Title
Flexible and Energy-Efficient Circuits for Implantable Biomedical Systems

Author
Rozgic, Dejan

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Flexible and Energy-Efficient Circuits for Implantable Biomedical Systems

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical Engineering

by

Dejan Rozgić

2017

ABSTRACT OF THE DISSERTATION

Flexible and Energy-Efficient Circuits for Implantable Biomedical Systems

by

Dejan Rozgić
Doctor of Philosophy in Electrical Engineering
University of California, Los Angeles, 2017
Professor Dejan Marković, Chair

Biomedical implant-scale electronics have gained a lot of attention in recent years. Particularly, neuromodulation implants are an important tool in treating drug-resistant neurological conditions, while also improving our understanding of the brain. Although demands for adding more functionality to the implant are constantly increasing, their power consumption and size are usually limiting factors that determine longevity of the battery and dictates the overall throughput of brain data. Therefore, in order to gain more insight into brain dynamics while keeping device small, it is crucial to increase number of accessing channels and to improve the overall device efficiency.

Enabling better platform technologies that would greatly impact the field of neuroscience and enhance the quality of life of patients with neurological disorders is a difficult task. This work seeks to address some of the design challenges related to a variety of biomedical applications, while providing the power efficiency and flexibility needed for implantable devices.

First, a new self-powered, thermo-electric harvesting architecture is proposed and demonstrated. The miniaturized system, accompanied with efficient energy processing circuits was able to achieve a cold startup with a few 10's of mV of input voltage while achieving good

end-to-end efficiency. This design was further verified in real environment (in-vivo, rat) and showed a good trade-off between the form factor and extracted power.

Second, we demonstrated a ‘holy grail’ implant-scale neuromodulation interface with high linear input range that enables concurrent sensing and stimulation. Our 64-channel interface meets the requirements of human-quality implants at an unprecedented level of electronic miniaturization as compared to prior art. It offers major new clinical perspectives: it supports different power delivery options, always-on sensing for enhanced closed-loop therapy, multi-channel arbitrary stimulation waveforms with user-friendly programming, high-resolution neural interface for more precise target localization.

Finally, a new neural recording paradigm based on the fast calcium imaging is described. This technology can provide communication between the brain and the external world at the resolution of individual neurons. We propose a hardware friendly approach for analyzing 1000’s of neurons in a single pipeline and in real-time, while relaxing the memory and computational requirements. This method is capable of delivering two orders of magnitude higher brain coverage as compared to the state-of-the-art electrophysiological approach, leading to a high-resolution, high-data-rate neural interface.

The dissertation of Dejan Rozgić is approved.

Sudhakar Pamarti

Gregg Pottie

Nanthia Suthana

Dejan Marković, Committee Chair

University of California, Los Angeles

2017

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my gratitude to Prof. Dejan Marković for providing me the opportunity to conduct my research at UCLA. It was a privilege to work and collaborate with him over past 6 years. His meticulous day-to-day guidance, support and bright spirit made my work possible and exciting.

There are many others that I am also thankful not just for their advices and technical suggestions, but also for their company and support. Without them this journey would not be complete. I am particularly grateful to Vaibhav Karkare, Hariprasad Chandrakumar, Neha Sinha, Richard Dorrance, Yuta Toriyama, Sina Basir-Kazeruni, Wenlong Jiang, Qaiser Nehal, Rachid Mansour, Vahagn Hokhikyan, Chaitali Biswas, Cheng C. Wang, Fang-Li Yuan, Jiacheng Pan and Fengbo Ren for many discussions during my UCLA time.

Special gratitude goes to Prof. Ippei Akita who participated in one of the chip designs and helped with the valuable suggestions. I would like to thank Vladimir Petrović who continued the work on imaging project and gave important contributions.

This research was supported by a few DARPA projects and Broadcom Foundation. Work on DARPA SUBNETS/RAM programs reminded me on true meaning of the motto – no pain no gain.

Also, I want to thank the whole EE Department staff, whose help made my life at UCLA and in Los Angeles much easier – particularly Kyle Jung and Jamie Khang, to whom I am eternally grateful.

And at the end, I must deeply thank to my parents who supported me in every step with love and encouragement. They were the source of energy and strength during hard times and desperate moments – I dedicate this thesis to them.

VITA

2009	B.E., Electrical Engineering – Electronics University of Belgrade, Serbia
2011	M. Eng. Sc., Electrical Engineering – Electronics University of Belgrade, Serbia
2014	Teaching Assistant Electrical Engineering Department University of California, Los Angeles
2016	Broadcom Fellowship
2011-2017	Graduate Student Research Assistant Electrical Engineering Department University of California, Los Angeles

PUBLICATIONS

D. Rozgić and D. Marković, "Micro-TEG Voltage Supplies for Spin Torque Oscillators", *IEEE TRANSACTIONS ON ELECTRON DEVICES*, VOL. 60, NO. 9, Sep. 2013.

D. Rozgić and D. Marković, "A 0.78mW/cm^2 Autonomous Thermoelctric Energy-Harvesting Platform for Biomedical Sensors", *Symposium on VLSI Circuits, C278 - C279, (VLSI'15).*

D. Rozgić and D. Marković, "A Miniaturized 0.78mW/cm^2 Autonomous Thermoelectric Energy-Harvesting Platform for Biomedical Sensors ", *Transactions on Biomedical Circuits and Systems,* pp. 773-783, VOL. 11, NO. 4, Aug. 2017.

V. Karkare, H. Chandrakumar, D. Rozgić, and D. Marković, "Robust, Reconfigurable, and Power-Efficient Biosignal Recording Systems," *IEEE Custom Integrated Circuits Conference*, Sep. 2014.

J. Pan, Asad A. Abidi, D. Rozgić, H. Chandrakumar and D. Marković, “An inductively-coupled wireless power-transfer system that is immune to distance and load variations”, in *IEEE ISSCC Dig. Tech. Papers 2017*, pp. 382-383, Feb. 2017.

CHAPTER 1

Thesis Overview

This dissertation is concerned with energy efficient and flexible circuits for implant-scale biomedical systems and consists of 3 different parts. Chapter 2 introduces and analyze a new architecture for miniaturize thermo-electric harvesting. Chapter 3 presents the work on implantable multi-channel neuromodulation platform that can support various power delivery options, while chapter 4 explains the path towards dedicated hardware for neural recording paradigm based on the fast brain imaging. A more detail explanation of every chapter is offered bellow. Each chapter presents a separated topic and therefore results and conclusions are derived at the end of each chapter.

Chapter 2: A Fully Autonomous TE Energy-Harvesting Platform for Biomedical Sensors

In order to use thermoelectric energy harvesters (TEHs) as a truly *autonomous energy* source for size-limited sensing applications, it is essential to improve the power conversion efficiency and energy density. This chapter presents a thin-film, array-based TEH with a surface area of 0.83cm^2. The TEH autonomously supplies a power management IC fabricated in a 65nm CMOS technology. The IC utilizes a single-inductor topology with integrated analog maximum power point tracking (MPPT), resulting in a 68% peak end-to-end efficiency (92% converter efficiency) and less than 20ms MPP tracking time. In an in-vivo test, a 645µW regulated output power (effective 3.5K of temperature gradient) was harvested from a rat implanted with our TEH, demonstrating true energy independence in a real environment while showing a 7.9x improvement in regulated power density compared to the state-of-the-art. The system showed autonomous operation down to 65mV of TEH input.

Chapter 3: A Miniaturized 64-Channel Neuromodulation Platform for Simultaneous Stimulation and Sensing

Brain machine interfaces (BMI) have the opportunity to advance our understanding of the brain, restore motor function, and improve the quality of life to patients with neurological conditions. For example, deep-brain stimulation (DBS) can provide symptomatic relief for neurological patients by emitting electrical pulses. For human use, a neuromodulation (NM) implant should be minimally invasive, with high-precision interface that can record neural activity in presence of stimulation.

In chapter 3 a first full duplex implant-scale NM unit, with extreme miniaturization packing a 32/64-channel interface in $0.135cm^3/0.22cm^3$ is demonstrated while meeting human-grade implant requirements. As an integrative part of the platform, integrated and flexible power management unit is shown. Power-management circuits in the NM should have high power conversion efficiency (PCE) to operate with smaller received power, but also should show a high level of integration. Circuits techniques that led to improvements in the PCE for wireless/wired implantable devices are analyzed. Specifically, neural stimulating systems should perform with high stimulation efficiency with a minimum amount of energy while ensuring charge-balanced stimulation, providing advantages such as a wide range of stimulus currents, a longer battery life, reconfigurability, etc. are demonstrated.

Chapter 4: Hardware accelerator for simultaneous, real-time neuronal recording of large ensembles for brain imaging

In this chapter, we report an alternative approach for neuronal recording. With recent advances in fluorescent imaging sensors, and their improved speed (100's of fps), we can simultaneously track and record data from a large number of neurons (100~10 000). However, the image sensors generate a large amount of data (0.1GB/s~1GB/s), while its real-time hardware implementation is bounded by large memory and heavy computation requirements, since the system performs frame-level processing. There are a few obstacles that prevent a wider use of this technology: i) The camera receives the frames with motion jitter ii) The position and shapes of neurons are unknown iii) Valuable information (spiking signals) have to be extracted from the raw fluorescence traces which are contaminated with high baseline noise and convolved with other unwanted content. So far, all data processing has been performed offline. In this chapter, we proposed a hardware approach that solves all these issues in a single pipeline and in real-time. Motion Correction and Blind Neuron Detection are realized by employing modified computer vision algorithms such as Maximally Stable Extremal Regions and Template Matching. By exploiting the sparse nature of neurons and spiking signals both in the spatial and time domains, specialized dedicated units that map the Sparse Approximation algorithm into hardware, are able to extract spikes and achieve 100x data reduction. The envisioned system consists of a fast photonic neural transducer and a smart DSP unit for low-power signal processing capable of spatio-temporal localization and tracking of single units in real-time. Every frame from the video is processed independently and does not require loading the whole frame into memory, thus reducing memory requirements.

CHAPTER 2

A Fully Autonomous TE Energy-Harvesting Platform for Biomedical Sensors

2.1 Introduction

Harvesting thermal energy and its usage as a potential source for miniaturized electronic systems, has attracted a lot of attention in recent years. Many studies showed that extracting thermal energy can potentially supply hundreds of microwatts of useful power. Even though the power levels are adequate, such harvesters produce very low voltage levels, 10's-100's of mV, which are insufficient to power CMOS electronics. The focus of the research community [1]-[9] has been on improving the harvester's efficiency and low-power circuit design. Their main goal was to achieve high efficiency of processing circuits and to reduce the number of off-chip components, so that the system is optimized for size, power and cost. However, prior work lacks power density, with state-of-the-art power density below 200μW/cm^2. Equivalently, a 1mW of power would require a 5-6cm^2 surface area, which is unacceptable for minimally-invasive implantable devices. The TEH conversion efficiency and power density need to be improved in order to have a miniaturized autonomous energy source. Also, it is necessary to miniaturize the thermoelectric transducer and integrate it with the power management IC. Previous designs lack a system-level design and optimization approach, which is offered here.

Four major obstacles prevent autonomous thermal energy usage, as described below. First, since the output voltage from the transducer (that is responsible for thermal-to-electrical energy conversion) can be very low, harvesting systems should have a cold startup ability, i.e. the circuit should trigger (startup) its operation without any stored energy. A few prior designs have demonstrated this ability for thermal harvesters, [1]-[7]. Their startup units require off-chip components, which makes them unattractive for miniaturization. Some designs require a battery

[1] or the output storage element to be charged to certain voltage [2], which can be used as the initial trigger. No prior work has reported an *autonomous-integrated* startup, *from a fully-discharged device*. A mechanical off-chip switch [3] is used in the boost converter design that is able to harvest energy even from a 30mV voltage input. The use of a mechanical (motion based) switch is not autonomous and hence has limited utility. Further, it achieves peak efficiency at lower voltages; the range of high efficiency (above 50%) is quite narrow, with the efficiency dropping at higher voltages. In order to reduce the startup voltage, the authors in [6], showed post-fabricated trimmed oscillator operation down to 90mV. Transformer based cold startup was demonstrated in [4]; however, this method requires a large volume to accommodate the transformer, limiting the practical usage of the system and affecting the maximum available efficiency. To facilitate startup, the authors in [7] have recently proposed a multiple-ambient-sources-harvesting approach, where the system would start operation by using one of the energy sources and then continue to harvest higher power from another source. A pre-calibration scheme and explicit control over inductive peaking current were employed to improve efficiency; however, this approach requires RF-assisted startup, which does not qualify as a fully autonomous TE self-start. Second, maximum power point tracking (MPPT) scheme has to be implemented to match the impedance of the harvester's circuit with that of the heat source, in order to maximize the available power. It is important to note that most of prior research focused on increasing converter efficiency [1]-[9], demonstrating sub-100mV operation limited to a controlled lab environment. References [2]-[4] are rare exceptions that report end-to-end efficiencies. Still, their PCB + TEG systems occupy a large area and harvesting in natural environments (where temperature differences are <3K) would be very difficult and unreliable. Third, providing a stable thermal gradient, with minimal heat leakage in a small footprint, is very challenging and it further hindered previous attempts at truly

autonomous energy harvesting. Lastly, the number of off-chip parts has to be reduced for better miniaturization.

Our work addresses the aforementioned challenges related to thermal harvesting from low input voltages, but takes into account miniaturization demands for biomedical implants. The overall system is presented in Section 2.2, features an on-chip startup CMOS circuit that is assembled with TE platform capable of extracting autonomous power. Section 2.3 describes our power management solution that makes power extraction efficient and agnostic to the harvester environment. With the fast closed-loop control techniques, described in Section 2.4, the low-power PM circuitry achieves high efficiency across a wide range of load currents and PVT. The high efficiency is due to accurate detection of inductor current zero-crossing and low-power comparator design. A miniaturized, custom TEH platform is described in Section 2.5. Together with a 65nm CMOS chip, the platform was tested in-vivo on a rat. Measurement results, discussed in Section 2.6, are the new state-of-the-art in autonomous thermoelectric harvesting. Our system achieves the highest level of integration, including both the PCB (circuit innovation) and TEH (materials, physics, mechanical, assembly, and surgery).

2.2 System Architecture

Fig. 2.1 shows the proposed system-level single-inductor hybrid-type architecture. A thermo-electric harvester, to a first-order approximation, is depicted as a DC-voltage source, V_{TEH}, with its internal resistance, R_{TEH}. The system comprises of an analog-domain MPPT circuit, a cold startup block based on inductive-load ring oscillator (ILRO) and mode controller. Charge transfer is done through the main boost branch comprised of an off-chip inductor, active diode AD, main

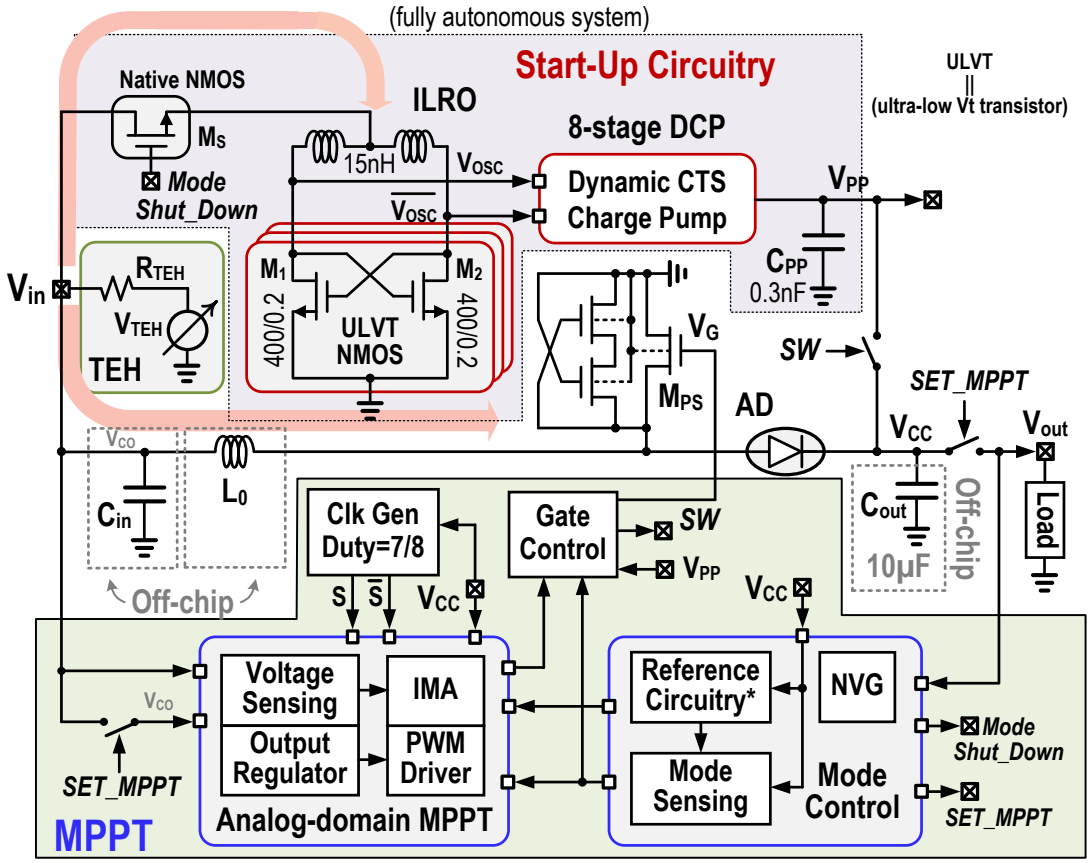


Fig. 2.1: Proposed thermoelctric harvesting architecture.

boost switch M_{PS} and a storage element C_{OUT}. Active diode implementation is crucial for an efficient and low-leakage power delivery. The details of the boost operation will be discussed in Section III. During the self-startup mode, an ILRO, a charge-transfer-switch (CTS) charge pump, and native NMOS mode switch with negative V_T, are employed. Once the startup block charges the output voltage (V_{CC}) to an intermediate level (0.8V), a negative voltage generator (NVG) shuts down the startup block and the boost converter transitions to the MPPT Mode. To extract maximum power, the MPPT block is enabled; the active control of M_{PS} periodically turns the switch off whenever the inductor current reaches zero in the falling charge-transfer operation. The MPPT operation is detailed in Section 2.4. In the MPPT mode, a dedicated output regulation unit

is used for output voltage control. Since the main boost switch (M_{PS}) carries 10's of mA of current, active body control is employed to prevent reverse current flow and to mitigate the leakage current.

