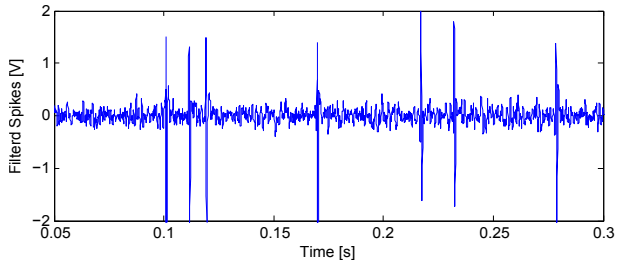


(a) Unfiltered synthetic data for LFP and EAP recordings.

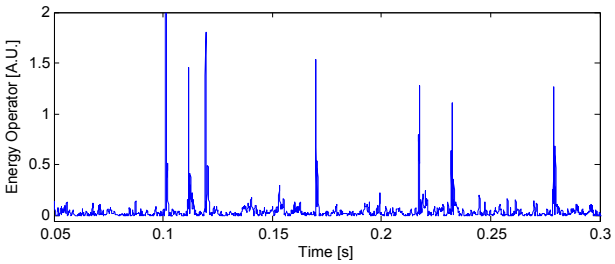


(b) Spiking activity after IIR reaction of slow varying component.

Figure D.1: Example of synthetic data derived from typical neural recordings.



(a) FIR filtered waveform of the spiking activity.



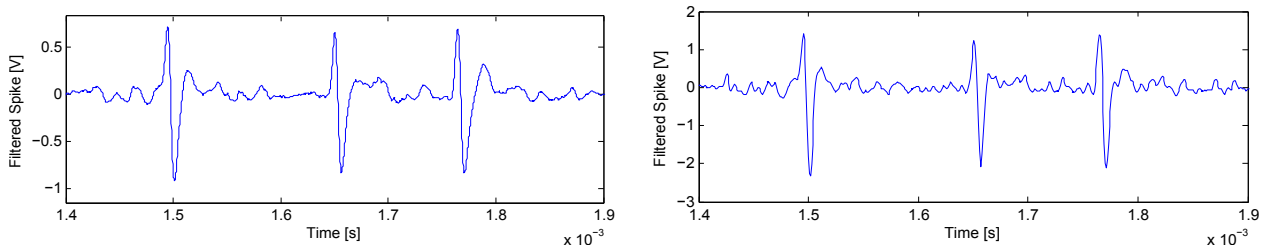
(b) Resulting waveform of the detection operator which rejects noise components.

Figure D.2: Filtered components of spiking activity and detection operator output.

Fig. D.1 depicts the typical recordings that are used before preprocessing. Whether we extract the finer signal components in the digital domain or in the analogue domain it is necessary to carefully consider the possible dynamic range. Particularly as buffer overflows are not explicitly modelled unless hardware specific data types are used.

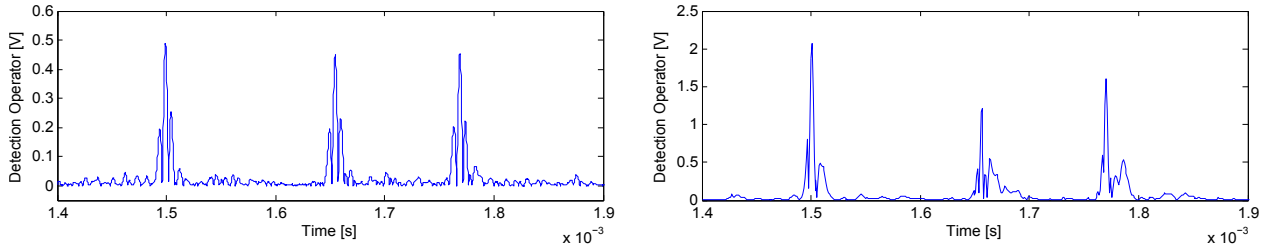
Once the spiking activity is extracted it may appear like that in Fig. D.2. Because spikes have a board bandwidth relative to the sampling rate, the white noise is typically more evident than what is presented. Alongside the spiking waveform we present the energy operator. Here there is a clear difference in how apparent the noise is in contrast to the spike amplitude. This disparity improves the more energy each spike had in the high-frequency bands.

Fig. D.3 compares the characteristic difference between analogue and digital filtering strategies. While we typically expect the group delay is well controlled in digital systems the analogue



(a) Detailed spike waveform after analogue filtering. (b) Detailed spike waveform after digital filtering.

Figure D.3: Comparison of digital and analogue pre-filtering methods.



(a) Analogue estimation of spiking energy. (b) Digital estimation of spiking energy.

Figure D.4: Comparison of digital and analogue detection methods.

implementation exhibits more systematic rejection of out of band components. Low order FIR structures in particular suffer from limited leaking these components but allow finer adjustments that accentuate separating features in the signal.

Similarly Fig. D.4 compares the digital and analogue realizations of the detection operators. Since we only use linear blocks to realize the analogue operator we lose the suppression of white noise. Moreover the digital operator can fine tune delay used to correlate derivative with amplitude components allowing for a very explicit single maximum in each waveform.

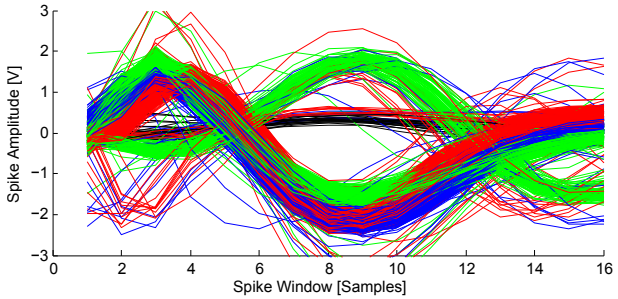


Figure D.5: Colour coded spike-waveforms for 16-sample spike windows where the three classes of neurons are coloured (ref,blue,green) and the false positives labelled (black).

Finally we illustrate a sub-set of aggregate waveforms captured by the digital methods in Fig. D.5. This data is used in the RVD and template-matching methods.

D.2 Implementation structure

Here we will detail the Matlab code implementation for the proposed methods which were used for validation in association with the data. Note alpha represent the smallest representable value which reflects the depth of the registers.

D.2.1 Introducing noise and LFP content

This code section exemplifies the method by which noise is systematically added to each recording in addition to LFP content. We note that the synthetic spike trains have well defined average rates and consistent amplitudes which allows the additive noise to remain consistent across all tests.

```
%Get LDP and EAP synthetic data
load(LFP_content)
LFP.Time=LFP.Time-min(LFP.Time);
load(SpikeTrain_dataset)

%Set low pass filter & sampling rate at 24KS*scale/2 ~ 32KS/s
scale=3;
lpflt = designfilt('lowpassiir', 'FilterOrder', 1,                    ...
'HalfPowerFrequency', 6500, 'SampleRate',                             ...
scale*10^(3)/samplingInterval, 'DesignMethod', 'butter');

% Resample Data
AT=resample(LFP,[1:size(data,2)].*(samplingInterval*10^(-3)));
rdata=awgn(resample(data(1:size(data,2))',scale,2),snr);
rdata=rdata+resample(AT.Data,scale,2);

% Analogue Filters
rdata=filter(lpflt,rdata);
```

D.2.2 Analogue detection operator

The analogue detection operator simply operated on the bandpass filtered sub-components of the original signal and tracks a means threshold at three times that of the average. Because we specifically upsample the waveform sectioning it make sense that the adaptive threshold updates close or below the Nyquist frequency in order to capture slowly varying changed in spiking intensity.

```
% Process samples for Test Size (TS)
i=34;
while(i<TS)
i=i+1;
        %Update Threshold value at low frequency
        if(mod(i,128)==0)
                        if(2.5.*made(i)>mei)
                        mei=mei+alpha;
                        else
                        mei=mei-alpha;
                        end
                %Set treshold at 3x rms with minimum at ~0.01
                drms=3*mei+1/64;
        end
        %Upon peak-detection above threshold process data
        if(made(i)>drms && made(i)>made(i+1))
                G(1:200,k)=rdata(i-32:i+167);
                IDX(k)=i*2/scale;
                k=k+1;
                i=i+180;
        end
end
```

D.2.3 Digital detection operator

The digital operator is slightly more sophisticated as it is not sensitive to the rate or spiking when estimating its threshold. While we similarly remove low-frequency content equivalent to the analogue filters. The threshold updates are only performed upon detecting peaks in the energy operator.

```
% Process samples for Test Size (TS)
while(i<TS)
        i=i+1;
        %track low-frequency signal
        trv=trv+(rdata(i)-trv)/32;
        %Perform IIR + FIR filtering
        T(i)=rdata(i)-trv+T(i-1).*0.25;
        G(i)= T(i-a3:i)*FIR;
        %Estimate spiking energy and track its noise
        DO(i)=abs((0.25+G(i)).*T(i-round(a3/2+1))); %+DO(i-1).*0.25;
        nest=nest+(a2*DO(i)-nest)/512;
        if(cnt<9)
                %if new local minima is detected reset analysis else
                    continue
                if(DO(i)>lm)
                        lm=DO(i);
                        thv = thv + (a1*nest+DO(i)-a1*thv)/512;
                        F(k,1:8)=G(i+[-10:2:-4  -3:0]);
                        cnt=1;
                else
                        F(k,8+cnt)=G(i);
                        cnt=cnt+1;
                end
        elseif(cnt<scale*15)
                %do signal analysis clasification here
                cnt=cnt+1;
        elseif(DO(i) > thv )
                %upon detecting threshold crossing initialize
                   buffers
                lm=DO(i);
                k=k+1;
                thv = thv + (a1*nest+DO(i-2)-a1*thv)/512;
                F(k,1:8)=G(i+[-10:2:-4  -3:0]);
                cnt=1;
        end
end
```