2.3 Power Management and Timing Control

2.3.1 Inductive Load Ring Oscillator (ILRO)

As mentioned before, starting up CMOS circuits with sub-100mV input presents a difficult task in cold startup circuit design. Below 100mV, active circuitry (transistors) operates in weak inversion (WI). In order to decrease the startup voltage, circuit designers usually connect the harvester output directly to some kind of an oscillator which acts as the system activation unit. Such oscillators demand either bulky off-chip components or their transistors require some post-fabrication tuning. Motivated by the work in [10], we leverage the fact that the inductive-load ring oscillator (ILRO) architecture can push the oscillation amplitude above the supply rails, allowing it to be triggered with very low input voltages, Fig. 2.2a. We employed a 2-stage ILRO due to its simplicity and the good trade-off between the performance (low-voltage startup) and active chip area. Using the EKV Model [11], and the small-signal equivalent circuit for ILRO, Fig. 2.2b, it can be shown that the minimum startup voltage for a 2-stage ILRO is half of the startup voltage for a classical CMOS inverter-based oscillator. The lower bound for the classical oscillator startup that operates in subthreshold region is described in [12] and given by:

$$V_{DD}(\min) = 2\phi_t \ln(1 + n), \tag{2.1}$$

where $\phi_t = kT/q$ is the thermal voltage and n represents the subthreshold slope.

Analog high-performance (native-depleted) transistors have threshold voltages around zero, resulting in a high current drive and high output gain. Thus, their ILRO implementation can produce high output frequencies while supplied with low input voltages. The upper bound for the oscillator frequency is dictated by the transistor's unity-gain frequency f_T and by the load attached

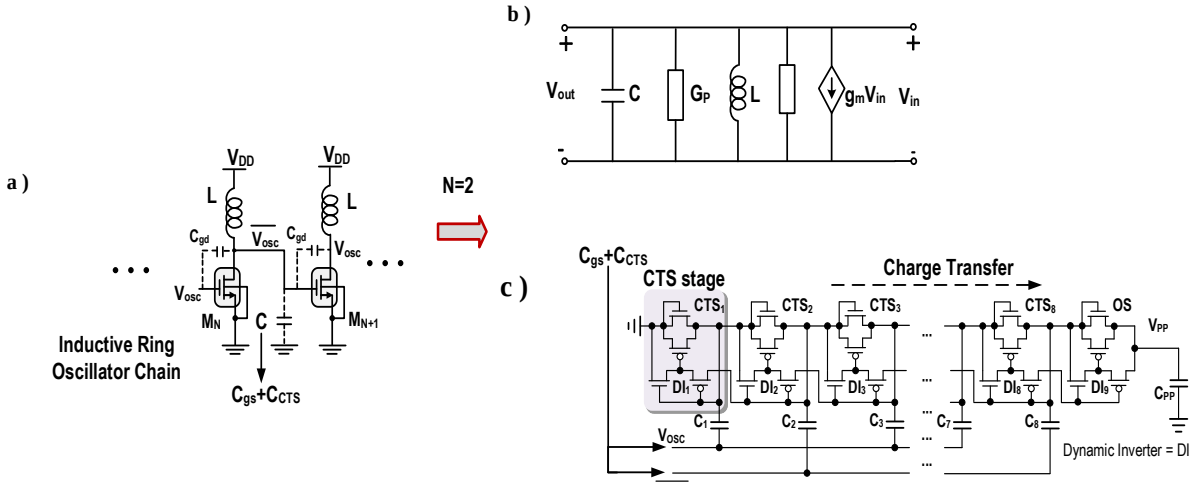


Fig. 2.2: (a) Inductive-load ring oscillator chain and its (b) small-signal circuit equivalent. (c) Schematic of 8-stage CTS charge pump.

to the output. In a 65nm technology, f_T for the native transistors is in the 100MHz– 1GHz range, for V_{GS} of several 10's of mV (10mV-40mV). To derive the relationship between the minimum startup voltage for the ILRO and the transistor's geometry, we refer to Fig. 2.2b, [10]-[11]. The single-stage transfer characteristic implies:

$$\frac{V_o}{V_{in}} = -\frac{g_m}{g_{md}+G_L}\frac{1}{1-j\tan\varphi}. \tag{2.2}$$

The g_m, g_{md} and g_{ms} represent the gate, drain and source transconductances of the transistor. The G_L denotes the inductor losses.

The phase shift φ for the single stage is assumed to be π without lost of generality. Assuming that $Q = \frac{\omega C}{G_L}$ is the quality factor, the Barkhausen's criterion for oscillation startup is:

$$\frac{g_m - g_{md}}{C} - \frac{\omega}{Q} > 0 \tag{2.3}$$

The capacitance C in (2.3), which represents the ILRO load, is given as $C = C_{CTS} + C_{DS}$, where C_{CTS} is the input capacitance of the CTS charge pump and C_{DS} is the equivalent drain-source capacitance of the ILRO transistor.

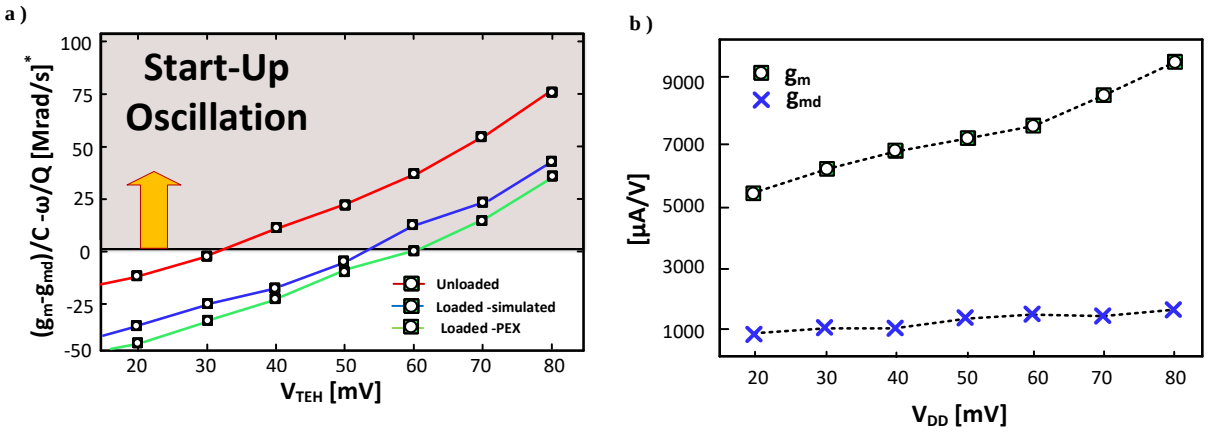


Fig. 2.3: (a) Simulated startup condition for a different V_{TEH}; (b) Simulated gate and drain transconductances of the ULVT transistors (V_S=V_B=0).

The EKV model of the transistor in WI [11], implies the relationship between g_{md}, g_{ms}, g_m and the drain-source voltage is given by (2.4)-(2.5):

$$ng_m = g_{ms} - g_{md}, \tag{2.4}$$

$$\frac{g_{ms}}{g_{md}} = e^{\frac{Vds}{\varphi_t}}. \tag{2.5}$$

From (2.2), (2.4) and (2.5), the minimum supply voltage needed for ILRO startup is given by:

$$V_{DD}(\min) = V_{DS}(\min) = \phi_t \ln\left[1 + n\left(1 + \frac{G_L}{g_{md}}\right)\right]. \tag{2.6}$$

In the ideal case, the minimum supply voltage for oscillations to occur is $\phi_t \ln(1+n)$, which is exactly one-half of the minimum supply voltage needed to startup an inverter-based ring oscillator. Fig. 2.3a shows the simulated startup condition (for sustained oscillations) in terms of the minimum harvester voltage (V_{TEH}) for the oscillator using a 65 nm technology.

In order to get more insight into the properties of the native MOS transistor, Fig. 2.3b plots g_m

and g_{md} transconductances versus drain-source voltage V_{DS} that is swept from 0 to 80mV.

The ILRO is built with inductors with L ≈ 15nH and transistors with W/L=2400μm/0.2μm. The inductor was chosen such that its G_L value was as low as possible within the expected frequency of operation (300 to 500 MHz). Additional headroom allocated for PVT variation and layout parasitics marginally increased the startup voltage and contributed to the drop in oscillation frequency from 350MHz (designed) to 300MHz (measured).

The minimum voltage needed to start the oscillation was measured to be 65 mV, closely matching the value of 60 mV obtained from PEX simulations. The efficiency of ILRO is 15% for the minimum startup voltage (simulations showed $I_{DC\text{-}ILRO} \approx 0.13$mA @ V_{IN}=60mV, $I_{DC\text{-}ILRO} \approx$ 0.39mA @ V_{IN}=100mV).

2.3.2 Charge-Transfer-Switch-Charge Pump

The design goal for the CTS charge pump (CP) is to achieve sufficient output DC voltage and to be able to supply the control circuitry while minimizing the equivalent input capacitance. The schematic of the CTS charge pump in the proposed startup circuit is shown in Fig. 2.2c. The dynamic CTS CP uses the backward and forward control for NMOS and PMOS pass transistors respectively. This scheme employs the high voltages generated in the succeeding stage to control the NMOS transistor and low voltages generated in preceding stage for the PMOS transistor. The body effect in the last stage is successfully eliminated by PMOS CTS. The CTS CP shifts the charge stage-by-stage synchronously with negligible voltage drop. The pass transistors in the charge pump are completely turned off by V_{OSC} and completely turned on by higher voltages from the following stage. This leads to higher efficiency since the reverse current flow is significantly reduced. The CTS also uses low-V_T (≈0V) transistors and their aspect ratio increases in consecutive stages in order to keep the output impedance low. Our 8-stage CTS charge pump

shows 41% and 71% simulated power efficiency for V_{OSC}=100mV_{pk-pk} and V_{OSC}=500mV_{pk-pk}, respectively.

2.3.3 Active Diode with Low-Voltage Drop

Since the current from the harvester flows through the diode, in all operating modes, its realization should be energy efficient and yet it should show sufficient performance. Using off-chip Schottky diodes would prevent the reverse current leakage, but their threshold voltage is bounded to 0.2V-0.3V and they would occupy extra PCB space. For area-limited applications, active diode implementation is a must, provided that static power is minimized. Previous solutions on active diodes [13]-[14] show fast circuits consisting of comparators, which with help of feedback actively control the diode (NMOS/PMOS) switch. The advantage of these approaches is that the diode will achieve almost zero current switching and the circuit will compensate any delay in the switch response. On the other hand, these implementations consume a lot of power in comparators and auxiliary circuitry; for biomedical applications, this approach does not offer a good trade-off between design needs and power consumption. Our active diode schematic is shown in Fig. 2.4. The circuit is designed so that the bias current I_B presents a very small fraction of the forward current I_F; this ratio is dictated by transistor geometry. During the conduction period (V_S> V_{out}+V_{DROP}), the bias current exists and the bulk of M_D is tied to the highest available potential to mitigate the current leakage though M_D. During the blocking period, both bias and conduction currents drop down to zero due to the positive feedback in the circuit. Transistors M_{R1} and M_{R2} act as large resistors and they additionally limit the static current consumption. If the forward currents are low, transistor M_D operates in subthreshold regime and with proper sizing of M_D, M_1 and M_2, the voltage drop V_{DROP} can be very low, [15]. If the forward currents are high (10's of mA), the

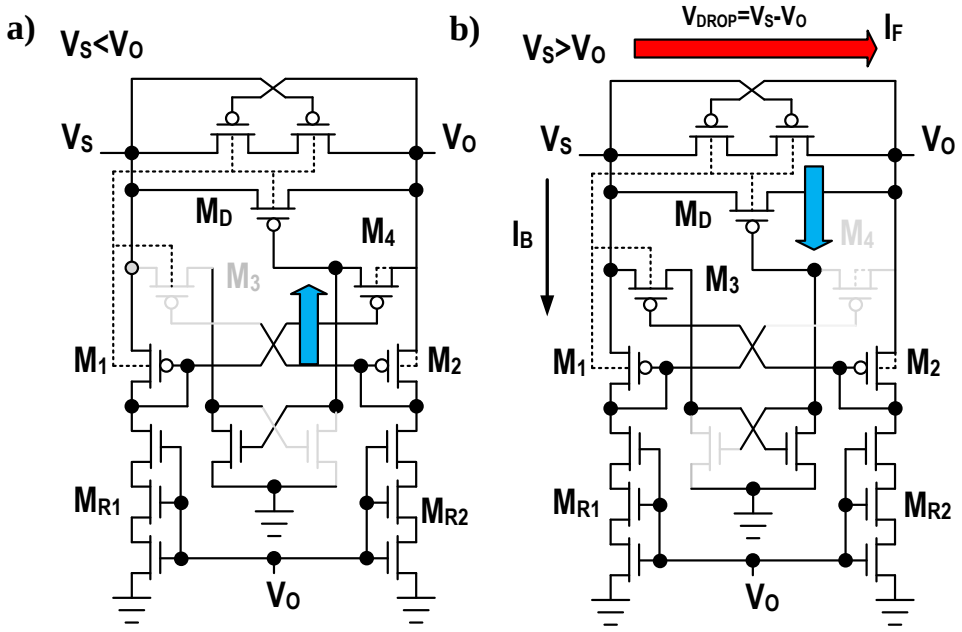


Fig. 2.4: Active diode (AD) during (a) OFF and (b) ON states.

voltage drop (normally 10's of mV) is inversely proportional to the forward current. This simple design approach can reach the performance similar to the ideal Schottky diode, but with lower power. The active diode can carry 10's of mA of forward current while providing sufficiently fast signal switching.

2.3.4 Startup Mode and Relevant Waveforms

As previously explained, the inductive-load ring oscillator was employed due to its low voltage startup and compact area. Bulky off-chip inductors would additionally increase the startup time and voltage, and also increase the load at the ILRO output. At t=0, since there is no accumulated energy in the system, the current flows from the harvester through the startup path. The native-NMOS transistor with large negative threshold voltage V_T is used as the mode switch; initially (t=0) it is ON. Cross-coupled transistors $M_{1\text{-}2}$ in the ILRO are realized with high-performance (HP) ultralow-V_T analog transistors. These transistors have enough voltage gain and high current drive

even at low voltage supplies; our simulations showed Q_{ILRO}=10.5 at 300MHz and 60mV startup voltage.

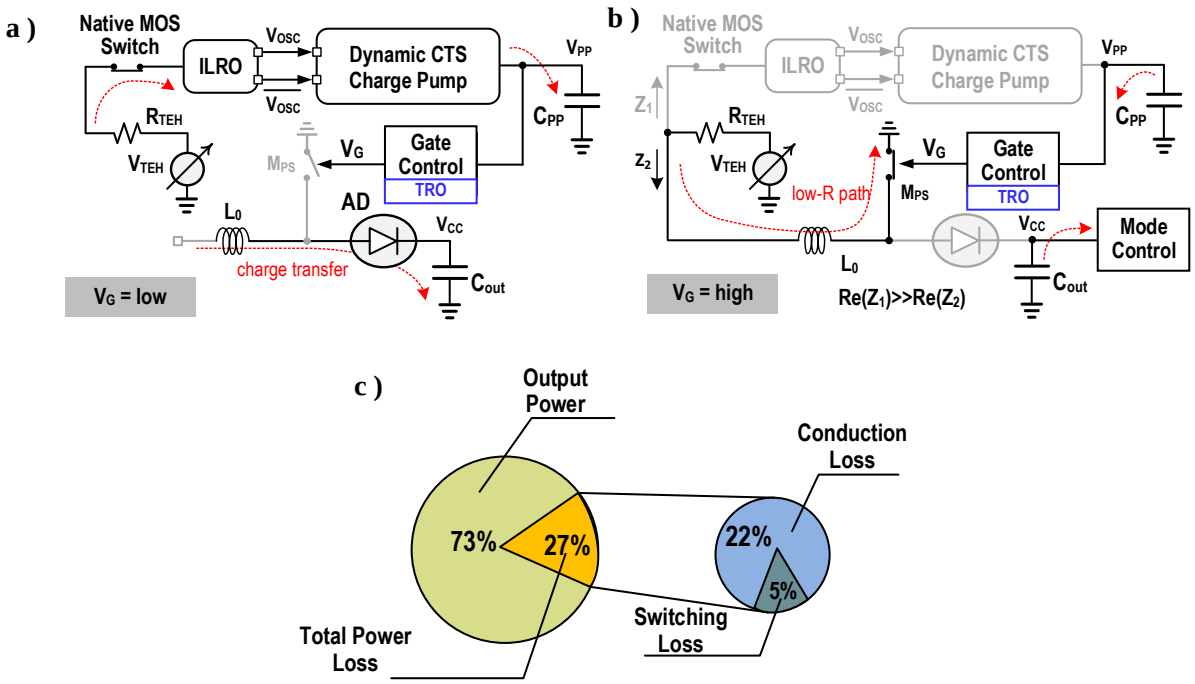


Fig. 2.5: Active circuitry during the startup mode in (a) discharging and (b) charging phases c) Simulated power distribution after design optimization.

For a main boost switch M_{PS}, a low-V_T transistor was used, as it lowers the leakage current as compared to the native one, when the switch is OFF. As the loading of this converter is a simple digital logic, the load current (I_{LOAD}) is relatively low. Discontinuous conduction mode (DCM) is the preferable operation mode of the boost converter if the output voltage of the harvester is low. In DCM – the boost converter can still have a large boosting ratio even with a light load current and low input voltage $(1 + \frac{V_{in}D^2}{2L_0 f_{SW} I_{LOAD}})$, [16]. The duty cycle is set to be ¾. The efficiency of the boost converter, during this mode, is inherently bounded by losses during conduction (P_α) due to resistance on the current path, and losses due to switching (P_{SW}). The effective resistance on the current path is given by $R_\alpha = R_{SW} + R_L + R_{TEH}$, where R_L is the series resistance of the inductor and R_{SW} is the on-resistance of the main boost switch.

The power (P_R) accumulated in the inductor during the current rising can be approximated with

$$P_R = \frac{1}{2}\frac{L_0 i_{peak}^2}{t_R}, \quad (2.7),$$

where i_{peak} is the inductor current at the end of the rising period t_R. The current i_R can be expressed as

$$i_{peak} = \frac{V_{TEH}}{R_\alpha}(1 - e^{-\frac{R_\alpha}{L_0}t_R}). \quad (2.8)$$

Substituting (2.8) in (2.7) yields

$$P_R = \frac{L_0 {V_{TEH}}^2}{{R_\alpha}^2}\frac{(1-e^{-\frac{R_\alpha}{L_0}t_R})^2}{2t_R}. \quad (2.9)$$

The falling time t_F in DCM during the current drop can be approximated with

$$t_F = \frac{V_{TEH}}{R_\alpha}\frac{L_0}{V_{CC}+V_{DROP}} \approx \frac{V_{TEH}}{R_\alpha}\frac{L_0}{V_{CC}}, \quad (2.10)$$

if we assume $V_{DROP} \ll V_{CC}$. In order to maximize the average power P_R delivered during one period, main switch ON-resistance and the inductor ESR have to satisfy $R_{SW} + R_L \ll R_{TEH}$. Larger switch will mitigate the ON resistance and result in a higher dynamic power dissipation. Preservation of energy during the startup mode gives us relation between the energy stored in the inductor when M_{PS} is ON and the energy dissipated on the diode during its forward bias (t_F) and on the load during the entire period:

$$\frac{1}{2}L_0 i_{peak}^2 \geq E_{ESR}|_{V_{CC}=0.8V} + E_{DIODE}|_{V_{CC}=0.8V} + V_{CC}I_{LOAD}T_S|_{V_{CC}=0.8V}. \quad (2.11)$$

With maximizing power P_R (i.e. $\frac{dP_R}{dt_R} = 0$) and given (2.7)-(2.11) we can determine the lower bound of the inductance L_0 and desired DCM period T_S. We employed T_S = 80μs and $L_0 = 150\mu H$

(footprint: 6mm x 5.6mm). Note that L_0 does not have a linear impact on the footprint; the footprint is more sensitive to R_L than to L_0. Fig. 2.5c shows the power distribution during one converter cycle after parameter optimization. Out of 27% total power loss, 22% is in the conduction loss. Simulation results indicate an available load current close to 0.4µA, which is sufficient to drive the auxiliary control circuits.

Fig. 2.5 shows the startup mode during both phases. After voltage V_{PP} passes 0.3V, the thyristor-based oscillator (TRO) will start driving the buffer in the gate-control (GC) block which will conduct charging/discharging of the power switch M_{PS}. In the charging phase, the startup block is turned OFF since the current flows through the path of lower resistance (Z_2 in this case). In the discharging phase (V_G is low), the energy stored in L_0 is transferred through AD to the output capacitance C_{PP}. Concurrently, the current from the harvester closes the loop through ILRO again; starts oscillations and the charge pump additionally recharges the auxiliary capacitance C_{PP}. Control circuitry that is biased from C_{PP} consumes less than 100nA over one period. The startup block is turning on periodically while the output voltage V_{CC} keeps increasing.

The true single-phase latch (TPSC) keeps the MPPT controller in idle mode during the startup phase. Post-layout simulated waveforms during startup mode are shown in Fig. 2.6. The MPPT controller becomes active after the output voltage is boosted to 0.8V.

2.4 MPPT Mode and Timing Diagrams

The MPPT mode requires a low-power comparator scheme that can achieve adequately fast state transitioning. In order to satisfy these requirements, a two-stage OTA-based comparator is employed, as shown in Fig. 2.7a.

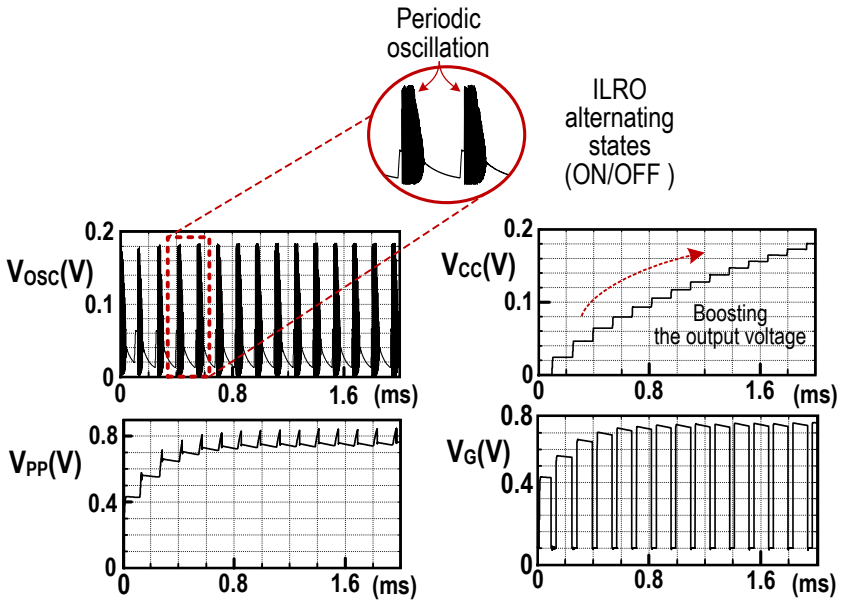


Fig. 2.6: PEX-simulated startup for a different V_{TEH}.

The comparator is designed so that the current consumption is less than 470nA for a supply voltage of 1.8V while achieving switching frequency of 0.1MHz.

The maximum power point (MPP) of a thermoelectric harvester is attained when the input voltage is at one half of the open-circuit voltage. Because the MPP changes the value with the environment conditions such as the pressure, temperature fluctuations and also it varies with load requirements, a control circuit for MPP tracking is employed to sense half of the open-circuit voltage and to adaptively follow the MPP.

The maximum output power, P_{max}, can be expressed as:

$$P_{max} = \frac{V_{TEH}{}^2}{4R_{TEH}}, \text{ with } V_{in} = \frac{V_{TEH}}{2}, \tag{2.12}$$

The MPPT loop and all active circuits during this mode are shown in Fig. 2.7b. After the activation of the MPPT controller (SET_MPPT = high), the negative voltage generator produces −0.4V at its output to turn-off the native transistor (the mode switch) and the startup block. A clock

generator outputs complementary signals S and $\bar{S}$ with a 7/8 duty ratio and a 600µs period. When S = low, the open-circuit voltage V_{TEH} is sampled, while the PWM and GC blocks keep

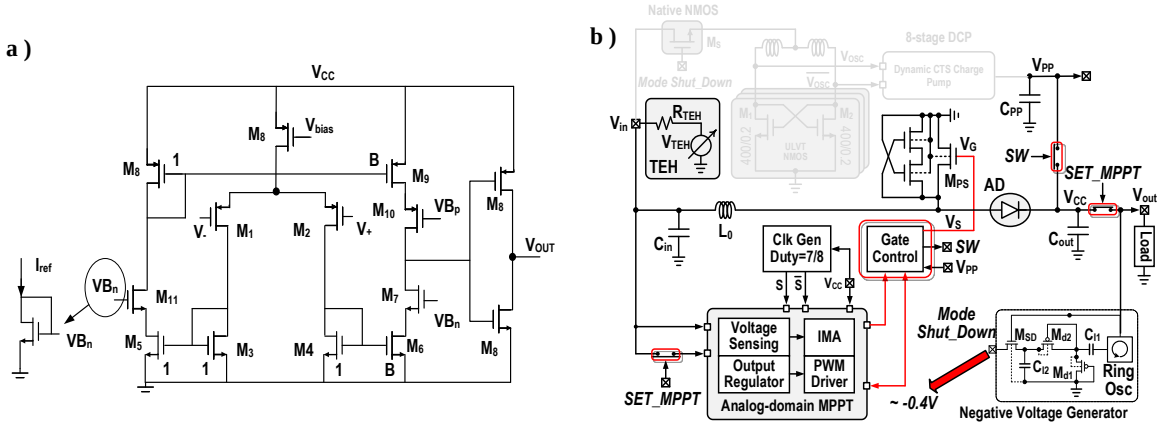


Fig. 2.7: a) Two-stage comparator design. b) Activation of MPPT block and shut-down of low-voltage starter.

CMP_3 dynamically matches the input voltage V_{in} and $V_{TEH}/2$, by accommodating the pulse width of the gate voltage V_G through the feedback loop formed by CMP_3, inductor, the PWM and GC blocks. For precise control, it is important to minimize offset of CMP_3 comparator. The energy stored in inductor is transferred to C_{OUT} during the $\bar{S}$ phase, Fig. 2.8a.

Fig. 2.8b shows the active circuitry in the feedback loop when S is high. If the input voltage is higher than $V_{TEH}/2$, CMP_3 will turn-on the main boost switch M_{PS} through TG_1, TPSC and GC blocks. As the current through inductor keeps increasing, the input voltage is decreasing. After

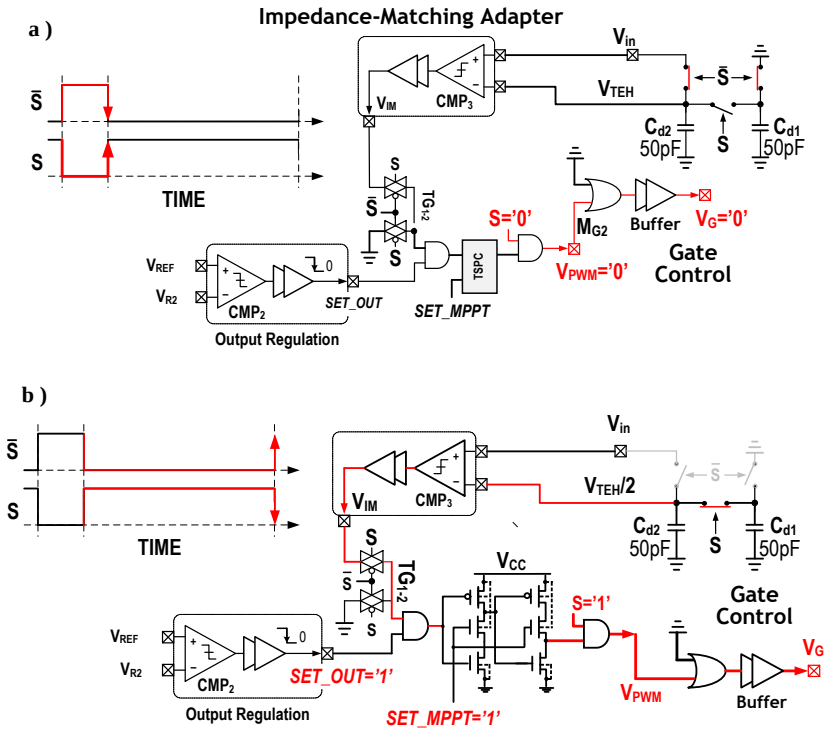


Fig. 2.8: MPPT operation: a) open-circuit condition, b) feedback loop during the PWM phase.

signals V_{PWM} and V_G at zero, Fig 2.8a. During this period, the active diode prevents the reverse current flow from output capacitor C_{OUT}. When S = high, the capacitive divider gives $V_{TEH}/2$ by sharing the charge between the capacitors C_{d1} and C_{d2}.

When the input voltage reaches $V_{TEH}/2$, the MPPT controller turns off the main switch M_{PS}. Potential V_S becomes higher than the output voltage, and the energy stored in the inductor is transferred into C_{OUT} via AD. Then the current through the inductor starts to decrease, while V_{in} goes further below V_{TEH}. After a full period, the input voltage will go up while the inductor current will go down; this sequence starts repeating periodically, after V_{in} becomes higher than $V_{TEH}/2$.

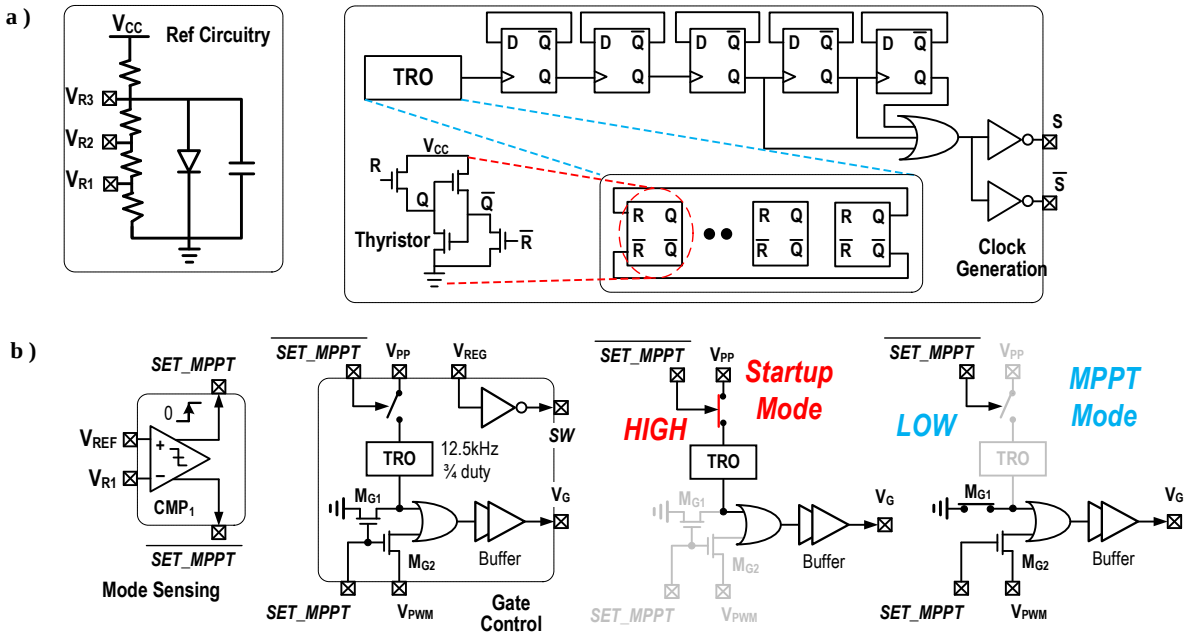


Fig. 2.9: (a) Reference and clock generation, (b) Gate control block during startup (left) and MPPT (right) modes.