D.2.4 Ω_3 optimization

The mixed signal optimization that selected the best sections in the aggregate waveforms is performed iteratively. Here we simply introduce normalization factors K and section pointers T.

```
% Use the collected spike windows in G
for i = repmat(1:size(G,2),1,1)
```

```
        % for each $\Omega$ feature find the best section
        for j = 1:3
                %offset with reference point
                ref=3.*sum(G(RP(j),i));
                %integrate current section current section
                rng=round(K(j)*10)+1;
                acc=sum(G([-rng:rng]+round(T(j)),i));
                %nomalize the integration to reference
                if(ref < acc && K(j) > 0.1)
                        K(j)=K(j) - alpha;
                elseif(K(j)<1)
                        K(j)=K(j) + alpha;
                end
                %Compare signal deviation and increment on sections
                G1=abs(ref-sum(G([-rng:rng]+round(T(j)+2),i)));
                G2=abs(ref-sum(G([-rng:rng]+round(T(j)-2),i)));
                if(T(j)>70 && G2>G1)
                        T(j)=T(j)-alpha;
                elseif(T(j)<90 && G1>G2)
                        T(j)=T(j)+alpha;
                end
        end
end
```

D.2.5 K-means with centroid splitting

Here we present the means by which we explicitly multiplex active clusters and the iterative splitting of each centroid.

```
%intitialize first cluster arbitrarily
seed=randi(size(F,1)-5,1);
U(1:4,:)=F(seed+[1:4],:);
%for each iteration of increasing clusters track centroids
for NM=[1, 2, 4]
        %if new iteration duplicate new centroid from previous
        switch NM
                case 2
                        U(2,:)=U(1,:)-ones(1,size(U,2)).*alpha;
                case 4
                        U(3,:)=U(1,:)-ones(1,size(U,2)).*alpha;
                        U(4,:)=U(2,:)-ones(1,size(U,2)).*alpha;
        end
        % Track centroids using data in F
        for i = repmat(1:size(F,1),1,1)
```

```
16                  %multiplex centroid adjustment
17                  if(CS==0)
18                          CA=mod(CA,NM)+1;
19                  end
20                  CS=mod(CS+1,4);
21                  [dist, mc]=min(sum(abs(repmat(F(i,:),NM,1)'-U(1:NM
                        ,:)')));
22                  %if current centroid is minimum then adjust
23                  if(CA==mc)
24                          U(mc,:)=U(mc,:)+sign(F(i,:)-U(mc,:) ).*alpha
                            ;
25                  end
26          end
27  end
```

D.2.6 Recursive variance decomposition

Similarly this implementation realizes the estimation of loading vectors that provide a basis for the spikes in F.

```
1  %Estimate loading vectors based on spikes in F
2  for i = repmat(1:size(F,1),1,1)
3          %First track mean xu and first loading vetor s1
4          for j = 1:size(F,2)
5                  d(j) = F(i,j)-xu(j);
6                  xu(j) = xu(j) + sign(d(j)-xu(j))*alpha;
7                  ad(j) = abs(d(j));
8                  s1(j) = s1(j) + sign(ad(j).*g1-s1(j))*alpha*2;
9          end
10         %get dot product s and project on it-self to check gain g1
11         s=ad*s1;
12         if( (ad'-s1.*(s))'*s1 > 0 )g1=g1+alpha;
13                 else g1=g1-alpha;
14         end
15         %using varience residues esitimate second loading vector s2
16         for j = 1:size(F,2)
17                 ad(j) = ad(j)-s1(j).*s;
18                 s2(j) = s2(j) + sign(ad(j).*g2-s2(j))*alpha*2;
19         end
20         %get dot product of s2 and project to normalize with g2
21         s=ad*s2;
22         if( (ad'-s2.*(s))'*s2 > 0 )
23                 g2=g2+alpha;
24         else
```

```
25                      g2=g2-alpha;
26              end
27  end
```

Bibliography

[1] R. Q. Quiroga, Z. Nadasdy, and Y. Ben-Shaul, "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering," *Neural Computation*, vol. 16, pp. 1661–1687, April 2004. [Online]: http://dx.doi.org/10.1162/089976604774201631

[2] R. A. Normann, "Technology insight: future neuroprosthetic therapies for disorders of the nervous system," *Nature Clinical Practice Neurology*, vol. 3, pp. 444–452, August 2007. [Online]: http://dx.doi.org/10.1038/ncpneuro0556

[3] K. Birmingham, V. Gradinaru, P. Anikeeva, W. M. Grill, B. Pikov, Victor McLaughlin, P. Pasricha, K. Weber, Douglas Ludwig, and K. Famm, "Bioelectronic medicines: a research roadmap," *Nature Reviews Drug Discovery*, vol. 13, pp. 399–400, May 2014. [Online]: http://dx.doi.org/10.1038/nrd4351

[4] "Bridging the bio-electronic divide," Defense Advanced Research Projects Agency, Arlington, Texas, January 2016. [Online]: http://www.darpa.mil/news-events/2015-01-19

[5] G. Fritsch and E. Hitzig, "ber die elektrische erregbarkeit des grosshirns," *Archiv fr Anatomie, Physiologie und Wissenschaftliche Medicin.*, vol. 37, pp. 300–332, 1870.

[6] G. E. Loeb, "Cochlear prosthetics," *Annual Review of Neuroscience*, vol. 13, no. 1, pp. 357–371, 1990, pMID: 2183680. [Online]: http://dx.doi.org/10.1146/annurev.ne.13.030190.002041

[7] "Annual update bcig uk cochlear implant provision," British Cochlear Implant Group, London WC1X 8EE, UK, pp. 1–2, March 2015. [Online]: http://www.bcig.org.uk/wp-content/uploads/2015/12/CI-activity-2015.pdf

[8] M. Alexander, "Neuro-numbers," Association of British Neurologists (ABN), London SW9 6WY, UK, pp. 1–12, April 2003. [Online]: http://www.neural.org.uk/store/assets/files/20/original/NeuroNumbers.pdf

[9] A. Jackson and J. B. Zimmermann, "Neural interfaces for the brain and spinal cord restoring motor function," *Nature Reviews Neurology*, vol. 8, pp. 690–699, December 2012. [Online]: http://dx.doi.org/10.1038/nrneurol.2012.219

[10] M. Gilliaux, A. Renders, D. Dispa, D. Holvoet, J. Sapin, B. Dehez, C. Detrembleur, T. M. Lejeune, and G. Stoquart, "Upper limb robot-assisted therapy in cerebral palsy: A single-blind randomized controlled trial," *Neurorehabilitation AND Neural Repair*, vol. 29, no. 2, pp. 183–192, February 2015. [Online]: http://nnr.sagepub.com/content/29/2/183.abstract

[11] P. Osten and T. W. Margrie, "Mapping brain circuitry with a light microscope," *Nature Methods*, vol. 10, pp. 515–523, June 2013. [Online]: http://dx.doi.org/10.1038/nmeth.2477

[12] S. M. Gomez-Amaya, M. F. Barbe, W. C. de Groat, J. M. Brown, J. Tuite, Gerald F.AND Corcos, S. B. Fecho, A. S. Braverman, and M. R. Ruggieri Sr, "Neural reconstruction methods of restoring bladder function," *Nature Reviews Urology*, vol. 12, pp. 100–118, February 2015. [Online]: http://dx.doi.org/10.1038/nrurol.2015.4

[13] H. Yu, W. Xiong, H. Zhang, W. Wang, and Z. Li, "A parylene self-locking cuff electrode for peripheral nerve stimulation and recording," *IEEE/ASME Journal of Microelectromechanical Systems*, vol. 23, no. 5, pp. 1025–1035, Oct 2014. [Online]: http://dx.doi.org/10.1109/JMEMS.2014.2333733

[14] J. S. Ho, S. Kim, and A. S. Y. Poon, "Midfield wireless powering for implantable systems," *Proceedings of the IEEE*, vol. 101, no. 6, pp. 1369–1378, June 2013. [Online]: http://dx.doi.org/10.1109/JPROC.2013.2251851

[15] R. D. KEYNES, "Excitable membranes," *Nature*, vol. 239, pp. 29–32, September 1972. [Online]: http://dx.doi.org/10.1038/239029a0

[16] A. D. Grosmark and G. Buzsáki, "Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences," *Science*, vol. 351, no. 6280, pp. 1440–1443, March 2016. [Online]: http://science.sciencemag.org/content/351/6280/1440

[17] B. Sakmann and E. Neher, "Patch clamp techniques for studying ionic channels in excitable membranes," *Annual Review of Physiology*, vol. 46, no. 1, pp. 455–472, October 1984, pMID: 6143532. [Online]: http://dx.doi.org/10.1146/annurev.ph.46.030184.002323

[18] M. P. Ward, P. Rajdev, C. Ellison, and P. P. Irazoqui, "Toward a comparison of microelectrodes for acute and chronic recordings," *Brain Research*, vol. 1282, pp. 183 – 200, July 2009. [Online]: http://www.sciencedirect.com/science/article/pii/S0006899309010841

[19] J. E. B. Randles, "Kinetics of rapid electrode reactions," *Discuss. Faraday Soc.*, vol. 1, pp. 11–19, 1947. [Online]: http://dx.doi.org/10.1039/DF9470100011

[20] M. E. Spira and A. Hai, "Multi-electrode array technologies for neuroscience and cardiology," *Nature Nanotechnology*, vol. 8, pp. 83 – 94, February 2013. [Online]: http://dx.doi.org/10.1038/nnano.2012.265

[21] G. E. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, January 1998. [Online]: http://dx.doi.org/10.1109/JPROC.1998.658762

[22] I. Ferain, C. A. Colinge, and J.-P. Colinge, "Multigate transistors as the future of classical metal-oxide-semiconductor field-effect transistors," *Nature*, vol. 479, pp. 310–316, November 2011. [Online]: http://dx.doi.org/10.1038/nature10676

[23] I. H. Stevenson and K. P. Kording, "How advances in neural recording affect data analysis," *Nature neuroscience*, vol. 14, no. 2, pp. 139–142, February 2011. [Online]: http://dx.doi.org/10.1038/nn.2731

[24] C. Thomas, P. Springer, G. Loeb, Y. Berwald-Netter, and L. Okun, "A miniature microelectrode array to monitor the bioelectric activity of cultured cells," *Experimental cell research*, vol. 74, no. 1, pp. 61–66, September 1972. [Online]: http://dx.doi.org/0.1016/0014-4827(72)90481-8

[25] R. A. Andersen, E. J. Hwang, and G. H. Mulliken, "Cognitive neural prosthetics," *Annual review of Psychology*, vol. 61, pp. 169–190, December 2010, pMID: 19575625. [Online]: http://dx.doi.org/10.1146/annurev.psych.093008.100503

[26] L. A. Jorgenson, W. T. Newsome, D. J. Anderson, C. I. Bargmann, E. N. Brown, K. Deisseroth, J. P. Donoghue, K. L. Hudson, G. S. Ling, P. R. MacLeish *et al.*, "The brain initiative: developing technology to catalyse neuroscience discovery," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 370, no. 1668, p. 20140164, 2015.

[27] E. DAngelo, G. Danese, G. Florimbi, F. Leporati, A. Majani, S. Masoli, S. Solinas, and E. Torti, "The human brain project: High performance computing for brain cells hw/sw simulation and understanding," in *Proceedings of the Digital System Design Conference*, August 2015, pp. 740–747. [Online]: http://dx.doi.org/10.1109/DSD.2015.80

[28] K. Famm, B. Litt, K. J. Tracey, E. S. Boyden, and M. Slaoui, "Drug discovery: a jump-start for electroceuticals," *Nature*, vol. 496, no. 7444, pp. 159–161, April 2013. [Online]: http://dx.doi.org/0.1038/496159a

[29] K. Deisseroth, "Optogenetics," *Nature methods*, vol. 8, no. 1, pp. 26–29, January 2011. [Online]: http://dx.doi.org/10.1038/nmeth.f.324

[30] M. Velliste, S. Perel, M. C. Spalding, A. S. Whitford, and A. B. Schwartz, "Cortical control of a prosthetic arm for self-feeding," *Nature*, vol. 453, no. 7198, pp. 1098–1101, June 2008. [Online]: http://dx.doi.org/10.1038/nature06996

[31] T. N. Theis and P. M. Solomon, "In quest of the "next switch" prospects for greatly reduced power dissipation in a successor to the silicon field-effect transistor," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2005–2014, December 2010. [Online]: http://dx.doi.org/10.1109/JPROC.2010.2066531

[32] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities, reprinted from the afips conference proceedings, vol. 30 (atlantic city, n.j., apr. 18-20), afips press, reston, va., 1967, pp. 483-485, when dr. amdahl was at international business machines corporation, sunnyvale, california," in *AFIPS Conference Proceedings, Vol. 30 (Atlantic City, N.J., Apr. 18-20)*, vol. 12, no. 3. IEEE, Summer 2007, pp. 19–20. [Online]: http://dx.doi.org/0.1109/N-SSC.2007.4785615

[33] J. G. Koller and W. C. Athas, "Adiabatic switching, low energy computing, and the physics of storing and erasing information," in *IEEE Proceedings of the Workshop on Physics and Computation*. IEEE, October 1992, pp. 267–270. [Online]: http://dx.doi.org/10.1109/PHYCMP.1992.615554

[34] E. P. DeBenedictis, J. E. Cook, M. F. Hoemmen, and T. S. Metodi, "Optimal adiabatic scaling and the processor-in-memory-and-storage architecture (oas :pims)," in *IEEE Proceedings of the International Symposium on Nanoscale Architectures*. IEEE, July 2015, pp. 69–74. [Online]: http://dx.doi.org/10.1109/NANOARCH.2015.7180589

[35] S. Houri, G. Billiot, M. Belleville, A. Valentian, and H. Fanet, "Limits of cmos technology and interest of nems relays for adiabatic logic applications," *IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications*, vol. 62, no. 6, pp. 1546–1554, June 2015. [Online]: http://dx.doi.org/10.1109/TCSI.2015.2415177

[36] S. K. Arfin and R. Sarpeshkar, "An energy-efficient, adiabatic electrode stimulator with inductive energy recycling and feedback current regulation," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 1, pp. 1–14, February 2012. [Online]: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6036003&isnumber=6138606

[37] P. R. Kinget, "Scaling analog circuits into deep nanoscale cmos: Obstacles and ways to overcome them," in *IEEE Proceedings of the Custom Integrated Circuits Conference*. IEEE, September 2015, pp. 1–8. [Online]: http://dx.doi.org/10.1109/CICC.2015.7338394

[38] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, "High-performance cmos variability in the 65-nm regime and beyond," *IBM Journal of Research AND Development*, vol. 50, no. 4.5, pp. 433–449, July 2006. [Online]: http://dx.doi.org/10.1147/rd.504.0433

[39] L. L. Lewyn, T. Ytterdal, C. Wulff, and K. Martin, "Analog circuit design in nanoscale cmos technologies," *Proceedings of the IEEE*, vol. 97, no. 10, pp. 1687–1714, October 2009. [Online]: http://dx.doi.org/10.1109/JPROC.2009.2024663

[40] Y. Xin, W. X. Y. Li, Z. Zhang, R. C. C. Cheung, D. Song, and T. W. Berger, "An application specific instruction set processor (asip) for adaptive filters in neural prosthetics," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 5, pp. 1034–1047, September 2015. [Online]: http://dx.doi.org/10.1109/TCBB.2015.2440248

[41] G. Schalk, P. Brunner, L. A. Gerhardt, H. Bischof, and J. R. Wolpaw, "Brain-computer interfaces (bcis): detection instead of classification," *Journal of neuroscience methods*, vol. 167, no. 1, pp. 51–62, 2008, brain-Computer Interfaces (BCIs). [Online]: http://www.sciencedirect.com/science/article/pii/S0165027007004116

[42] Z. Li, J. E. O'Doherty, T. L. Hanson, M. A. Lebedev, C. S. Henriquez, and M. A. Nicolelis, "Unscented kalman filter for brain-machine interfaces," *PloS one*, vol. 4, no. 7, pp. 1–18, 2009. [Online]: http://dx.doi.org/10.1371/journal.pone.0006243

[43] A. L. Orsborn, H. G. Moorman, S. A. Overduin, M. M. Shanechi, D. F. Dimitrov, and J. M. Carmena, "Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control," *Neuron*, vol. 82, pp. 1380 – 1393, March 2016. [Online]: http://dx.doi.org/10.1016/j.neuron.2014.04.048

[44] Y. Yan, X. Qin, Y. Wu, N. Zhang, J. Fan, and L. Wang, "A restricted boltzmann machine based two-lead electrocardiography classification," in *IEEE Proceedings of the International Conference on Wearable and Implantable Body Sensor Networks*. IEEE, June 2015, pp. 1–9. [Online]: http://dx.doi.org/10.1109/BSN.2015.7299399

[45] B. M. Yu and J. P. Cunningham, "Dimensionality reduction for large-scale neural recordings," *Nature Neuroscience*, vol. 17, pp. 1500 – 1509, November 2014. [Online]: http://dx.doi.org/10.1038/nn.3776

[46] S. Makeig, C. Kothe, T. Mullen, N. Bigdely-Shamlo, Z. Zhang, and K. Kreutz-Delgado, "Evolving signal processing for brain: Computer interfaces," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1567–1584, May 2012. [Online]: http://dx.doi.org/10.1109/JPROC.2012.2185009

[47] G. Indiveri and S. C. Liu, "Memory and information processing in neuromorphic systems," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1379–1397, August 2015. [Online]: http://dx.doi.org/10.1109/JPROC.2015.2444094

[48] Y. Chen, E. Yao, and A. Basu, "A 128-channel extreme learning machine-based neural decoder for brain machine interfaces," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 3, pp. 679–692, June 2016. [Online]: http://dx.doi.org/10.1109/TBCAS.2015.2483618

[49] V. Karkare, S. Gibson, and D. Markovi, "A 75- μw, 16-channel neural spike-sorting processor with unsupervised clustering," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 9, pp. 2230–2238, September 2013. [Online]: http://dx.doi.org/10.1109/JSSC.2013.2264616

[50] T. C. Chen, W. Liu, and L. G. Chen, "128-channel spike sorting processor with a parallel-folding structure in 90nm process," in *IEEE Proceedings of the International Symposium on Circuits and Systems*, May 2009, pp. 1253–1256. [Online]: http://dx.doi.org/10.1109/ISCAS.2009.5117990

[51] G. Baranauskas, "What limits the performance of current invasive brain machine interfaces?" *Frontiers in Systems Neuroscience*, vol. 8, no. 68, April 2014. [Online]: http://www.frontiersin.org/systems_neuroscience/10.3389/fnsys.2014.00068

[52] E. F. Chang, "Towards large-scale, human-based, mesoscopic neurotechnologies," *Neuron*, vol. 86, pp. 68–78, March 2016. [Online]: http://dx.doi.org/10.1016/j.neuron.2015.03.037

[53] M. A. L. Nicolelis and M. A. Lebedev, "Principles of neural ensemble physiology underlying the operation of brain-machine," *Nature Reviews Neuroscience*, vol. 10, pp. 530–540, July 2009. [Online]: http://dx.doi.org/10.1038/nrn2653