Due to the fast voltage sensing (C_{d1}=C_{d2}=50pF) and comparator fast transition, CMP_3 enables the system to find MPP very quickly; less than 20ms is needed for complete MPP regulation. The fast feedback-loop response results in a small voltage ripple at V_{in} with a small (2nF) input capacitance C_{in} and short settling time (3µs). The amount of ripple is dependent on the input

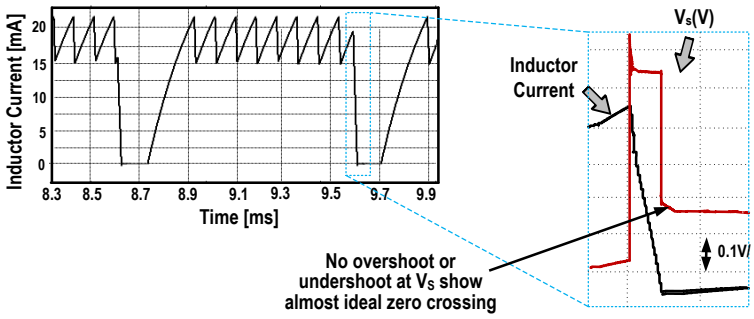


Fig. 2.10: Inductor switching waveform for V_S at V_{TEH}=180mV showing almost perfect zero switching.

capacitance (C_{in}). The regulated output of the TEH is kept at its MPP with negligible voltage ripple. Any voltage fluctuations at the harvester side (up to 50Hz) can be captured by the feedback loop

in MPPT controller. The input signal periodically moves between V_{TEH} and $V_{TEH}/2$ confirming the correct impedance matching. Until the output voltage doesn't reach 1.8V, which is the target value (to power neural recording interfaces), MPPT mode is active. Comparator CMP_2 keeps the output voltage at the desired value by dynamically alternating the control signal SET_OUT. Auxiliary circuitry and MPPT controller consume less than 2.9µA during active mode and 0.07µA during idle mode, which directly translates into high converter efficiency. Circuit details of auxiliary blocks used in MPPT block and GC are shown in Fig. 2.9. The clock generation block uses thyristor-based cells, while the reference on the chip employs simple, low power diode-based circuitry.

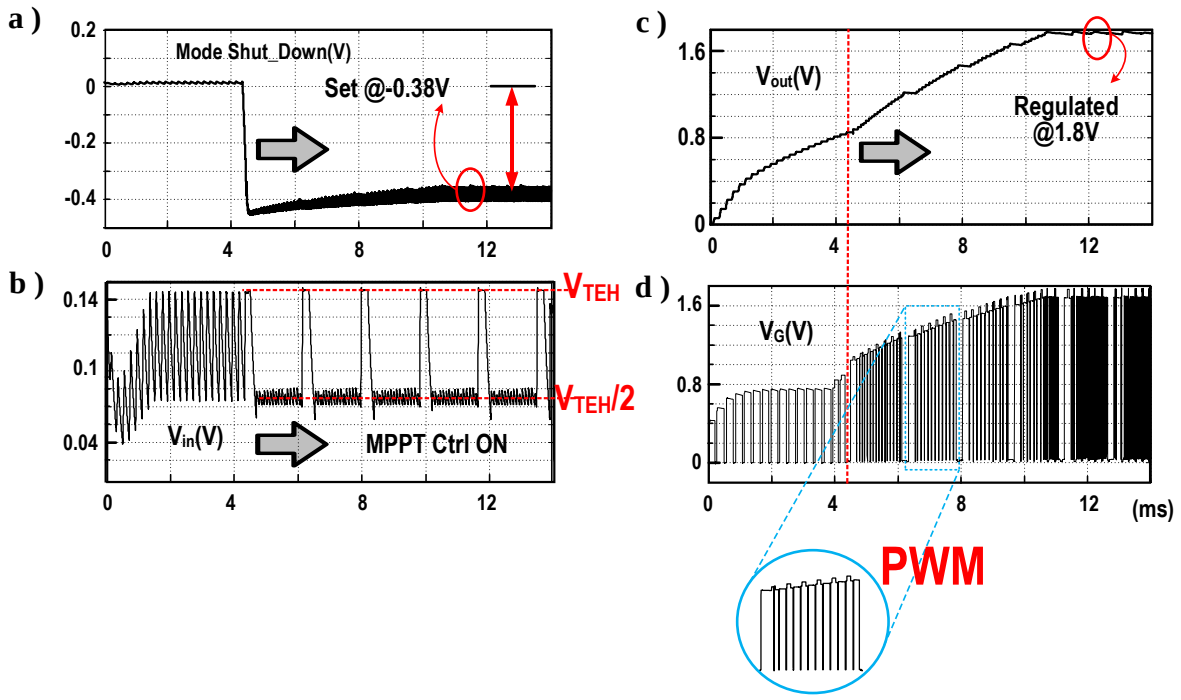


Fig. 2.11: Relevant waveforms during MPPT operation: (a) shut-down voltage, (b) input voltage, (c) output voltage, (d) gate voltage.

The measured inductor current and the voltage V_S at the inductor current zero-crossing are shown in Fig. 2.10. The potential at the node V_S shows no undershoots or overshoots while crossing zero which implies almost perfect zero detection during $\bar{S}$ period. Fig. 2.11 shows the simulated

waveforms during MPPT control. With this MPPT regulation scheme, we can approximate (to a first order) the available average load current during one period as:

$$I_{LOAD,avg} \approx \gamma B f(L_0, D, f, R_\alpha) \frac{V_{TEH}}{2R_{TEH}} \frac{L_0}{T}, \quad (2.13)$$

where $B = \frac{V_{TEH}}{V_{CC}}$ is the reciprocal boosting ratio, $f = \left(1 - e^{\frac{DR_\alpha}{fL_0}}\right)^2$ and γ depends on the input ripple and the speed of the feedback loop [16], [18]. As (2.13) implies, the higher output current and boosting ratio require higher inductance value.

2.5 Compound TEH Platform

We have designed and fabricated compound TEH module (Fig. 2.12) in order to meet the stringent anatomical and biophysical confinements of living subjects including but not limited to

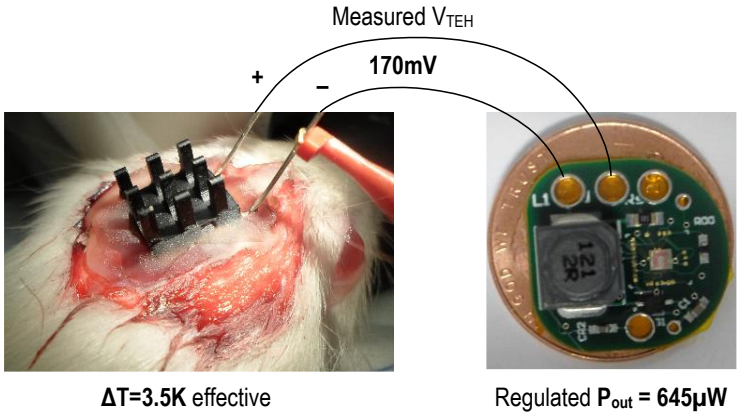


Fig. 2.12: Implanted TEH module shows 170mV in-vivo, with 645μW regulated output power.

rats. The animal's cerebrovascular system is directly in contact with the bottom part of the TEH platform which is made of bio-friendly material – titanium. At the bottom titanium plate, we have arranged 3x3 thermo-electrical elements (μTEGs [19]); each μTEG behaves as an independent voltage source. The μTEG array is attached to the titanium plate with a thin layer of thermo-conductive glue.

Each µTEG element, [19], is composed of n thermocouples, where every thermocouple consists of two thermoelectric bars that are made of different materials, and joined at one end. Because of the thermoelectric Seebeck effect, the thermoelectric electromotive force, is created in the presence of a temperature difference between these two materials. The voltage is proportional to the junction temperature difference ΔT and to the difference between the Seebeck coefficients $S = S1 - S2$, and $U = S\Delta T$. Thermocouples are usually made of semiconductors and connected electrically in series to obtain higher output power and voltage. The generating performance of a µTEG is primarily evaluated in terms of its output power. More power at the output means more thermocouples connected in series (bigger area) and/or higher temperature gradient ΔT.

By serially stacking three µTEGs and connecting these stacks in parallel the output power can be increased while maintaining the equivalent source impedance of a single µTEG source. Post-

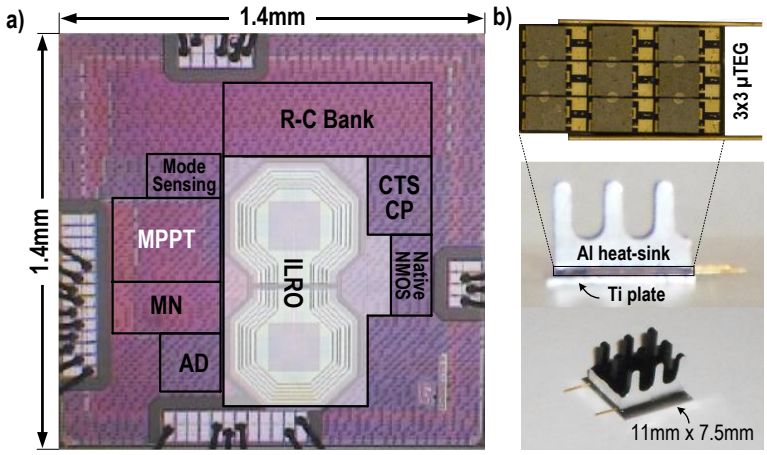


Fig. 2.13: (a) Chip micrograph, (b) Fabricated compound TEH platform.

fabrication measurements of our TEH structure showed an equivalent 6.3Ω of internal impedance. Slight increase in impedance is due to the bond wires and Ohmic contacts. Also, 11mm x 7.5mm heat sink is utilized, which is large enough to cover all three TEGs, while the bottom plate extends 1mm on both sides to accommodate skull-fixing screws. Further, in order to confine the heat flow and prevent unwanted heat leakage on the side, the exposed space between the heat sink and the

bottom plate is filled with a biocompatible insulator. By controlling the output resistance, we have an explicit control over the power delivered to the load. Compared to standard animal head-stages, our design occupies a smaller volume and does not present a burden to animal behavior.

2.6 Measurements Results

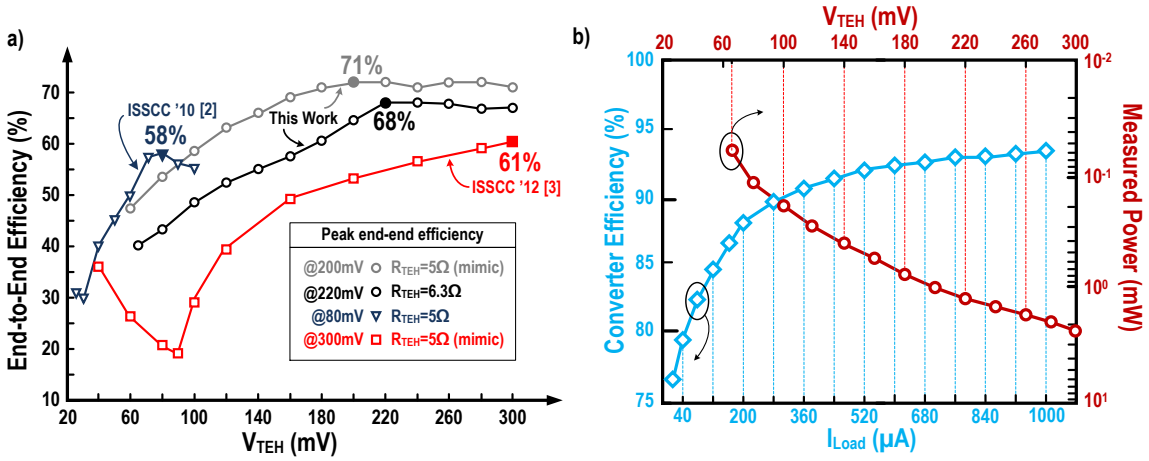


Fig. 2.14: (a) End-to-end efficiency comparison with state-of-the-art. (b) Measured converter efficiency as a function of the output load current (left vertical axis), and measured output power (right vertical axis) as a function of the source voltage.

The proposed low-power, boost-converter for TE harvesting applications was implemented in a 65nm CMOS technology. Fig. 2.13 shows the micrograph of the chip with the MPPT controller occupying 0.06mm^2 while the ILRO-based startup block takes 0.65mm^2 of the chip area. As suggested in [10], [20], designing the Colpitt's-based or multi-stage ILRO would require more on-chip inductance, with inevitable increase in the chip area. For bench-top evaluation, a voltage DC-source together with serial resistance was employed to mimic the TEH. The peak end-to- end efficiency is defined as the ratio between the maximum available power delivered to the load during the impedance matching ($V_{in} = V_{TEH}/2$) and the maximum available power from the thermo-electric harvester

$$\text{Peak End to End Efficiency} = \frac{P_{out|V_{out}=1.8V}}{P_{in|V_{in}=V_{TEH}/2}}. \quad (2.14)$$

The chip-verification comprises of two parts: cold startup, and MPPT operation with mode change. Peak end-to-end efficiency is 68% at V_{TEH} = 220mV, outperforming prior art, as shown in Fig. 2.14. Measured waveforms during the startup mode and the MPPT control (Fig. 2.15) imply fully autonomous operation down to V_{TEH} = 65mV. In our bench-top setup, we measured the efficiency for V_{TEH} from 60mV to 300mV.

V_{TEH} is sampled when $\bar{S}$ is high and stays around $V_{TEH}/2$ when $\bar{S}$ is low. In Fig. 2.15, V_{TEH} and V_{IN} are 65 mV and 32 mV, respectively, demonstrating the functionality of the MPPT control. In the MPPT mode, the main contributors to energy loss are the inductor resistance and switching losses associated with the M_{PS} switch, as predicted by simulations. In order to demonstrate fully-autonomous operation, we also conducted in-vivo testing. Collaborators from the UCLA Department of Neurology have provided us with adequate infrastructure for the in-vivo test. There was no craniotomy on the animal (for microelectrode insertion). The entire experiment lasted about 20 minutes, after which the animal got stitched and returned to its habitat, fully recovered. Our

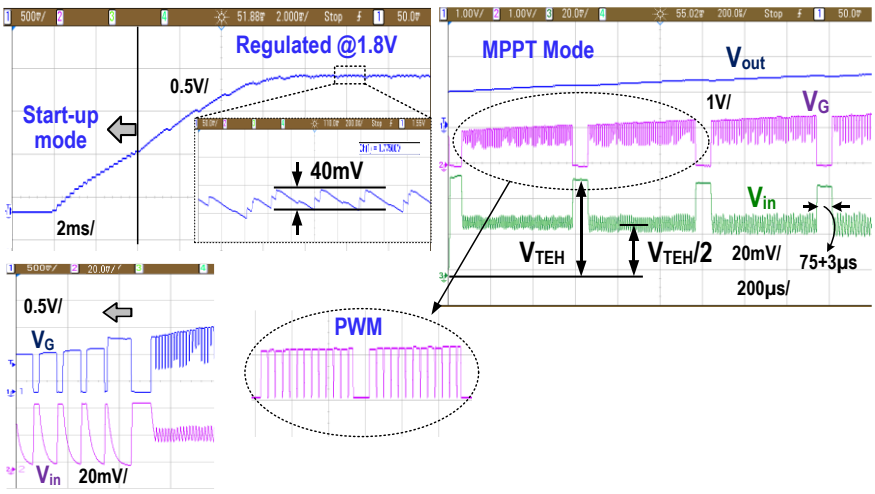


Fig. 2.15: Measured lab waveforms show V_{TEH}=65mV and regulation to 1.8V in less than 20ms.

experiment was far less invasive than a typical animal surgery involving craniotomy and cementing of the head stage. The heat sink is only 9 mm tall and it is smaller and lighter than head-caps used in animal neuroscience, hence it does not negatively impact the animal behavior. It is necessary to ensure thermal flow through a small area, hence the need for a thermal antenna. In fact, our TEH heat-sink is much less invasive than head-caps used today. We have demonstrated feasibility of our technology in a neuroscience application. Further opportunities exist in environmental monitoring and similar areas.

Table 2.1: Comparison with state-of-the-art thermal energy harvesters.