[54] Z. Fekete, "Recent advances in silicon-based neural microelectrodes and microsystems: a review," *Sensors AND Actuators B: Chemical*, vol. 215, pp. 300 – 315, 2015. [Online]: http://www.sciencedirect.com/science/article/pii/S092540051500386X

[55] N. Saeidi, M. Schuettler, A. Demosthenous, and N. Donaldson, "Technology for integrated circuit micropackages for neural interfaces, based on goldsilicon wafer bonding," *Journal of Micromechanics AND Microengineering*, vol. 23, no. 7, p. 075021, June 2013. [Online]: http://stacks.iop.org/0960-1317/23/i=7/a=075021

[56] K. Seidl, S. Herwik, T. Torfs, H. P. Neves, O. Paul, and P. Ruther, "Cmos-based high-density silicon microprobe arrays for electronic depth control in intracortical neural recording," *IEEE Journal of Microelectromechanical Systems*, vol. 20, no. 6, pp. 1439–1448, December 2011. [Online]: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6033040&isnumber=6075219

[57] T. D. Y. Kozai, N. B. Langhals, P. R. Patel, X. Deng, H. Zhang, K. L. Smith, J. Lahann, N. A. Kotov, and D. R. Kipke, "Ultrasmall implantable composite microelectrodes with bioactive surfaces for chronic neural interfaces," *Nature Materials*, vol. 11, pp. 1065–1073, December 2012. [Online]: http://dx.doi.org/10.1038/nmat3468

[58] D. A. Schwarz, M. A. Lebedev, T. L. Hanson, D. F. Dimitrov, G. Lehew, J. Meloy, S. Rajangam, V. Subramanian, P. J. Ifft, Z. Li, A. Ramakrishnan, A. Tate, K. Z. Zhuang, and M. A. L. Nicolelis, "Chronic, wireless recordings of large-scale brain

activity in freely moving rhesus monkeys," *Nature Methods*, vol. 11, pp. 670–676, April 2014. [Online]: http://dx.doi.org/10.1038/nmeth.2936

[59] 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, no. 5, pp. 505–515, 2010. [Online]: http://dx.doi.org/10.1002/tee.20566

[60] H.-W. Kang, S. J. Lee, I. K. Ko, C. Kengla, J. J. Yoo, and A. Atala, "A 3d bioprinting system to produce human-scale tissue constructs with structural integrity," *Nature Biotechnology*, vol. 34, pp. 312–319, March 2016. [Online]: http://dx.doi.org/10.1038/nbt.3413

[61] C. Xie, J. Liu, T.-M. Fu, X. Dai, W. Zhou, and C. M. Lieber, "Three-dimensional macroporous nanoelectronic networks as minimally invasive brain probes," *Nature Materials*, vol. 14, pp. 1286–1292, May 2015. [Online]: http://dx.doi.org/10.1038/nmat4427

[62] 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 Journal of Solid-State Circuits*, vol. 42, no. 1, pp. 123–133, Jan 2007. [Online]: http://dx.doi.org/10.1109/JSSC.2006.886567

[63] J. Guo, W. Ng, J. Yuan, S. Li, and M. Chan, "A 200-channel area-power-efficient chemical and electrical dual-mode acquisition ic for the study of neurodegenerative diseases," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 3, pp. 567–578, June 2016. [Online]: http://dx.doi.org/10.1109/TBCAS.2015.2468052

[64] W. Biederman, D. J. Yeager, N. Narevsky, J. Leverett, R. Neely, J. M. Carmena, E. Alon, and J. M. Rabaey, "A 4.78 mm 2 fully-integrated neuromodulation soc combining 64 acquisition channels with digital compression and simultaneous dual stimulation," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 4, pp. 1038–1047, April 2015. [Online]: http://dx.doi.org/10.1109/JSSC.2014.2384736

[65] C. M. Lopez, A. Andrei, S. Mitra, M. Welkenhuysen, W. Eberle, C. Bartic, R. Puers, R. F. Yazicioglu, and G. G. E. Gielen, "An implantable 455-active-electrode 52-channel cmos neural probe," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 248–261, January 2014. [Online]: http://dx.doi.org/10.1109/JSSC.2013.2284347

[66] P. D. Wolf, *Thermal considerations for the design of an implanted cortical brain–machine interface (BMI).* CRC Press Boca Raton, FL, 2008, pMID: 21204402. [Online]: http://www.ncbi.nlm.nih.gov/books/NBK3932

[67] T. Denison, K. Consoer, W. Santa, A. T. Avestruz, J. Cooley, and A. Kelly, "A 2 μw 100 nv/rthz chopper-stabilized instrumentation amplifier for chronic measurement of neural field potentials," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 12, pp. 2934–2945, December 2007. [Online]: http://dx.doi.org/10.1109/JSSC.2007.908664

[68] B. Johnson, S. T. Peace, A. Wang, T. A. Cleland, and A. Molnar, "A 768-channel cmos microelectrode array with angle sensitive pixels for neuronal recording," *IEEE Sensors Journal*, vol. 13, no. 9, pp. 3211–3218, Sept 2013. [Online]: http://dx.doi.org/10.1109/JSEN.2013.2266894

[69] H. Kassiri, A. Bagheri, N. Soltani, K. Abdelhalim, H. M. Jafari, M. T. Salam, J. L. P. Velazquez, and R. Genov, "Battery-less tri-band-radio neuro-monitor and responsive neurostimulator for diagnostics and treatment of neurological disorders," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 5, pp. 1274–1289, May 2016. [Online]: http://dx.doi.org/10.1109/JSSC.2016.2528999

[70] M. Ballini, J. Mller, P. Livi, Y. Chen, U. Frey, A. Stettler, A. Shadmani, V. Viswam, I. L. Jones, D. Jckel, M. Radivojevic, M. K. Lewandowska, W. Gong, M. Fiscella, D. J. Bakkum, F. Heer, and A. Hierlemann, "A 1024-channel cmos microelectrode array with 26,400 electrodes for recording and stimulation of electrogenic cells in vitro," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 11, pp. 2705–2719, Nov 2014. [Online]: http://dx.doi.org/10.1109/JSSC.2014.2359219

[71] J. Scholvin, J. P. Kinney, J. G. Bernstein, C. Moore-Kochlacs, N. Kopell, C. G. Fonstad, and E. S. Boyden, "Close-packed silicon microelectrodes for scalable spatially oversampled neural recording," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 1, pp. 120–130, Jan 2016. [Online]: http://dx.doi.org/10.1109/TBME.2015.2406113

[72] M. Han, B. Kim, Y. A. Chen, H. Lee, S. H. Park, E. Cheong, J. Hong, G. Han, and Y. Chae, "Bulk switching instrumentation amplifier for a high-impedance source in neural signal recording," *IEEE Transactions on Circuits and Systems—Part II: Express Briefs*, vol. 62, no. 2, pp. 194–198, Feb 2015. [Online]: http://dx.doi.org/10.1109/TCSII.2014.2368615

[73] R. Muller, S. Gambini, and J. M. Rabaey, "A 0.013 mm^2, 5 μw, dc-coupled neural signal acquisition ic with 0.5 v supply," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 232–243, Jan 2012. [Online]: http://dx.doi.org/10.1109/JSSC.2011.2163552

[74] "Rhd2164 digital electrophysiology interface chip - data sheet," Intan Technologies, Los Angeles, California, December 2013. [Online]: http://www.intantech.com/files/Intan_RHD2164_datasheet.pdf

[75] 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 Transactions on Biomedical Circuits and Systems*, vol. 6, no. 5, pp. 403–413, October 2012. [Online]: http://dx.doi.org/10.1109/TBCAS.2012.2218105

[76] D. Han, Y. Zheng, R. Rajkumar, G. S. Dawe, and M. Je, "A 0.45 v 100-channel neural-recording ic with sub-μw/channel consumption in 0.18μm cmos," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 6, pp. 735–746, December 2013. [Online]: http://dx.doi.org/10.1109/TBCAS.2014.2298860

[77] R. Muller, S. Gambini, and J. M. Rabaey, "A 0.013mm^2, 5 μw, dc-coupled neural signal acquisition ic with 0.5v supply," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 232–243, Jan 2012. [Online]: http://dx.doi.org/10.1109/JSSC.2011.2163552

[78] 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," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 4, no. 6, pp. 360–371, Dec 2010. [Online]: http://dx.doi.org/10.1109/TBCAS.2010.2078814

[79] J. Yoo, L. Yan, D. El-Damak, M. A. B. Altaf, A. H. Shoeb, and A. P. Chandrakasan, "An 8-channel scalable eeg acquisition soc with patient-specific seizure classification and recording processor," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 214–228, Jan 2013. [Online]: http://dx.doi.org/10.1109/JSSC.2012.2221220

[80] M. A. B. Altaf and J. Yoo, "A 1.83 μj/classification, 8-channel, patient-specific epileptic seizure classification soc using a non-linear support vector machine," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 1, pp. 49–60, Feb 2016. [Online]: http://dx.doi.org/10.1109/TBCAS.2014.2386891

[81] K. Abdelhalim, H. M. Jafari, L. Kokarovtseva, J. L. P. Velazquez, and R. Genov, "64-channel uwb wireless neural vector analyzer soc with a closed-loop phase synchrony-triggered neurostimulator," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 10, pp. 2494–2510, Oct 2013. [Online]: http://dx.doi.org/10.1109/JSSC.2013.2272952

[82] A. Bagheri, S. R. I. Gabran, M. T. Salam, J. L. P. Velazquez, R. R. Mansour, M. M. A. Salama, and R. Genov, "Massively-parallel neuromonitoring and neurostimulation rodent headset with nanotextured flexible microelectrodes," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 5, pp. 601–609, Oct 2013. [Online]: http://dx.doi.org/10.1109/TBCAS.2013.2281772

[83] H. G. Rhew, J. Jeong, J. A. Fredenburg, S. Dodani, P. G. Patil, and M. P. Flynn, "A fully self-contained logarithmic closed-loop deep brain stimulation soc with wireless telemetry and wireless power management," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2213–2227, Oct 2014. [Online]: http://dx.doi.org/10.1109/JSSC.2014.2346779

[84] W. Biederman, D. J. Yeager, N. Narevsky, J. Leverett, R. Neely, J. M. Carmena, E. Alon, and J. M. Rabaey, "A 4.78 mm 2 fully-integrated neuromodulation soc

combining 64 acquisition channels with digital compression and simultaneous dual stimulation," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 4, pp. 1038–1047, April 2015. [Online]: http://dx.doi.org/10.1109/JSSC.2014.2384736

[85] A. Mendez, A. Belghith, and M. Sawan, "A dsp for sensing the bladder volume through afferent neural pathways," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 4, pp. 552–564, Aug 2014. [Online]: http://dx.doi.org/10.1109/TBCAS.2013.2282087

[86] T. T. Liu and J. M. Rabaey, "A 0.25 v 460 nw asynchronous neural signal processor with inherent leakage suppression," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 4, pp. 897–906, April 2013. [Online]: http://dx.doi.org/10.1109/JSSC.2013.2239096

[87] D. Han, Y. Zheng, R. Rajkumar, G. S. Dawe, and M. Je, "A 0.45 v 100-channel neural-recording ic with sub-μw/channel consumption in 0.18 μm cmos," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 6, pp. 735–746, Dec 2013. [Online]: http://dx.doi.org/10.1109/TBCAS.2014.2298860

[88] R. Muller, H. P. Le, W. Li, P. Ledochowitsch, S. Gambini, T. Bjorninen, A. Koralek, J. M. Carmena, M. M. Maharbiz, E. Alon, and J. M. Rabaey, "A minimally invasive 64-channel wireless μecog implant," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 344–359, Jan 2015. [Online]: http://dx.doi.org/10.1109/JSSC.2014.2364824

[89] B. Vigraham, J. Kuppambatti, and P. R. Kinget, "Switched-mode operational amplifiers and their application to continuous-time filters in nanoscale cmos," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 12, pp. 2758–2772, December 2014. [Online]: http://dx.doi.org/10.1109/JSSC.2014.2354641

[90] V. Karkare, H. Chandrakumar, D. Rozgi, and D. Markovi, "Robust, reconfigurable, and power-efficient biosignal recording systems," in *IEEE Proceedings of the Custom Integrated Circuits Conference*, Sept 2014, pp. 1–8. [Online]: http://dx.doi.org/10.1109/CICC.2014.6946018

[91] L. B. Leene and T. G. Constandinou, "A 0.45v continuous time-domain filter using asynchronous oscillator structures," in *IEEE Proceedings of the International Conference on Electronics, Circuits and Systems*, December 2016.

[92] R. Mohan, L. Yan, G. Gielen, C. V. Hoof, and R. F. Yazicioglu, "0.35 v time-domain-based instrumentation amplifier," *Electronics Letters*, vol. 50, no. 21, pp. 1513–1514, October 2014. [Online]: http://dx.doi.org/10.1049/el.2014.2471

[93] X. Zhang, Z. Zhang, Y. Li, C. Liu, Y. X. Guo, and Y. Lian, "A 2.89 μw dry-electrode enabled clockless wireless ecg soc for wearable applications," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 10, pp. 2287–2298, Oct 2016. [Online]: http://dx.doi.org/10.1109/JSSC.2016.2582863

[94] M. Elia, L. B. Leene, and T. G. Constandinou, "Continuous-time micropower interface for neural recording applications," in *IEEE Proceedings of the International Symposium on Circuits and Systems*, May 2016, pp. 534–537. [Online]: http://dx.doi.org/10.1109/ISCAS.2016.7527295

[95] N. Guo, Y. Huang, T. Mai, S. Patil, C. Cao, M. Seok, S. Sethumadhavan, and Y. Tsividis, "Energy-efficient hybrid analog/digital approximate computation in continuous time," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 7, pp. 1514–1524, July 2016. [Online]: http://dx.doi.org/10.1109/JSSC.2016.2543729

[96] B. Bozorgzadeh, D. R. Schuweiler, M. J. Bobak, P. A. Garris, and P. Mohseni, "Neurochemostat: A neural interface soc with integrated chemometrics for closed-loop regulation of brain dopamine," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 3, pp. 654–667, June 2016. [Online]: http://dx.doi.org/10.1109/TBCAS.2015.2453791

[97] E. B. Myers and M. L. Roukes, "Comparative advantages of mechanical biosensors," *Nature nanotechnology*, vol. 6, no. 4, pp. 1748–3387, April 2011. [Online]: http://dx.doi.org/10.1038/nnano.2011.44

[98] R. Machado, N. Soltani, S. Dufour, M. T. Salam, P. L. Carlen, R. Genov, and M. Thompson, "Biofouling-resistant impedimetric sensor for array high-resolution extracellular potassium monitoring in the brain," *Biosensors*, vol. 6, no. 4, p. 53, October 2016. [Online]: http://dx.doi.org/10.3390/bios6040053

[99] J. Guo, W. Ng, J. Yuan, S. Li, and M. Chan, "A 200-channel area-power-efficient chemical and electrical dual-mode acquisition ic for the study of neurodegenerative diseases," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 3, pp. 567–578, June 2016. [Online]: http://dx.doi.org/10.1109/TBCAS.2015.2468052

[100] D. A. Dombeck, A. N. Khabbaz, F. Collman, T. L. Adelman, and D. W. Tank, "Imaging large-scale neural activity with cellular resolution in awake, mobile mice." *Neuron*, vol. 56, no. 1, pp. 43–57, October 2007. [Online]: http://dx.doi.org/10.1016/j.neuron.2007.08.003

[101] T. York, S. B. Powell, S. Gao, L. Kahan, T. Charanya, D. Saha, N. W. Roberts, T. W. Cronin, J. Marshall, S. Achilefu, S. P. Lake, B. Raman, and V. Gruev, "Bioinspired polarization imaging sensors: From circuits and optics to signal processing algorithms and biomedical applications," *Proceedings of the IEEE*, vol. 102, no. 10, pp. 1450–1469, Oct 2014. [Online]: http://dx.doi.org/10.1109/JPROC.2014.2342537

[102] K. Paralikar, P. Cong, O. Yizhar, L. E. Fenno, W. Santa, C. Nielsen, D. Dinsmoor, B. Hocken, G. O. Munns, J. Giftakis, K. Deisseroth, and T. Denison, "An implantable optical stimulation delivery system for actuating an excitable biosubstrate," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 321–332, Jan 2011. [Online]: http://dx.doi.org/10.1109/JSSC.2010.2074110

[103] N. Ji and S. L. Smith, "Technologies for imaging neural activity in large volumes," *Nature Neuroscience*, vol. 19, pp. 1154–1164, September 2016. [Online]: http://dx.doi.org/10.1038/nn.4358

[104] S. Song, K. D. Miller, and L. F. Abbott, "Competitive hebbian learning through spike-timing-dependent synaptic plasticity," *Nature Neuroscience*, vol. 3, pp. 919–926, September 2000. [Online]: http://dx.doi.org/10.1038/78829

[105] T. Kurafuji, M. Haraguchi, M. Nakajima, T. Nishijima, T. Tanizaki, H. Yamasaki, T. Sugimura, Y. Imai, M. Ishizaki, T. Kumaki, K. Murata, K. Yoshida, E. Shimomura, H. Noda, Y. Okuno, S. Kamijo, T. Koide, H. J. Mattausch, and K. Arimoto, "A scalable massively parallel processor for real-time image processing," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 10, pp. 2363–2373, October 2011. [Online]: http://dx.doi.org/10.1109/JSSC.2011.2159528

[106] J. Y. Kim, M. Kim, S. Lee, J. Oh, K. Kim, and H. J. Yoo, "A 201.4 gops 496 mw real-time multi-object recognition processor with bio-inspired neural perception engine," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 1, pp. 32–45, Jan 2010. [Online]: http://dx.doi.org/10.1109/JSSC.2009.2031768

[107] C. C. Cheng, C. H. Lin, C. T. Li, and L. G. Chen, "ivisual: An intelligent visual sensor soc with 2790 fps cmos image sensor and 205 gops/w vision processor," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 127–135, Jan 2009. [Online]: http://dx.doi.org/10.1109/JSSC.2008.2007158

[108] H. Noda, M. Nakajima, K. Dosaka, K. Nakata, M. Higashida, O. Yamamoto, K. Mizumoto, T. Tanizaki, T. Gyohten, Y. Okuno, H. Kondo, Y. Shimazu, K. Arimoto, K. Saito, and T. Shimizu, "The design and implementation of the massively parallel processor based on the matrix architecture," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 1, pp. 183–192, Jan 2007. [Online]: http://dx.doi.org/10.1109/JSSC.2006.886545

[109] M. S. Chae, W. Liu, and M. Sivaprakasam, "Design optimization for integrated neural recording systems," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 9, pp. 1931–1939, September 2008. [Online]: http://dx.doi.org/10.1109/JSSC.2008.2001877

[110] K. J. Miller, L. B. Sorensen, J. G. Ojemann, and M. den Nijs, "Power-law scaling in the brain surface electric potential," *PLoS Comput Biol*, vol. 5, no. 12, pp. 1–10, 12 2009. [Online]: http://dx.doi.org/10.1371%2Fjournal.pcbi.1000609