Reference	[2]+	[3]	[4]++	[5]	[7]+++	[22]	[23]	[24]	[25]+	This work
Process	0.13µm	0.35µm	0.13µm	65nm	0.13µm	0.18µm	65nm	0.18µm	0.13µm	**65nm**
Startup mechanism	External voltage	Mechanical	White noise	Electrical	RF-Kick startup	No Start-Up Unit	Electrical	Electrical	Electrical	**Electrical**
Min $V_{start-up}$	650mV	35mV	40mV	50mV	220mV		80mV	350mV	150mV	**65mV**
Regulated V_{out}	1V	1.8V	2V	1.2V	1.2V	0.5V	0.7V-1V	1.8V	1.8V	**1.8V**
Peak efficiency end-end (conv.)	63% e-e (75% cnv)	58% e-e (91% cnv)	61% e-e (N/A cnv)	N/A (73% cnv)	N/A (83% cnv)	N/A (83.6% cnv)	N/A (73% cnv)	N/A (80% cnv)w/o regulation	N/A (73% cnv)	**68% e-e (92% cnv)**
Off-chip L+C+R	1+3+0	3+4+0	2+5+0	3+4+0	1+2+0	1+3+0	4+2+0	0+7+4	0+6+0	**1+2+0**
Regulated Power Density @ΔT=4K (µW/cm²)	22	34	N/A**	162	80*	N/A**	N/A**	128	N/A**	**1285**
Tracking Time	N/A	~20ms	~20s	~25ms	~ 50s	N/A	~20ms	<180ms	N/A	**<20ms**
In-Vivo	NO	YES	NO	NO	YES	NO	NO	NO	NO	YES

$V_{start-up}$ refers to V_{TEH}
(e-e) end-to-end
(cnv) converter

(+) no MPPT
(++) uses transformer
(+++) operation down to10mV, but need 220mV for the startup.

(*) Harvested Power During In-Vivo experiment without reported area/volume of their system
(**) Partial Solutions without In-Vivo experiment and system reported

The harvesting platform is mounted on the head of a rat and temperature gradient of 3.5K is measured while the system was able to harvest 645µW regulated output power, with 61% end-to-end and 92% converter efficiencies, Fig. 2.15. The power level indicates that TE harvester outputs V_{TEH} = 170mV at stable state. The keys to the improved efficiency lie in integrated power-efficient startup unit, the compact TEH source, and the fast fully-analog MPPT controller. Our fully-autonomous thermoelectric harvester shows a 7.9x improvement in regulated power density from a 0.83cm^2 surface area (Table I) relative to the current state-of-the-art. With one storage cap and

only one off-chip inductor, the mote-PCB paves the road to the new level of miniaturization. The compound TEH together with a small PCB occupies less than $1cm^3$ of volume and weighs less than 3g. TEH platform presented in this work is the new state-of-the-art in the factor and power density levels. With the presented approach, elimination of bulky batteries in size-constrained neural recording sensors becomes possible and their integration presents the future work.

2.7. Conclusion

This work demonstrated a fully-integrated, electrical startup boost converter for autonomous thermo-electric harvesting. A standalone thermoelectric platform integrates our efficient power management IC with customized TEH into a single micro-system. We fabricated our TEH with tiny µTEGs, which have a great power levels (measured 645uW end-to-end), and with customized and optimized platform we were able to maintain stable temperature gradient over a 9mm thin platform. We have shown: 1) the most efficient single-ambient-source circuitry reported to date (68% vs. 61% in prior work) while achieving 2) the most compact PCB + TEG reported to date (6.3x smaller than prior art) and additionally providing 3) the first demonstration of fully autonomous TEG operation in real environment (vs. lab-bench). We require only 1 off-chip inductor and two small off-chip capacitors. Overall, this leads to ~6x smaller PCB footprint than previous work from [3] and [5]. Our analog MPPT minimizes energy loss and achieves <20ms output regulation (very important requirement in the event of temperature fluctuations).

CHAPTER 3

A Miniaturized 64-Channel Neuromodulation Platform for Simultaneous Stimulation and Sensing

3.1 Introduction

Today, only in USA about 40 million people suffer from various neurological disorders, like Parkinson disease, epilepsy, tremor, memory losses, depression, Alzheimer's disease, etc. Neural interfaces are used as a part of therapy to mitigate these conditions. They are not just effective in restoring various functions and improving the quality of life in patients, but they also help our understanding of the brain. Current neuromodulation (NM) devices, Fig. 3.1, are not only bulky in size, there is a lot of implanted hardware in human body and the wires that are sticking out are creating a lot of discomfort to the patient. They have a small number of low-precision contacts,

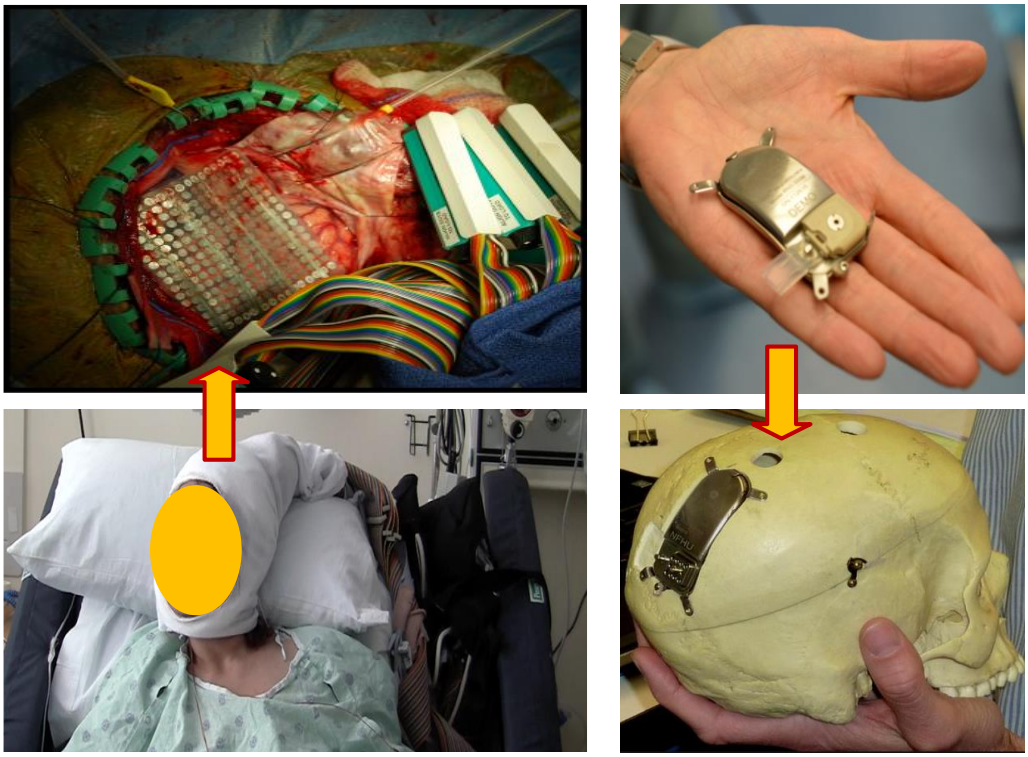


Fig. 3.1: Current NM devices. NeuroPace RNS-300.

and limited sensing capabilities. No NM device has the ability to record neural activity in the presence of stimulation artifacts [33]. This technology is decade old and it seems that these complicated disorders cannot be treated efficiently with these old tools. Essentially, there is a need for better platform technology that will reduce the form factor, introduce more flexibility and improve the power efficiency of the device, so the battery life can be extended.

In the core of every NM interface, we have a unit for sensing neural activity and another unit which is responsible for delivering responsive stimulation. Together, they are indispensable tool in treatment of the brain disorders. The next generation of NM devices would require concurrent stimulation and sensing abilities, where the stimulation parameters can be adapted in real-time based on the feedback provided from sensing unit, Fig. 3.2. A real-time stimulation parameter update would directly follow the dynamics of the brain and cause the better therapeutic results over time.

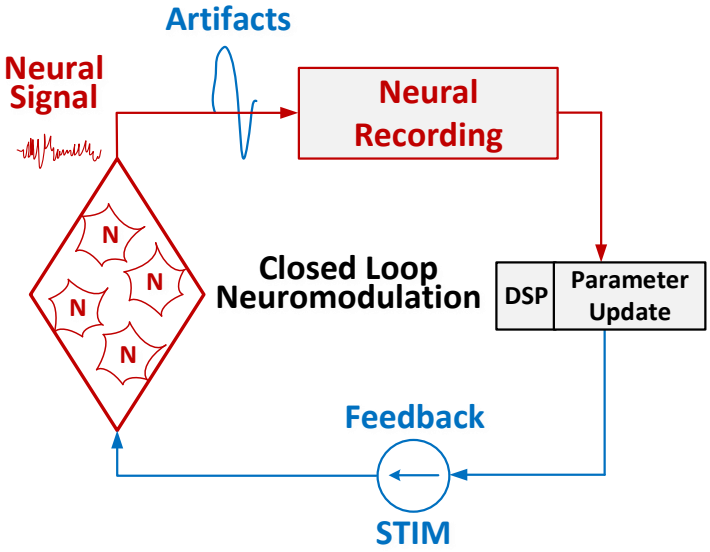


Fig. 3.2: Closed-Loop Neuromodulation.

State-of-the-art research [34-38] reports low-power neuromodulation (NM) units, mostly for animal use, with modest level of integration needing several cm^3. Requirements also include a high linear input range sensing unit (>100mV_{p-p} and THD<-80dB) and a differential stimulation

strategy to prevent tissue large common mode swings. Recently, authors in [39], proposed implantable NM module for human patients, but their approach introduces several shortcomings. Concurrent, charge-balanced multi-channel stimulation is not possible, while the front-end linearity is poor due to the limited THD performances for the high input signal. Also, since front-end is chopped, the input impedance is reduced. Further, front-end should be able to sample LFP signals at >5kHz frequency to allow for the removal of high-frequency stimulation artifacts. Such high-fidelity artifact removal is not possible with under-sampled input data at 1kHz as in [39]. Insufficient linear range and single-ended stimulation would imply huge voltage excursions in the tissue and incapability to perform simultaneous stimulation and recording.

Neural stimulation is purposeful modulation of nervous system activity. Today, it is widely used, from cochlear implants to neurological disorders treatment, and it is proven to have a potential to treat brain disorders in patients that do not respond to the medications. E.g., deep-brain stimulation (DBS) can provide symptomatic relief for neurological patients by emitting electrical

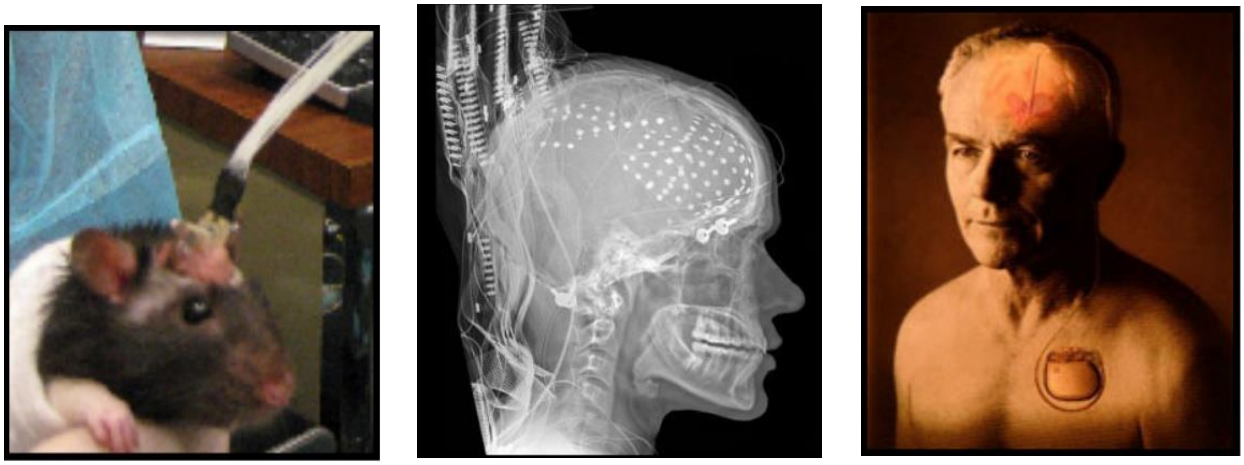


Fig. 3.3: Neural Interfaces Applications-Behavioral Neuroscience, Pre-Surgical Mapping, Decease Therapies.

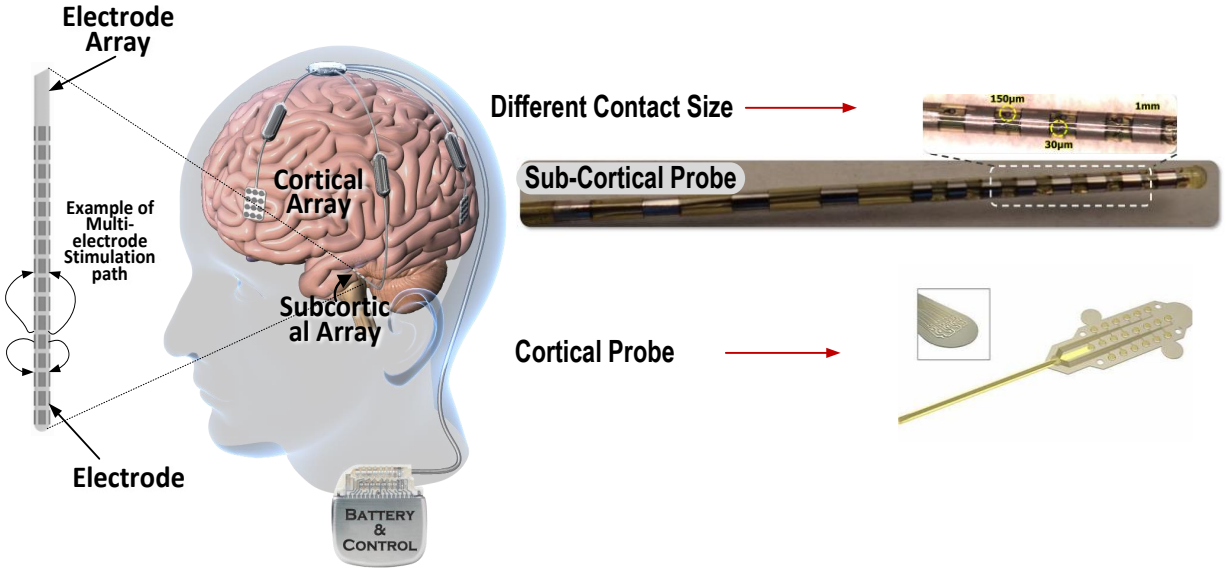


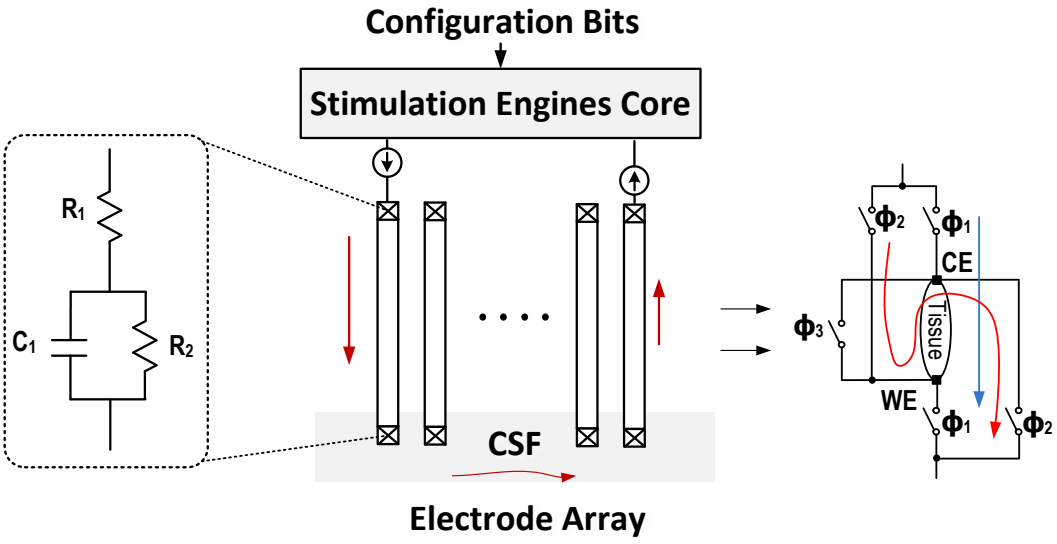
Fig. 3.4: High-precision multiscale Neural Probe. Cortical and Sub-Cortical Lead.

pulses. It is efficacious in Parkinson's disease and other movement disorders, which are anatomically focal, where open-loop stimulation on just one contact is sufficient. The same technology doesn't show therapeutic benefit in network-scale indications such as depression or Alzheimer's disease, where a more precise localization as well as distributed sensing and stimulation are necessary. Furthermore, various neurological conditions often stem in multiple brain regions, so modular neural interface with higher channel count is requirable. Also, continuous open-loop stimulation can lead to harmful outcome and it can lose positive effect during the time because of the changes in the brain. Closed-loop system that updates the parameters in real-time, will significantly enhance the effects of stimulation, mitigate the undesirable outcomes and improve our understanding of the hidden brain dynamics.

Over past decades, with the advances in technology, many types of neural stimulators have been proposed. The main purpose of stimulator is to create a desired neurological response by providing the charge from or into the neural cells. The charge amount needed to inhibit a neural response depends on many factors: tissue degeneration level, type of neuron, interface (neural

probe), etc. Also, while stimulators required to provide a wide range of stimulus energy to the tissue, the power needed for the stimulation is usually dominant and itself dictates the overall NM power consumption. Adopting cutting-edge power management circuits for the next generation NM devices, that can support different power delivery options (wireless, wired, rechargeable batteries, etc.) and can improve performance and power efficiency is an imperative.