[111] R. Harrison and C. Charles, "A low-power low-noise cmos amplifier for neural recording applications," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 6, pp. 958–965, June 2003. [Online]: http://dx.doi.org/10.1109/JSSC.2003.811979

[112] W. Sansen, "1.3 analog cmos from 5 micrometer to 5 nanometer," in *IEEE Proceedings of the International Solid-State Circuits Conference*. IEEE, February 2015, pp. 1–6. [Online]: http://dx.doi.org/10.1109/ISSCC.2015.7062848

[113] M. S. J. Steyaert and W. M. C. Sansen, "A micropower low-noise monolithic instrumentation amplifier for medical purposes," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 6, pp. 1163–1168, December 1987. [Online]: http://dx.doi.org/10.1109/JSSC.1987.1052869

[114] W. Wattanapanitch, M. Fee, and R. Sarpeshkar, "An energy-efficient micropower neural recording amplifier," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 1, no. 2, pp. 136–147, June 2007. [Online]: http://dx.doi.org/10.1109/TBCAS.2007.907868

[115] B. Johnson and A. Molnar, "An orthogonal current-reuse amplifier for multi-channel sensing," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 6, pp. 1487–1496, June 2013. [Online]: http://dx.doi.org/10.1109/JSSC.2013.2257478

[116] C. Qian, J. Parramon, and E. Sanchez-Sinencio, "A micropower low-noise neural recording front-end circuit for epileptic seizure detection," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 6, pp. 1392–1405, June 2011. [Online]: http://dx.doi.org/10.1109/JSSC.2011.2126370

[117] X. Zou, L. Liu, J. H. Cheong, L. Yao, P. Li, M.-Y. Cheng, W. L. Goh, R. Rajkumar, G. Dawe, K.-W. Cheng, and M. Je, "A 100-channel 1-mw implantable neural recording ic," *IEEE Transactions on Circuits and Systems—Part I: Regular Papers*, vol. 60, no. 10, pp. 2584–2596, October 2013. [Online]: http://dx.doi.org/10.1109/TCSI.2013.2249175

[118] V. Majidzadeh, A. Schmid, and Y. Leblebici, "Energy efficient low-noise neural recording amplifier with enhanced noise efficiency factor," *IEEE Transactions on*

Biomedical Circuits and Systems, vol. 5, no. 3, pp. 262–271, June 2011. [Online]: http://dx.doi.org/10.1109/TBCAS.2010.2078815

[119] C. C. Enz and E. A. Vittoz, *Charge-based MOS transistor modeling: the EKV model for low-power AND RF IC design.* John Wiley & Sons, August 2006. [Online]: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470855452.html

[120] Y. Yasuda, T.-J. K. Liu, and C. Hu, "Flicker-noise impact on scaling of mixed-signal cmos with hfsion," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 417–422, January 2008. [Online]: http://dx.doi.org/10.1109/TED.2007.910759

[121] S.-Y. Wu, C. Lin, M. Chiang, J. Liaw, J. Cheng, S. Yang, M. Liang, T. Miyashita, C. Tsai, B. Hsu, H. Chen, T. Yamamoto, S. Chang, V. Chang, C. Chang, J. Chen, H. Chen, K. Ting, Y. Wu, K. Pan, R. Tsui, C. Yao, P. Chang, H. Lien, T. Lee, H. Lee, W. Chang, T. Chang, R. Chen, M. Yeh, C. Chen, Y. Chiu, Y. Chen, H. Huang, Y. Lu, C. Chang, M. Tsai, C. Liu, K. Chen, C. Kuo, H. Lin, S. Jang, and Y. Ku, "A 16nm finfet cmos technology for mobile soc and computing applications," in *IEEE Proceedings of the International Electron Devices Meeting*, December 2013, pp. 9.1.1–9.1.4. [Online]: http://dx.doi.org/10.1109/IEDM.2013.6724591

[122] L. B. Leene, Y. Liu, and T. G. Constandinou, "A compact recording array for neural interfaces," in *IEEE Proceedings of the Biomedical Circuits and Systems Conference*, October 2013, pp. 97–100. [Online]: http://dx.doi.org/10.1109/BioCAS.2013.6679648

[123] Q. Fan, F. Sebastiano, J. Huijsing, and K. Makinwa, "A $1.8\,\mu w\,60\,nv/\sqrt{Hz}$ capacitively-coupled chopper instrumentation amplifier in 65 nm cmos for wireless sensor nodes," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 7, pp. 1534–1543, July 2011. [Online]: http://dx.doi.org/10.1109/JSSC.2011.2143610

[124] H. Chandrakumar and D. Markovic, "A simple area-efficient ripple-rejection technique for chopped biosignal amplifiers," *IEEE Transactions on Circuits and Systems—Part II: Express Briefs*, vol. 62, no. 2, pp. 189–193, February 2015. [Online]: http://dx.doi.org/10.1109/TCSII.2014.2387686

[125] H. Chandrakumar and D. Markovic, "A 2μw 40mvpp linear-input-range chopper-stabilized bio-signal amplifier with boosted input impedance of 300mohm and electrode-offset filtering," in *IEEE Proceedings of the International Solid-State Circuits Conference*. IEEE, January 2016, pp. 96–97. [Online]: http://dx.doi.org/10.1109/ISSCC.2016.7417924

[126] H. Rezaee-Dehsorkh, N. Ravanshad, R. Lotfi, K. Mafinezhad, and A. M. Sodagar, "Analysis and design of tunable amplifiers for implantable neural recording applications," *IEEE Transactions on Emerging and Selected Topics in Circuits and Systems*, vol. 1, no. 4, pp. 546–556, December 2011. [Online]: http://dx.doi.org/10.1109/JETCAS.2011.2174492

[127] X. Zou, X. Xu, L. Yao, and Y. Lian, "A 1-v 450-nw fully integrated programmable biomedical sensor interface chip," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1067–1077, April 2009. [Online]: http://dx.doi.org/10.1109/JSSC.2009.2014707

[128] L. Leene and T. Constandinou, "Ultra-low power design strategy for two-stage amplifier topologies," *Electronics Letters*, vol. 50, no. 8, pp. 583–585, April 2014. [Online]: http://dx.doi.org/10.1049/el.2013.4196

[129] H. G. Rey, C. Pedreira, and R. Q. Quiroga, "Past, present and future of spike sorting techniques," *Brain Research Bulletin*, vol. 119, Part B, pp. 106–117, October 2015, advances in electrophysiological data analysis. [Online]: http://www.sciencedirect.com/science/article/pii/S0361923015000684

[130] Y. Chen, A. Basu, L. Liu, X. Zou, R. Rajkumar, G. S. Dawe, and M. Je, "A digitally assisted, signal folding neural recording amplifier," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 4, pp. 528–542, August 2014. [Online]: http://dx.doi.org/10.1109/TBCAS.2013.2288680

[131] X. Yue, "Determining the reliable minimum unit capacitance for the dac capacitor array of sar adcs," *Microelectronics Journal*, vol. 44, no. 6, pp. 473 – 478, 2013. [Online]: http://www.sciencedirect.com/science/article/pii/S0026269213000815

[132] Y. Zhu, C.-H. Chan, U.-F. Chio, S.-W. Sin, S.-P. U, R. Martins, and F. Maloberti, "Split-sar adcs: Improved linearity with power and speed optimization," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 2, pp. 372–383, February 2014. [Online]: http://dx.doi.org/10.1109/TVLSI.2013.2242501

[133] L. Xie, G. Wen, J. Liu, and Y. Wang, "Energy-efficient hybrid capacitor switching scheme for sar adc," *Electronics Letters*, vol. 50, no. 1, pp. 22–23, January 2014. [Online]: http://dx.doi.org/10.1049/el.2013.2794

[134] P. Nuzzo, F. De Bernardinis, P. Terreni, and G. Van der Plas, "Noise analysis of regenerative comparators for reconfigurable adc architectures," *IEEE Transactions on Circuits and Systems—Part I: Regular Papers*, vol. 55, no. 6, pp. 1441–1454, July 2008. [Online]: http://dx.doi.org/10.1109/TCSI.2008.917991

[135] G. Heinzel, A. Rüdiger, and R. Schilling, "Spectrum and spectral density estimation by the discrete fourier transform (dft), including a comprehensive list of window functions and some new at-top windows," pp. 25–27, February 2002. [Online]: http://hdl.handle.net/11858/00-001M-0000-0013-557A-5

[136] F. Gerfers, M. Ortmanns, and Y. Manoli, "A 1.5-v 12-bit power-efficient continuous-time third-order sigma; delta; modulator," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 8, pp. 1343–1352, Aug 2003. [Online]: http://dx.doi.org/10.1109/JSSC.2003.814432

[137] Y. Chae, K. Souri, and K. A. A. Makinwa, "A 6.3 μw 20 bit incremental zoom-adc with 6 ppm inl and 1 μv offset," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 12, pp. 3019–3027, Dec 2013. [Online]: http://dx.doi.org/10.1109/JSSC.2013.2278737

[138] Y. S. Shu, L. T. Kuo, and T. Y. Lo, "An oversampling sar adc with dac mismatch error shaping achieving 105db sfdr and 101db sndr over 1khz bw in 55nm cmos," in *IEEE Proceedings of the International Solid-State Circuits Conference*, January 2016, pp. 458–459. [Online]: http://dx.doi.org/10.1109/ISSCC.2016.7418105

[139] P. Harpe, E. Cantatore, and A. van Roermund, "An oversampled 12/14b sar adc with noise reduction and linearity enhancements achieving up to 79.1db sndr," in *IEEE*