This work demonstrates a miniaturized, implant-scale NM implant for concurrent sensing and stimulation which includes flexible, electrode-agnostic, 8-driver-to-64-contact stimulator that can deliver up to 5.1mA per driver; the implant also houses a full-fledged, multi-mode power



Φ_1	Φ_2	Φ_3	Mode of operation
On	Off	Off	I_{stim} flows through electrodes
Off	On	Off	$-I_{stim}$ flows through electrodes
Off	Off	On	Removal of residual charge

Fig. 3.5: Electrode-Tissue Model. Biphasic Differential Neural Stimulation.

management unit that supports different NM applications (cochlea implants, DBS, retinal prosthesis, etc.) and can extend the battery lifetime compared to state-of-the art.

3.2 Types of Neural Stimulation and Biphasic Current Pulses

Neural stimulation is delivered as a train of controlled current pulses, which are usually zero-mean, into specific brain regions to modulate brain activity. Stimulation is performed through neural probe or micro-electrode array, Fig. 3.4, which serves as an interface between neurons and the electronic circuitry. As a first order approximation, electrode-tissue model can be depicted as it is shown in Fig. 3.5, where R_1 is the sum of Faradaic charge transfer resistance and trace resistance, C_1 models double layer capacitance, while R_2 depicts so called Warburg impedance, [41]. For all practical reasons during stimulator design, R_2 can be neglected.

Figure 3.6 illustrates different amplitude and timing parameters that can be set during the active stimulation. Also, different kinds of stimulus waveforms can be adopted, depending on the application. The physiological response generated by the stimulation is directly dependent on the waveform. Among different pulse shapes that can be employed, biphasic current pulses, are preferred due to the charge balancing property. During biphasic stimulus, anodic phase ensures positive charge delivery to the tissue, while cathodic phase provides a negative charge delivery. The inter-phase delay separates the cathodic (CP) and anodic pulse (AP) so that the AP does not change the effect of the CP. Ideally, these two charge amounts should be equal, so that after one bi-phasic pulse, there is no remaining charge in the tissue. Since, the ideal matching for all practical reasons is not possible, to ensure safe operation, the residual charge has to be removed. Hence, these two phases are followed with shorting phase, in which electrodes are shorted to the gnd or some other DC-level and all residual charge is removed. Switches Φ_1, Φ_2, Φ_3 are responsible for the bi-phasic stimulation control. Duration of the shorting phase is directly proportional to the R_1-C_1 constant of the electrode.

Safe operation of stimulator is necessary, otherwise the tissue damage may occur. Tissue

damage can be induced in several different ways: i) Heat Dissipation – Implanted hardware releases too much heat that can effectively cause the temperature rise at the electrode-tissue interface – FDA safe limit < 2^{o}C; ii) Charge imbalance during stimulation; iii) Excessive Charge Injection – Size (cross-section) and the electrode material dictates the limit for the safe charge density.

Neural stimulation can be monopolar and differential. In monopolar stimulation, usually there are several stimulating electrodes and one return electrode, which plays the role in the charge recovery. This method shows shortcomings if the precise stimulus localization is necessary. Figure 3.5, shows an example of differential stimulation, where the pair of electrodes is used at each

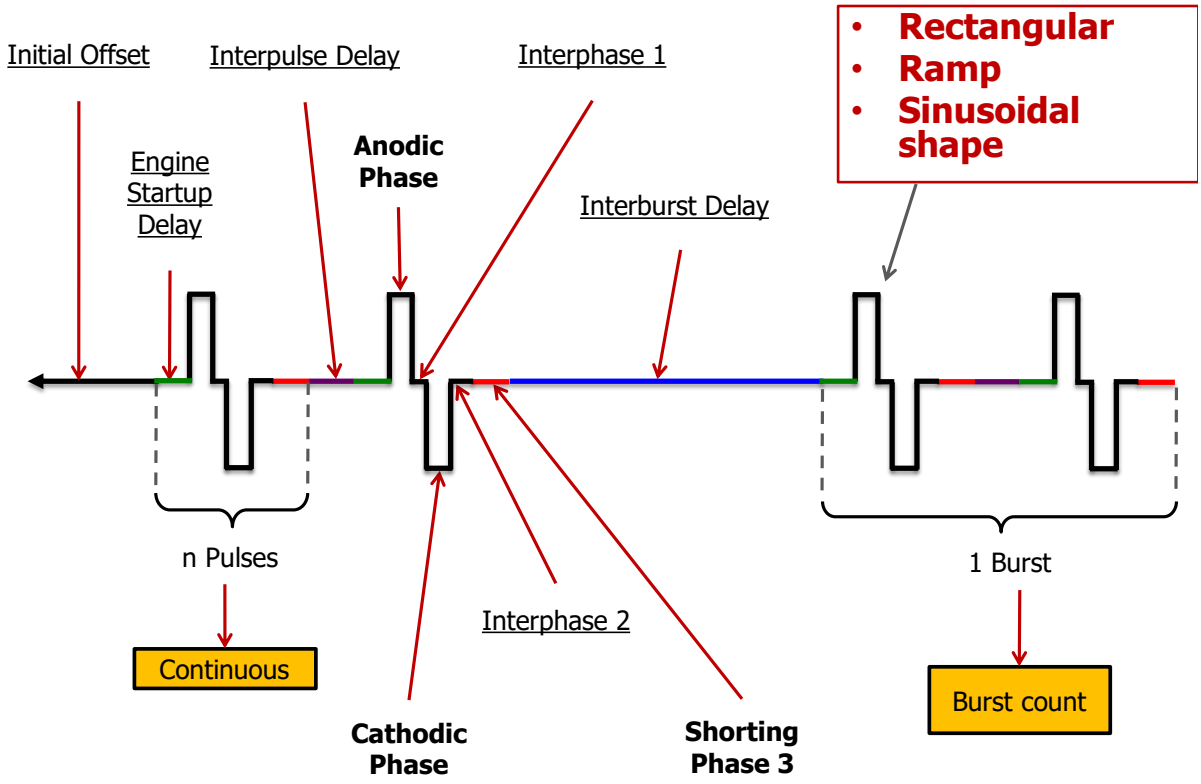


Fig. 3.6: Neural Stimulation – Waveform Shape.

stimulation cites; the current is pushed from current source through working electrode (WE) and it closes the loop through CSF (conductive cerebrospinal fluid), counter electrode (CE) and current

sink. Differential stimulation is preferable method during the simultaneous stimulation in which multiple stimulus drivers are used for concurrent stimulation on several electrode pairs. Also, diferential stimulation prevents large common mode swings.

There are several types of neural stimulation presented so far. Each type brings different type of drawbacks and benefits. Voltage Current Stimulation (VCS), proposed in [42], ensures power efficient stimulation, but since the electrode impedance may vary over time and position, the charge balancing is problematic. Recently proposed Switched-Capacitor Stimulation (SCS), [43], offers good tradeoff between safety and efficiency, but SCS requires a big number of off-chip capacitors and it cannot be used in multi-channel, simultaneous stimulation, since current splitting among channels is undesirable. As widely used method in neural stimulation, current-controlled stimulation (CCS) offers an accurate charge control, but it reduces power efficiency because of the voltage drops across current mirrors (sink/source) in the output stage of stimulators.

Different applications need different types of electrodes (deep brain stimulation (DBS), epiretinal stimulation, etc); macro and micro electrode contacts show big range in tissue-electrode capacitances – from a few nF to a few μF. To support various electrodes and allow a wide range of stimulation currents, it is crucial to have the stimulation mechanism that accounts for the "capacitance-dominant" electrodes and extends the on-chip voltage headroom for the stimulator circuit. Since, our target was a design of simultaneous, multichannel and electrode agnostic stimulation engine that will compensate for the variability of electrode-tissue impedance, efficient CCS design is a preferable choice.

3.3 Design requirements

Our work targeted the next generation neural interface, that is minimally invasive, and addresses the demands for limited area and power, Fig. 3.7. It should provide a real-time, full duplex communication during concurrent stimulation and recording of neural signals. Further, our

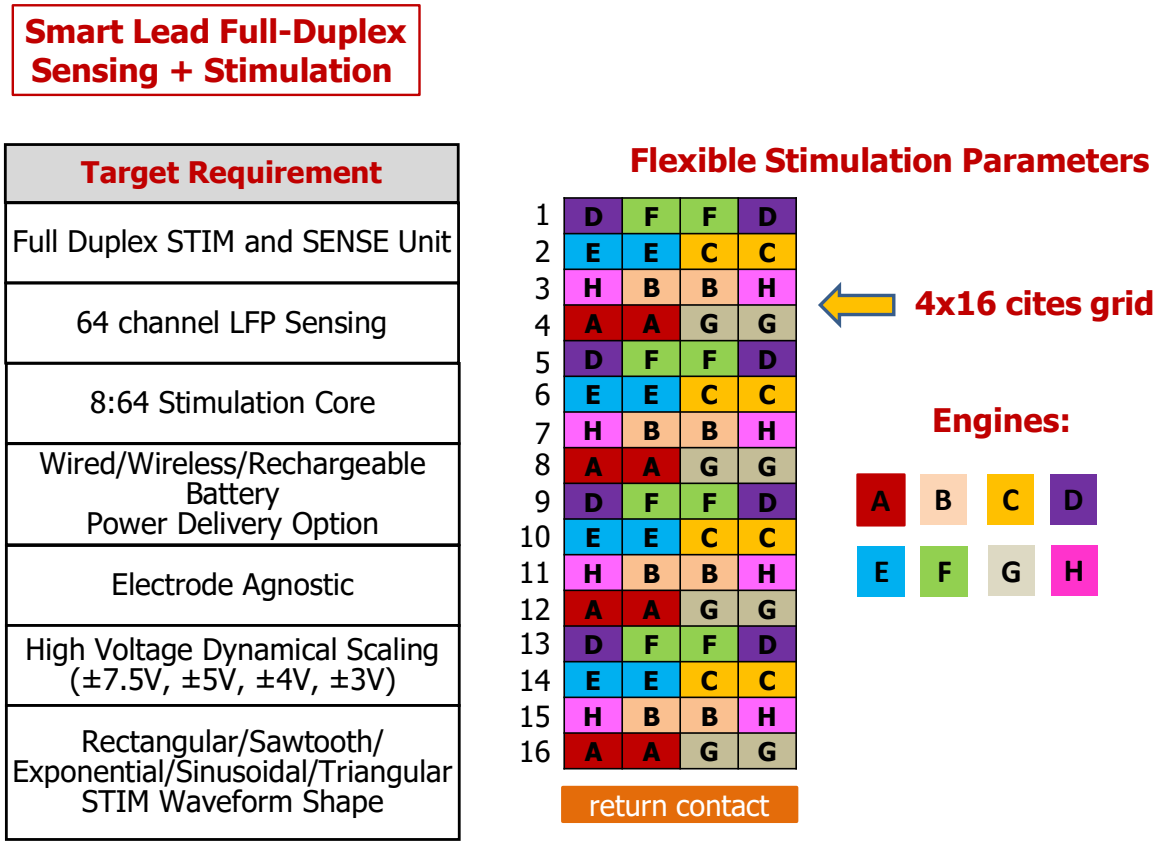


Fig. 3.7: Smart Lead Design Requirements.

modular approach and scalable architecture should allow gathering data from a grid of NM implants.

Our NM implant houses $100mV_{p\text{-}p}$ linear input range sensing unit recently demonstrated in [40]. In first stage, we develop a wired supplied 32-channel module together with 4-driver-to-32-channel fully flexible stimulation module delivering up to 3.1mA per engine in a 128-channel implantable closed-loop system. In stage-2, we scale up sensing capability to 64 channels,

stimulation to include 8 drivers (A-H, 5.1mA each), supported with highly efficient wireless power and data link. This interface will be integrated together with LLNL implantable electrode arrays and packages into a 64-ch modules that will be further assembled into a 256-channel system.

3.4. System Architecture

There are several primarily targeted applications. The first one considers the implantable system for the DBS treatment. The system (Fig. 3.8) consists of several NM "smart lead" units, each with *stim* and *sense* ICs assembled to cortical or sub-cortical leads using high-density

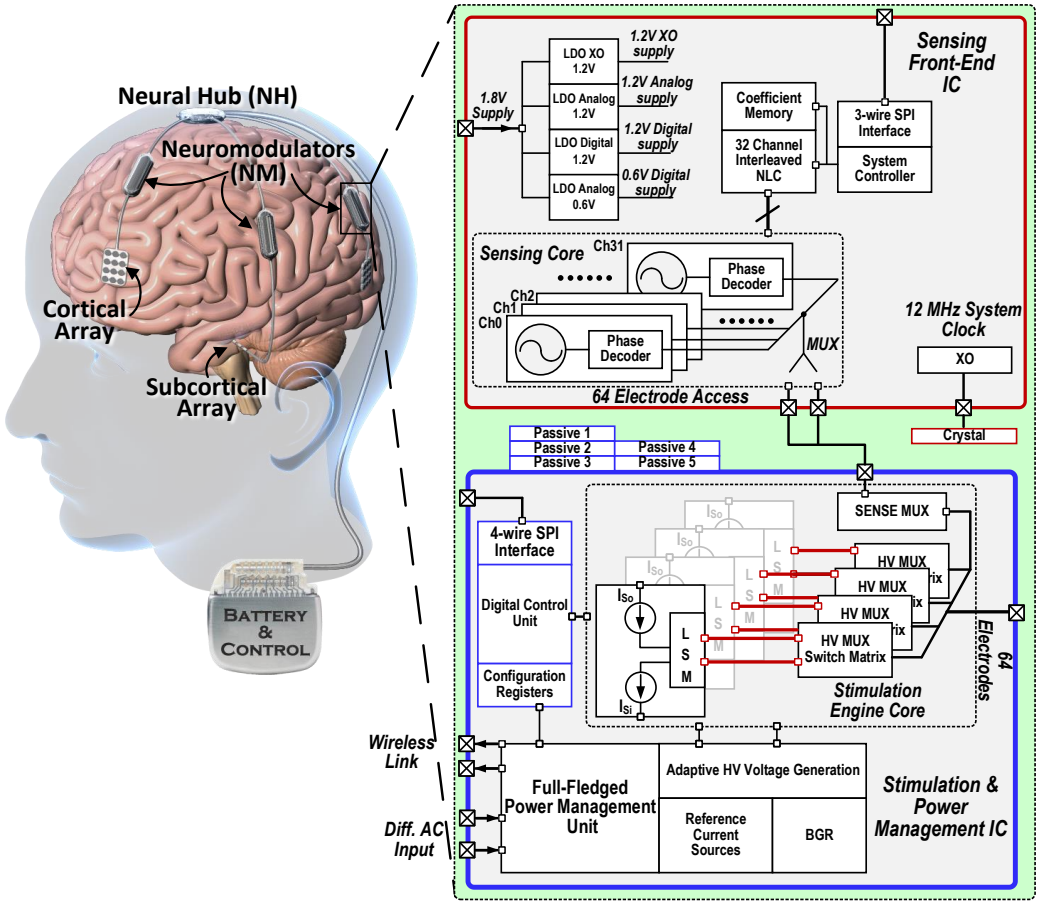


Fig. 3.8: Proposed implantable system with main STIM/PM IC and Sensing Front-End IC.

feedthroughs. The Neural Hub (NH) serializes data from 32-ch/64-ch NMs and communicates with control module in the chest. The second one (Fig. 3.9) extends the capabilities of the first one

and enables wireless power transfer (WPT) which is usually the only way of supplying fully implantable medical devices and plays unavoidable power solution for cochlear implants and retinal prosthesis. Our target was an implant for restoring active memory, placed at temporal lobe, that besides the NM core provides also the wireless data link, Fig. 3.9.