Proceedings of the International Solid-State Circuits Conference, February 2014, pp. 194–195. [Online]: http://dx.doi.org/10.1109/ISSCC.2014.6757396

[140] M. Braverman, J. Schneider, and C. Rojas, "Space-bounded church-turing thesis and computational tractability of closed systems," *Physical Review Letters*, vol. 115, August 2015. [Online]: http://link.aps.org/doi/10.1103/PhysRevLett.115.098701

[141] M. Verhelst and A. Bahai, "Where analog meets digital: Analog-to-information conversion and beyond," *IEEE Solid-State Circuits Magazine*, vol. 7, no. 3, pp. 67–80, September 2015. [Online]: http://dx.doi.org/10.1109/MSSC.2015.2442394

[142] H. A. Marblestone, M. B. Zamft, G. Y. Maguire, G. M. Shapiro, R. T. Cybulski, I. J. Glaser, D. Amodei, P. B. Stranges, R. Kalhor, A. D. Dalrymple, D. Seo, E. Alon, M. M. Maharbiz, M. J. Carmena, M. J. Rabaey, S. E. Boyden, M. G. Church, and P. K. Kording, "Physical principles for scalable neural recording," *Frontiers in Computational Neuroscience*, vol. 7, no. 137, 2013. [Online]: http://www.frontiersin.org/computational_neuroscience/10.3389/fncom.2013.00137

[143] L. Traver, C. Tarin, P. Marti, and N. Cardona, "Adaptive-threshold neural spike by noise-envelope tracking," *Electronics Letters*, vol. 43, no. 24, pp. 1333–1335, November 2007. [Online]: http://dx.doi.org/10.1049/el:20071631

[144] I. Obeid and P. Wolf, "Evaluation of spike-detection algorithms fora brain-machine interface application," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 905–911, June 2004. [Online]: http://dx.doi.org/10.1109/TBME.2004.826683

[145] P. Watkins, G. Santhanam, K. Shenoy, and R. Harrison, "Validation of adaptive threshold spike detector for neural recording," in *IEEE Proceedings of the International Conference on Engineering in Medicine and Biology Society*, vol. 2, September 2004, pp. 4079–4082. [Online]: http://dx.doi.org/10.1109/IEMBS.2004.1404138

[146] T. Takekawa, Y. Isomura, and T. Fukai, "Accurate spike sorting for multi-unit recordings," *European Journal of Neuroscience*, vol. 31, no. 2, pp. 263–272, 2010. [Online]: http://dx.doi.org/10.1111/j.1460-9568.2009.07068.x

[147] A. Zviagintsev, Y. Perelman, and R. Ginosar, "Low-power architectures for spike sorting," in *IEEE Proceedings of the International Conference on Neural Engineering*, March 2005, pp. 162–165. [Online]: http://dx.doi.org/10.1109/CNE.2005.1419579

[148] A. Rodriguez-Perez, J. Ruiz-Amaya, M. Delgado-Restituto, and A. Rodriguez-Vazquez, "A low-power programmable neural spike detection channel with embedded calibration and data compression," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 2, pp. 87–100, April 2012. [Online]: http://dx.doi.org/10.1109/TBCAS.2012.2187352

[149] U. Rutishauser, E. M. Schuman, and A. N. Mamelak, "Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo," *Journal of Neuroscience Methods*, vol. 154, no. 12, pp. 204 – 224, 2006. [Online]: http://www.sciencedirect.com/science/article/pii/S0165027006000033

[150] F. Franke, M. Natora, C. Boucsein, M. Munk, and K. Obermayer, "An online spike detection and spike classification algorithm capable of instantaneous resolution of overlapping spikes," *Journal of Computational Neuroscience*, vol. 29, no. 1-2, pp. 127–148, 2010. [Online]: http://dx.doi.org/10.1007/s10827-009-0163-5

[151] M. S. Chae, Z. Yang, M. Yuce, L. Hoang, and W. Liu, "A 128-channel 6 mw wireless neural recording ic with spike feature extraction and uwb transmitter," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 4, pp. 312–321, August 2009. [Online]: http://dx.doi.org/10.1109/TNSRE.2009.2021607

[152] P. H. Thakur, H. Lu, S. S. Hsiao, and K. O. Johnson, "Automated optimal detection and classification of neural action potentials in extra-cellular recordings," *Journal of Neuroscience Methods*, vol. 162, no. 12, pp. 364 – 376, 2007. [Online]: ttp://www.sciencedirect.com/science/article/pii/S0165027007000477

[153] J. Zhang, Y. Suo, S. Mitra, S. Chin, S. Hsiao, R. Yazicioglu, T. Tran, and R. Etienne-Cummings, "An efficient and compact compressed sensing microsystem for implantable neural recordings," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 4, pp. 485–496, August 2014. [Online]: http://dx.doi.org/10.1109/TBCAS.2013.2284254

[154] Y. Suo, J. Zhang, T. Xiong, P. S. Chin, R. Etienne-Cummings, and T. D. Tran, "Energy-efficient multi-mode compressed sensing system for implantable neural recordings," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 5, pp. 648–659, October 2014. [Online]: http://dx.doi.org/10.1109/TBCAS.2014.2359180

[155] B. Yu, T. Mak, X. Li, F. Xia, A. Yakovlev, Y. Sun, and C. S. Poon, "Real-time fpga-based multichannel spike sorting using hebbian eigenfilters," *IEEE Transactions on Emerging and Selected Topics in Circuits and Systems*, vol. 1, no. 4, pp. 502–515, December 2011. [Online]: http://dx.doi.org/10.1109/JETCAS.2012.2183430

[156] V. Ventura, "Automatic spike sorting using tuning information," *Neural computation*, vol. 21, no. 9, pp. 2466–2501, September 2009. [Online]: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4167425/

[157] D. Y. Barsakcioglu, A. Eftekhar, and T. G. Constandinou, "Design optimisation of front-end neural interfaces for spike sorting systems," in *IEEE Proceedings of the International Symposium on Circuits and Systems*, May 2013, pp. 2501–2504. [Online]: http://dx.doi.org/10.1109/ISCAS.2013.6572387

[158] A. M. Sodagar, K. D. Wise, and K. Najafi, "A fully integrated mixed-signal neural processor for implantable multichannel cortical recording," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 6, pp. 1075–1088, June 2007. [Online]: http://dx.doi.org/10.1109/TBME.2007.894986

[159] Y. Xin, W. X. Li, R. C. Cheung, R. H. Chan, H. Yan, D. Song, and T. W. Berger, "An fpga based scalable architecture of a stochastic state point process filter (ssppf) to track the nonlinear dynamics underlying neural spiking," *Microelectronics Journal*, vol. 45, no. 6, pp. 690 – 701, June 2014. [Online]: http://www.sciencedirect.com/science/article/pii/S0026269214000913

[160] C. Qian, J. Shi, J. Parramon, and E. Snchez-Sinencio, "A low-power configurable neural recording system for epileptic seizure detection," *IEEE Transactions on*

Biomedical Circuits and Systems, vol. 7, no. 4, pp. 499–512, August 2013. [Online]: http://dx.doi.org/10.1109/TBCAS.2012.2228857

[161] K. C. Chun, P. Jain, J. H. Lee, and C. H. Kim, "A 3t gain cell embedded dram utilizing preferential boosting for high density and low power on-die caches," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 6, pp. 1495–1505, June 2011. [Online]: http://dx.doi.org/10.1109/JSSC.2011.2128150

[162] R. E. Matick and S. E. Schuster, "Logic-based edram: Origins and rationale for use," *IBM Journal of Research AND Development*, vol. 49, no. 1, pp. 145–165, January 2005. [Online]: http://dx.doi.org/10.1147/rd.491.0145

[163] R. Nair, "Evolution of memory architecture," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1331–1345, August 2015. [Online]: http://dx.doi.org/10.1109/JPROC.2015.2435018

[164] C. E. Molnar and I. W. Jones, "Simple circuits that work for complicated reasons," in *IEEE Proceedings of the International Symposium on Advanced Research in Asynchronous Circuits and Systems*, 2000, pp. 138–149. [Online]: http://dx.doi.org/10.1109/ASYNC.2000.836995

[165] H. Schorr, "Computer-aided digital system design and analysis using a register transfer language," *IEEE Transactions on Electronic Computers*, vol. EC-13, no. 6, pp. 730–737, December 1964. [Online]: http://dx.doi.org/10.1109/PGEC.1964.263907

[166] D. Wang, A. Rajendiran, S. Ananthanarayanan, H. Patel, M. Tripunitara, and S. Garg, "Reliable computing with ultra-reduced instruction set coprocessors," *IEEE Micro*, vol. 34, no. 6, pp. 86–94, November 2014. [Online]: http://dx.doi.org/10.1109/MM.2013.130

[167] "Msp430g2x53 mixed signal microcontroller - data sheet," Texas Instruments Incorporated, Dallas, Texas, pp. 403–413, May 2013. [Online]: http://www.ti.com/lit/ds/symlink/msp430g2553.pdf

[168] F. L. Yuan, C. C. Wang, T. H. Yu, and D. Markovi, "A multi-granularity fpga with hierarchical interconnects for efficient and flexible mobile computing," *IEEE*

Journal of Solid-State Circuits, vol. 50, no. 1, pp. 137–149, January 2015. [Online]: http://dx.doi.org/10.1109/JSSC.2014.2372034

[169] B. Vigraham, J. Kuppambatti, and P. R. Kinget, "Switched-mode operational amplifiers and their application to continuous-time filters in nanoscale cmos," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 12, pp. 2758–2772, December 2014. [Online]: http://dx.doi.org/10.1109/JSSC.2014.2354641

[170] Y. Tsividis, "Event-driven data acquisition and continuous-time digital signal processing," in *IEEE Proceedings of the Custom Integrated Circuits Conference*, September 2010, pp. 1–8. [Online]: http://dx.doi.org/10.1109/CICC.2010.5617618

[171] I. Lee, D. Sylvester, and D. Blaauw, "A constant energy-per-cycle ring oscillator over a wide frequency range for wireless sensor nodes," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 3, pp. 697–711, March 2016. [Online]: http://dx.doi.org/10.1109/JSSC.2016.2517133

[172] B. Drost, M. Talegaonkar, and P. K. Hanumolu, "Analog filter design using ring oscillator integrators," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 12, pp. 3120–3129, December 2012. [Online]: http://dx.doi.org/10.1109/JSSC.2012.2225738

[173] V. Unnikrishnan and M. Vesterbacka, "Time-mode analog-to-digital conversion using standard cells," *IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications*, vol. 61, no. 12, pp. 3348–3357, December 2014. [Online]: http://dx.doi.org/10.1109/TCSI.2014.2340551

[174] K. Yang, D. Blaauw, and D. Sylvester, "An all-digital edge racing true random number generator robust against pvt variations," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 1022–1031, April 2016. [Online]: http://dx.doi.org/10.1109/JSSC.2016.2519383

[175] S. Chatterjee, Y. Tsividis, and P. Kinget, "0.5-v analog circuit techniques and their application in ota and filter design," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 12, pp. 2373–2387, December 2005. [Online]: http://dx.doi.org/10.1109/JSSC.2005.856280

[176] M. Alioto, "Understanding dc behavior of subthreshold cmos logic through closed-form analysis," *IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications*, vol. 57, no. 7, pp. 1597–1607, July 2010. [Online]: http://dx.doi.org/10.1109/TCSI.2009.2034233

[177] A. Hajimiri and T. Lee, "A general theory of phase noise in electrical oscillators," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 2, pp. 179–194, February 1998. [Online]: http://dx.doi.org/10.1109/4.658619

[178] A. Demir, A. Mehrotra, and J. Roychowdhury, "Phase noise in oscillators: a unifying theory and numerical methods for characterization," *IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications*, vol. 47, no. 5, pp. 655–674, May 2000. [Online]: http://dx.doi.org/10.1109/81.847872

[179] A. Hajimiri, S. Limotyrakis, and T. Lee, "Phase noise in multi-gigahertz cmos ring oscillators," in *IEEE Proceedings of the Custom Integrated Circuits Conference*, May 1998, pp. 49–52. [Online]: http://dx.doi.org/10.1109/CICC.1998.694905

[180] W. Jiang, V. Hokhikyan, H. Chandrakumar, V. Karkare, and D. Markovic, "A ±50mv linear-input-range vco-based neural-recording front-end with digital nonlinearity correction," in *IEEE Proceedings of the International Solid-State Circuits Conference*, January 2016, pp. 484–485. [Online]: http://dx.doi.org/10.1109/ISSCC.2016.7418118

[181] C. Weltin-Wu and Y. Tsividis, "An event-driven clockless level-crossing adc with signal-dependent adaptive resolution," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 9, pp. 2180–2190, September 2013. [Online]: http://dx.doi.org/10.1109/JSSC.2013.2262738

[182] H. Y. Yang and R. Sarpeshkar, "A bio-inspired ultra-energy-efficient analog-to-digital converter for biomedical applications," *IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications*, vol. 53, no. 11, pp. 2349–2356, November 2006. [Online]: http://dx.doi.org/10.1109/TCSI.2006.884463

[183] F. Corradi and G. Indiveri, "A neuromorphic event-based neural recording system for smart brain-machine-interfaces," *IEEE Transactions on Biomedical*

Circuits and Systems, vol. 9, no. 5, pp. 699–709, October 2015. [Online]: http://dx.doi.org/10.1109/TBCAS.2015.2479256

[184] K. A. Ng and Y. P. Xu, "A compact, low input capacitance neural recording amplifier," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 5, pp. 610–620, October 2013. [Online]: http://dx.doi.org/10.1109/TBCAS.2013.2280066

[185] J. Agustin and M. Lopez-Vallejo, "An in-depth analysis of ring oscillators: Exploiting their configurable duty-cycle," *IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications*, vol. 62, no. 10, pp. 2485–2494, October 2015. [Online]: http://dx.doi.org/10.1109/TCSI.2015.2476300

[186] K. Ng and Y. P. Xu, "A compact, low input capacitance neural recording amplifier," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 5, pp. 610–620, October 2013. [Online]: http://dx.doi.org/10.1109/TBCAS.2013.2280066

[187] M. Elia, L. B. Leene, and T. G. Constandinou, "Continuous-time micropower interface for neural recording applications," in *IEEE Proceedings of the International Symposium on Circuits and Systems*, May 2016.

[188] Y. W. Li, K. L. Shepard, and Y. P. Tsividis, "A continuous-time programmable digital fir filter," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 11, pp. 2512–2520, November 2006. [Online]: http://dx.doi.org/10.1109/JSSC.2006.883314

[189] B. Schell and Y. Tsividis, "A continuous-time adc/dsp/dac system with no clock and with activity-dependent power dissipation," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 11, pp. 2472–2481, November 2008. [Online]: http://dx.doi.org/10.1109/JSSC.2008.2005456

[190] S. Aouini, K. Chuai, and G. W. Roberts, "Anti-imaging time-mode filter design using a pll structure with transfer function dft," *IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications*, vol. 59, no. 1, pp. 66–79, January 2012. [Online]: http://dx.doi.org/10.1109/TCSI.2011.2161411

[191] X. Xing and G. G. E. Gielen, "A 42 fj/step-fom two-step vco-based delta-sigma adc in 40 nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 3, pp. 714–723, March 2015. [Online]: http://dx.doi.org/10.1109/JSSC.2015.2393814

[192] K. Reddy, S. Rao, R. Inti, B. Young, A. Elshazly, M. Talegaonkar, and P. K. Hanumolu, "A 16-mw 78-db sndr 10-mhz bw ct $\delta\sigma$ adc using residue-cancelling vco-based quantizer," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 12, pp. 2916–2927, December 2012. [Online]: http://dx.doi.org/10.1109/JSSC.2012.2218062

[193] J. Daniels, W. Dehaene, M. S. J. Steyaert, and A. Wiesbauer, "A/d conversion using asynchronous delta-sigma modulation and time-to-digital conversion," *IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications*, vol. 57, no. 9, pp. 2404–2412, September 2010. [Online]: http://dx.doi.org/10.1109/TCSI.2010.2043169

[194] F. M. Yaul and A. P. Chandrakasan, "A sub-μw 36nv/$\sqrt{Hz}$ chopper amplifier for sensors using a noise-efficient inverter-based 0.2v-supply input stage," in *IEEE Proceedings of the International Solid-State Circuits Conference*, January 2016, pp. 94–95. [Online]: http://dx.doi.org/10.1109/ISSCC.2016.7417923

[195] S. Patil, A. Ratiu, D. Morche, and Y. Tsividis, "A 3-10 fj/conv-step error-shaping alias-free continuous-time adc," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 908–918, April 2016. [Online]: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7433385&isnumber=7446371

[196] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1395–1408, July 2009. [Online]: http://dx.doi.org/10.1109/TIP.2009.2022459

[197] R. S. Schneider and H. C. Card, "Analog hardware implementation issues in deterministic boltzmann machines," *IEEE Transactions on Circuits and Systems—Part II: Analog and Digital Signal Processing*, vol. 45, no. 3, pp. 352–360, Mar 1998. [Online]: http://dx.doi.org/10.1109/82.664241

[198] J. Lu, S. Young, I. Arel, and J. Holleman, "A 1 tops/w analog deep machine-learning engine with floating-gate storage in 0.13μm cmos," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, January 2015. [Online]: http://dx.doi.org/10.1109/JSSC.2014.2356197

[199] M. T. Wolf and J. W. Burdick, "A bayesian clustering method for tracking neural signals over successive intervals," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 11, pp. 2649–2659, November 2009. [Online]: http://dx.doi.org/10.1109/TBME.2009.2027604

[200] D. Y. Barsakcioglu and T. G. Constandinou, "A 32-channel mcu-based feature extraction and classification for scalable on-node spike sorting," in *IEEE Proceedings of the International Symposium on Circuits and Systems*, May 2016.

[201] R. P. Feynman, "There's plenty of room at the bottom," *American Physical Society*, vol. 23, no. 5, pp. 22–36, February 1960. [Online]: http://www.zyvex.com/nanotech/feynman.html

[202] G. Leuba and L. J. Garey, "Comparison of neuronal and glial numerical density in primary and secondary visual cortex of man," *Experimental Brain Research*, vol. 77, no. 1, pp. 31–38, 1989. [Online]: http://dx.doi.org/10.1007/BF00250564

[203] I. Guideline, "Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 ghz)," *Health Physics*, vol. 74, no. 4, pp. 494–522, October 1998. [Online]: http://www.icnirp.org/cms/upload/publications/ICNIRPemfgdl.pdf

[204] L. B. Leene, S. Luan, and T. G. Constandinou, "A 890fj/bit uwb transmitter for soc integration in high bit-rate transcutaneous bio-implants," in *IEEE Proceedings of the International Symposium on Circuits and Systems*, May 2013, pp. 2271–2274. [Online]: http://dx.doi.org/10.1109/ISCAS.2013.6572330

[205] "Unconventional processing of signals for intelligent data exploitation (upside)," Defense Advanced Research Projects Agency, Arling-

ton, Texas, January 2016. [Online]: http://www.darpa.mil/program/unconventional-processing-of-signals-for-intelligent-data-exploitation