3.4.1. Stimulation Engine

In the core of every stimulation engine (SE) we have a current source and/or current sink depending on types of neural stimulation. The electrode-tissue impedance varies over time and its value also depends on electrode placement in the nerves. Also, different electrodes have different impedances depending on material they made of and depending on the size of contacts (range - 100's Ω-1MΩ). To support simultaneous multi-channel, electrode agnostic stimulation we need a very high output impedance current source/sink for a wide range of stimulus currents. Furthermore, the current mirrors should have a high output compliance to compensate the source/sink additional

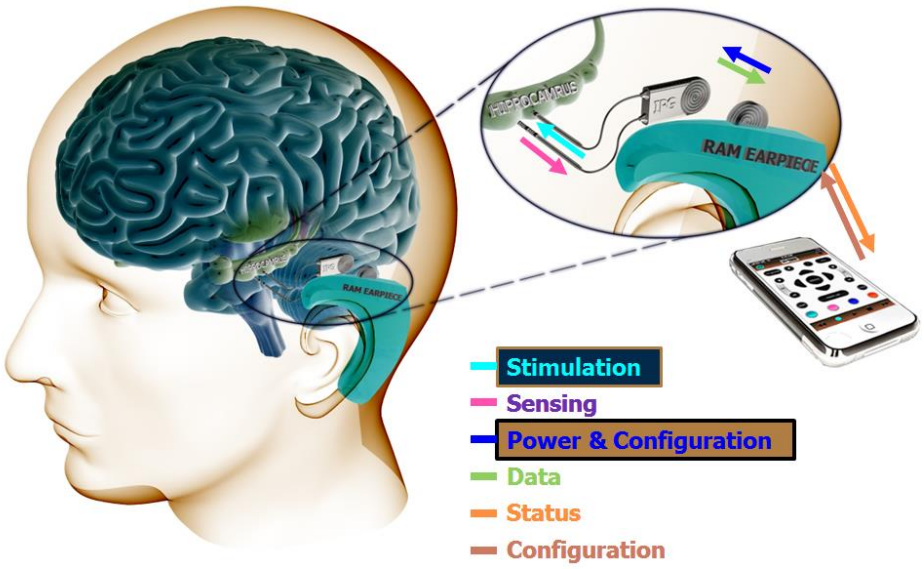


Fig. 3.9: Implantable RAM (Restoring Memory Device) Unit.

voltage headroom requirements in CCS, so that the most of rail-to-rail voltage can be dedicated to the output electrode pair (differential voltage). Motivated by the work in [44], Fig. 3.10(top) shows

the core of our (SE) – very precise, high-compliance and ultra-high output impedance current mirror for source/sink part of SE. These features are possible due to the combination of the positive and negative feedback loops employed in the circuit. This high-voltage, current mirror is superior in gathering super-high output impedance, high accuracy and high compliance ever achieved by any stimulation engine.

The core of this current mirror is essentially made of two feedback loops: first - positive

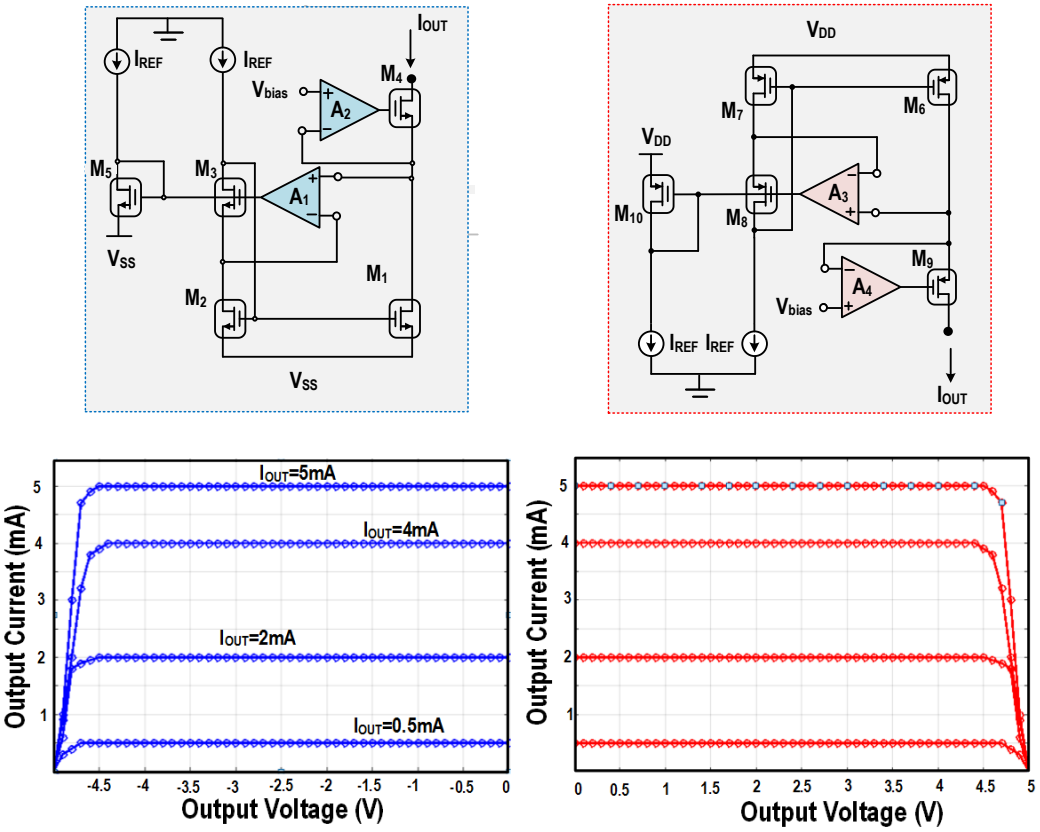


Fig. 3.10: Current Sink/Source and correspondent DC Output Characteristics.

feedback (PF) (made of the error amplifier A_1 and transistors M_3 and M_1) and the second with a negative feedback (NF) (A_1 and M_3). PF is always synchronized (in phase) with the input signal and it is determined as a positive loop gain (LG) around a feedback loop. Keeping only amplifier A_1 and transistors M_3 and M_1 as a current mirror will limit the output voltage by a single V_{DSAT}. But this structure has a serious drawback – for increased values of output voltage, M_3 goes into

linear region, since the input current source and aspect ratio of M_2 dictates the DC value of V_{G2}, [44]. The output voltage is bounded:

$$V_{OUT} \leq V_{G2} - V_{DSAT3}. \tag{3.1}$$

To prevent this, an extra NF that includes another operational amplifier (A_2) is added in the circuit. The plus terminal of amplifier A_2 is connected to a bias voltage, V_B. This would imply that the voltage at the plus terminal of A_1 is going to be set to a desired value and enlarging values of

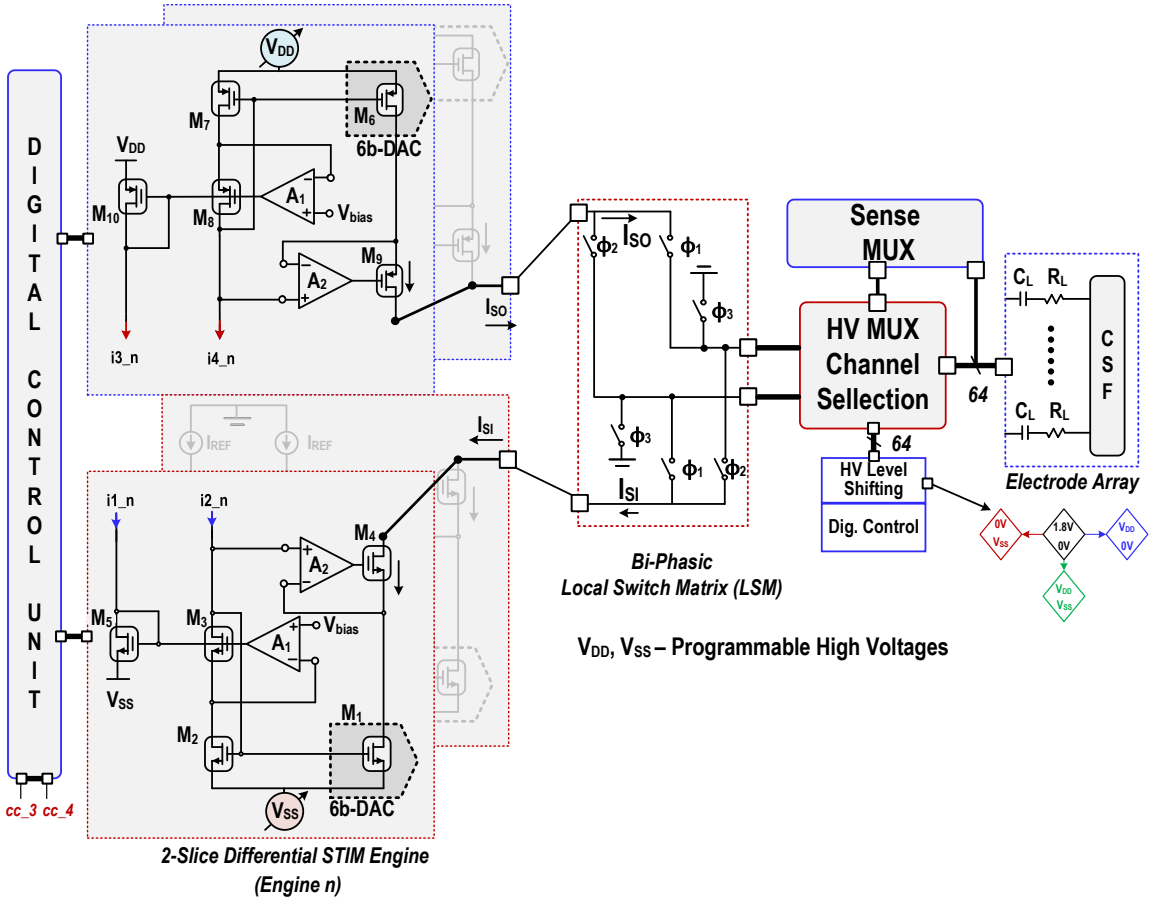


Fig. 3.11: Stimulation Engine Architecture.

V_{OUT} will not push the loop run by A1 to its bound. Connecting the plus terminal of amplifier A_2 to a V_B facilitates the V_{OUT} swing by shielding the input of the A_1. From the small signal analysis, the output resistance of the current mirror can be expressed as

$$R_{OUT} = r_{04} g_{m4} r_{01} g_{m2} A_2 R_{IN}, \tag{3.2}$$

where R_{IN} represents the input resistance of the mirror. This clearly shows that amplifier A_2 also contributes to the boosted output resistance. The output resistance is boosted and the voltage compliance is equal to V_{DD}-$2V_{DSAT}$. Folded cascode PMOS/NMOS amplifiers are employed to ensure the proper control loop operation at voltages close to V_{SS}/V_{DD}. Figure 3.10(bottom) shows the DC output characteristics for current source/sink across the wide range of output currents.

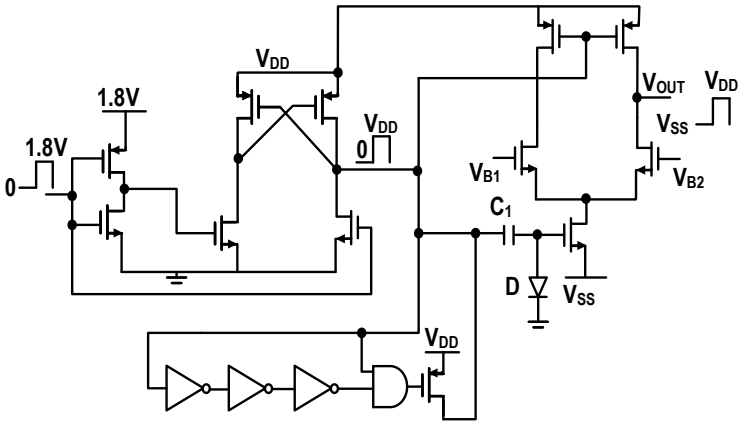


Fig. 3.12: Unipolar-to-Bipolar High Voltage Level Shifter.

Figure 3.11 shows the complete Stimulation Engine architecture. Digital Control Unit (DCU) activates the STIM engine only during active stimulation. The engine comprises of 2 driver slices, each with a 7-bit current source/sink for differential stimulation (to reduce artifacts), with integrated high-voltage (HV) level shifters (LS). By employing previously explained current mirrors, the output impedance of the current mirror is boosted to 100'sMΩ-1GΩ. This architecture ensures accurate current matching even for the large voltage swings (94% $V_{rail\text{-}to\text{-}rail}$) at the electrodes.

The shape, amplitude, various timing parameters and SENSE/STIM MUXs are configured by DCU and updated on-the-fly. The output of STIM engine is connected to the local switch matrix (LSM), which is employed for the biphasic control and post-stimulation active charge balancing. Any residual charge on the electrodes is cancelled out by shorting. The HV channel-selection MUX

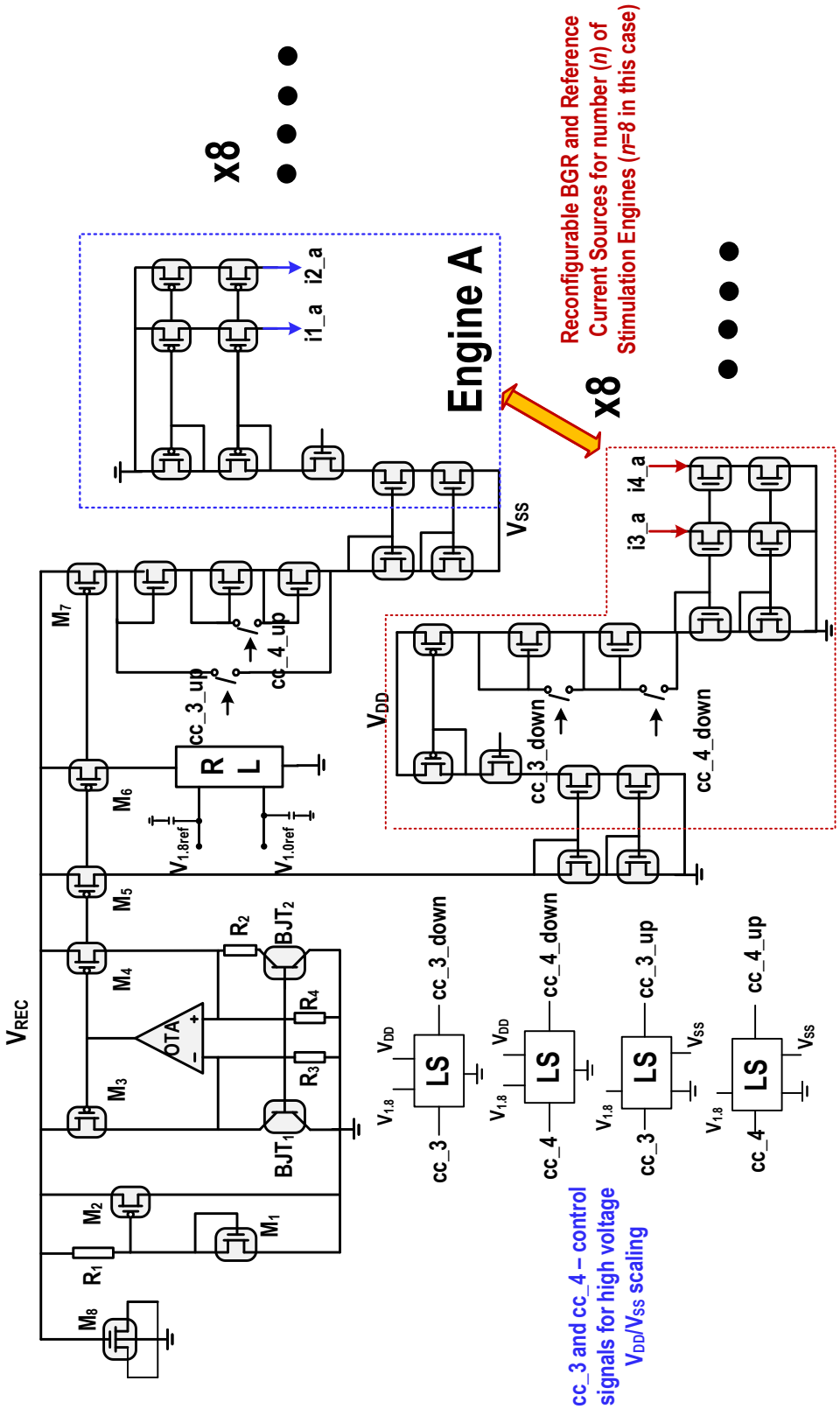


Fig. 3.13: Adaptive BGR and Reference Current Source.

provides a fine spatial granularity (8:64 MUX with integrated HV LS), for a multitude of

stimulation sites. Since, the control signals are coming in 1.8V level and HV STIM engine requires unipolar-to-bipolar voltage conversion, we employed the specific architecture (Fig. 3.12) for the HV Bipolar Level Shifting. The stim MUX determines the sense-IC accessibility to electrodes and protects the sense-IC from voltage overstress, since sense-IC is designed in lower node technology.

Power efficient stimulators are necessary in energy-limited systems. The stimulator engine dissipates significant power at its output stage, especially when it delivers small currents. This is the main drawback of the CCS. To additionally save the power, a high-voltage adaptive-rail (V_{DD}/V_{SS}) is provided to accommodate voltage drops across high electrode impedances. When the electrode cites are stimulating with specific current levels, DCU can configure (control bits cc_3, cc_4) the output of the high voltage generators to produce stimulation power supplies V_{DD} and V_{SS} according to the stimulation electrode needs (lower current – lower stimulation voltage and vice versa). Adaptive HV stimulation approach prolongs the battery life up to 10x. At the same time, reconfigurable BandGap Circuit and Reference Current Source (Fig. 3.13) is designed to be immune on these V_{DD}/V_{SS} changes.

3.4.2. Sensing Unit

Recording of neural activity plays an important role for diagnosing neurological conditions. Presence of biomarkers in recorded traces gives the neuroscientist valuable information. A frequency band occupied by the neural signals of interest, and picked by the electrodes, goes up to 6kHz. The local field potentials (LFPs) occupy a frequency band from 1Hz to 200Hz, while the action potentials (APs) fall within 200Hz to 6kHz frequency band. Also, the peak amplitude of LFP signals is 1 mV, and the peak amplitude of action potentials is up to 100 µV.

To ensure concurrent stimulation and recording, a high dynamic range sensing front-end unit is needed. The front-end must digitize neural signals to the required resolution of 8 bits in the

presence of stimulation artifacts and the required signal to noise and distortion ratio (SNDR) for differential signals is >12 bits, [45]. Saturation of the front-end would cause a loss of information (blanking).

Apart from the high input dynamic range, if the large DC offset is present at the sensing electrodes, it can create constant currents at the electrode because of the finite DC input-impedance of the sensing front-end unit. This will induce a tissue damage over time. To ensure a proper and safe functioning of front-end, its DC input impedance needs to be larger than 1GΩ.

Since we were targeting a universal and flexible STIM/PM IC for simultaneous stimulation and recording, that would be compatible with recent work on the high input dynamic range front-end units, [40], [45], we have designed a multichannel implantable NM interface, Fig. 3.8 that houses both STIM/PM and SENSE IC.

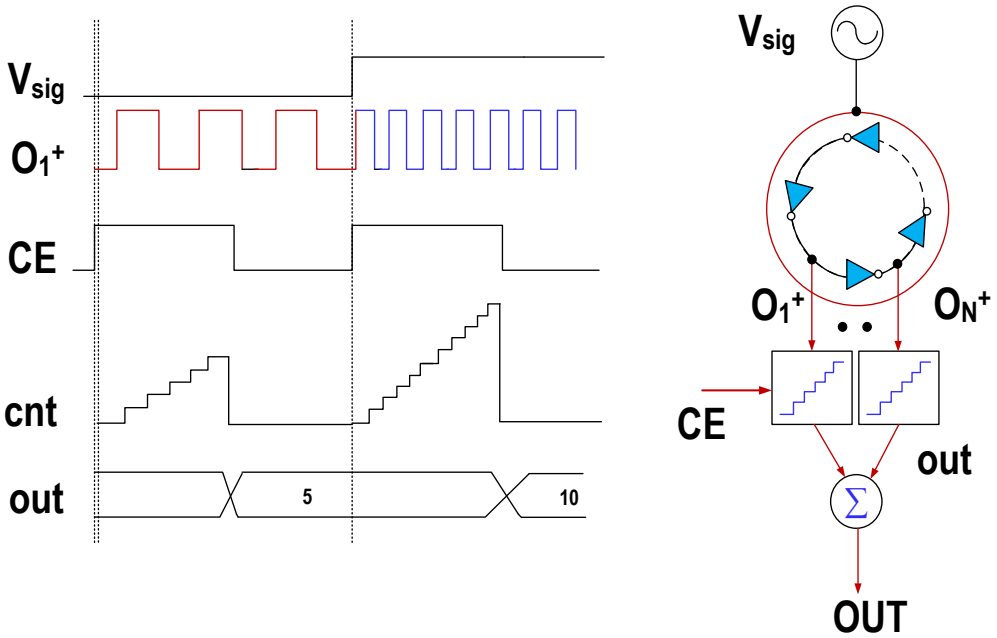


Fig. 3.14: Concept of the VCO-based ADC.

The sensing front-end IC features 32-channel (64-channel) VCO-based design with interleaved digital nonlinearity correction (NLC). In the conventional designs [37-38], area per sensing channel is dominated by the off-chip coupling capacitors, which dictates the overall size of the

implant and does not scale with CMOS technology. The VCO-based front-end, Fig. 3.14, processes LFP signals in the phase domain [40], thus allowing for low-noise linear digitization of voltage within a large input range on accessible electrodes, as controlled by channel selection MUXes in the *stim* IC. The digitized data is sent to the control module via a 3-wire SPI interface.

An implantable autonomous 2-chip system needs to have built-in clock generation for system timing. Therefore, a crystal oscillator driver (XO), high supply-rejection LDOs and associate BGRs are embedded in the *sense* IC, with LDO output voltages of 1.2V (VCO analog), 0.6V (VCO digital), 1.2V (NLC & system control) and 1.2V (XO). Power-on reset circuitry is designed, so that the 1.2V digital supply is ready first for system reset and configuration before the supplies to the VCO front-ends. The XO provides a 12MHz clock with lower jitter (9.8ps) which is sufficient for accurate (15 bits) signal capture.

3.4.3. Full-Fledged Power Management

To minimize the power consumption of a fully implantable biomedical device and to make the stimulator design compatible with the rest of the system, as an integrative part of the STIM chip, we proposed a full-fledged Implantable Power Management Unit (IPMU). IPMU is highly reconfigurable, can process and support different power transfers on-the-chip, depending on the application. STIM core and IPMU unit are made in HV technology, to accommodate large voltage swings at the electrodes, during stimulation. As a part of specification, we define several important targets which will be discussed in detail later: i) The IPMU should adapt the power delivery depending on the need at the load ii) Multiple modes of operation and smooth transition between

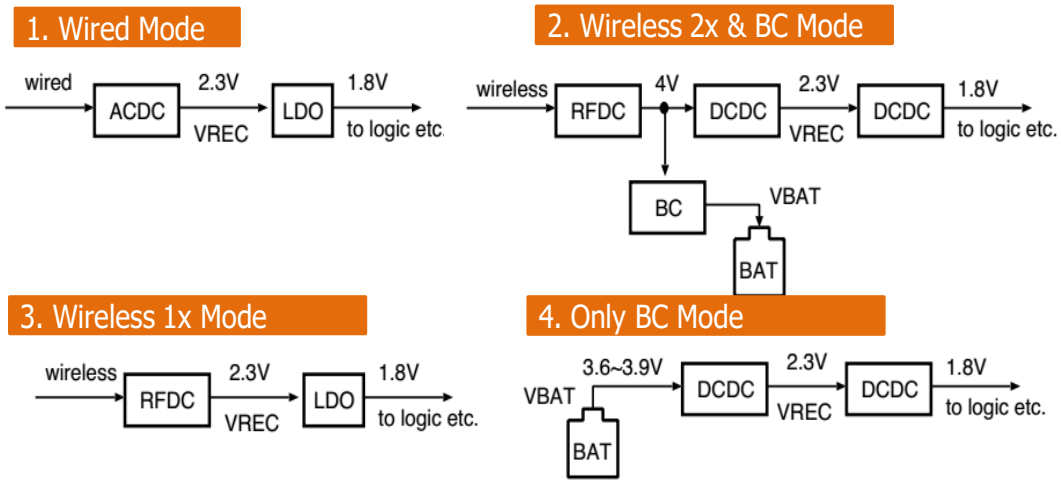


Fig. 3.15: Different Modes of Operation.

the modes iii) High power conversion efficiency (PCE>90%) iv) Small chip area and a few off-chip components to satisfy low cost and small volume (implantable interface) requirements.

IPMU supports 4 different modes of operation, Fig. 3.15., and it is controlled by 6 control signals, set through the DCU and user interface. The system can be configured to work in 1) Wired Mode – where the power is delivered to the implant, differentially through 2 wires; 2) Wireless 1X Mode- where the power is deliver through the near-field, inductive link; 3) Wireless 2X Mode – in which power is deliver through the inductive link, while simultaneously the rechargeable battery is charging and the implant is powered; 4) Battery Mode – where the whole implant is supplied from the battery. Figure 3.16 shows the complete block diagram of full-fledged IPMU. As the most power greedy blocks, efficient active rectifiers for both wired and wireless power transfer are imperative and they are covered in detail.

To improve the overall efficiency and maintain the efficacy of the NM interface of the inductively/wireline supplied stimulating medical devices, the efficiency of every stage in the power delivery path, such as the active rectifiers, high voltage generators, inductive link, etc., should be maximized. By adopting the system level approach and utilizing power-efficient circuit

techniques for both TX and RX side, we have designed IPMU that outperforms current state-of-the-art in flexibility and efficiency. Detail explanation follows.

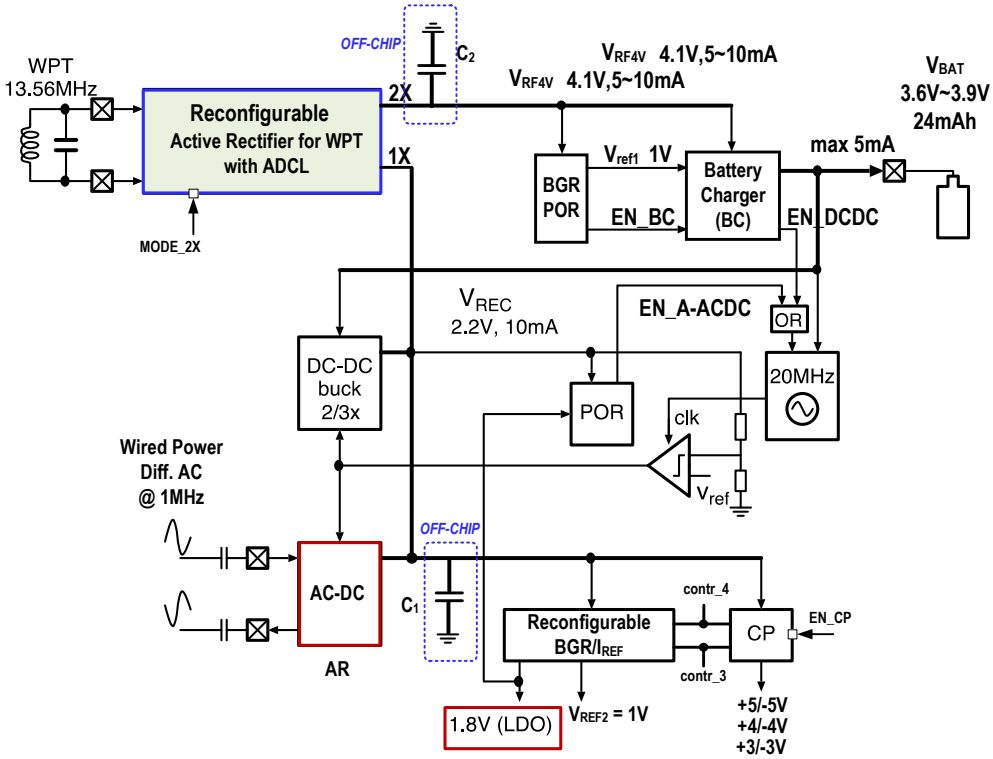


Fig. 3.16: Full-Fledged Power Management Unit.

3.4.3.1. Active Rectifier for Differential Wired Mode

During the operation in wired mode, the power management (PM) block is configured automatically and wireless power transfer & battery management units are turned-off, so there is no reverse current flow, Fig.3.17. The implant is powered by a differential AC input and active IMPU comprises of an active rectifier (AR-DC), scalable bandgap/reference current block (BGR/IR) and multiple-voltage generators for the various implant units. Two wires at the input carry sinusoidal signals shifted for 180 degrees to satisfy biomedical requirement, so that the net input voltage sum in the wires, is equal to zero at every moment in time. Peak-to-peak voltage is

6V at each wire. Duty-cycle control unit plays the role of a shunt regulator that adapts the power delivery to the load and also set the rectifier output voltage to the desired value – in our case 2.2V.

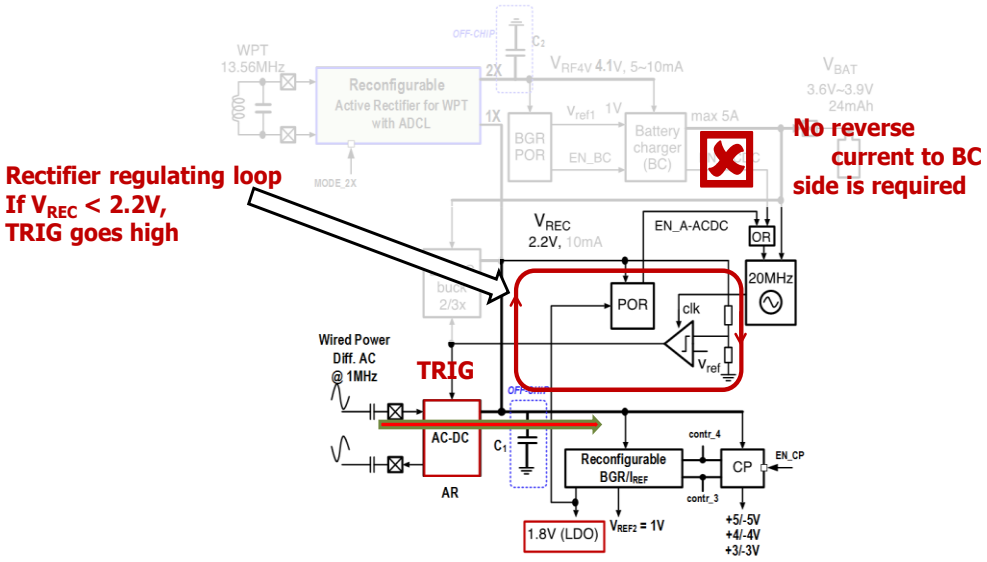


Fig. 3.17: Power Management Unit – Wired Mode.

Figure 3.18 depicts the adaptive, real-time on/off delay-compensated AR whose efficiency is improved and optimized for MHz-level inputs (PCE>80%). AR-DC also mitigates the substrate ringing and *di/dt* noise due to bondwire inductance. Output of the AR supplies the 1.8V LDO with high slew-rate and supply rejection. This LDO powers both ICs.

In the core of the active rectifier for differential wired power transfer is a full-bridge architecture. Every Active Diode (AD) inputs two control signals, which are necessary for transition from passive to active mode and for preventing excessive power dumping to the load. Also, since the targeted rectified voltage is 2.2V and the amplitude of the input signal is 3V, the source (drain) of power PMOS/NMOS transistors within the AD can reach 4.1V in the steady state. If the drivers inside the AD, are supplied from V_{REC} and gnd, turning off these diodes becomes problematic. To handle this, we proposed the active body biasing scheme (ABB), Fig. 3.18, to

mitigate any current leakage and prevent reverse current flow, which connects the bulk of every power transistor to the higher potential node. At the same time, the bulk node is used as a supply for the driver. The 1.1V offset shows up due to the isolation capacitances at the input of the active rectifier.

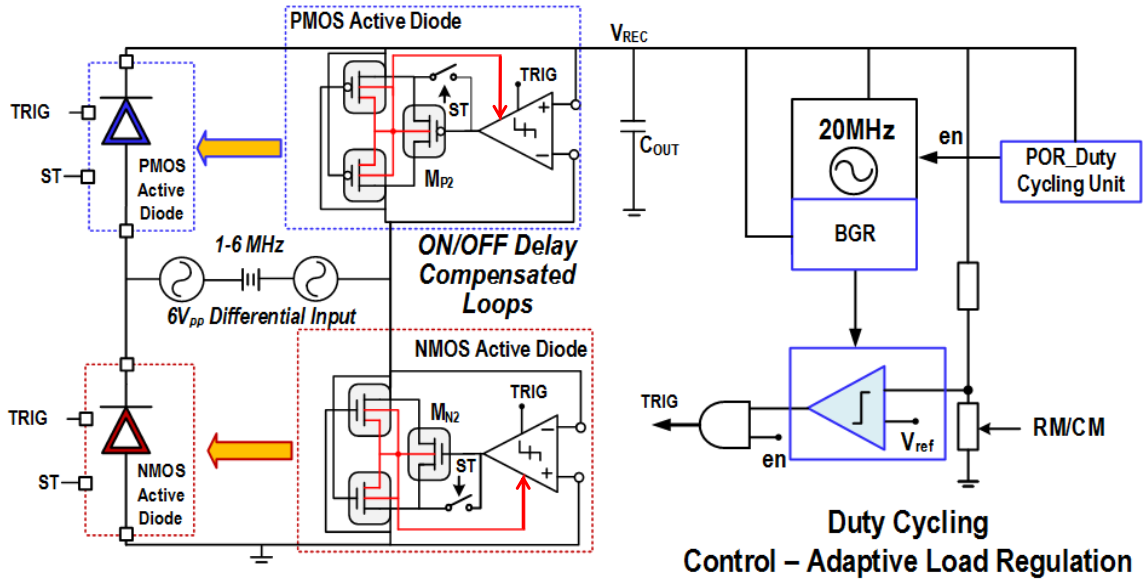


Fig. 3.18: Active Rectifier Scheme.

3.4.3.2. Wireless Power Link

For the near-skin implantable biomedical devices, wireless power transfer (WPT) is preferable power delivery option, which is usually based on the inductive near-field coupling due to its high efficiency. To be consistent with biomedical requirements, implantable applications usually use the frequencies from the ISM band, in which 13.56 MHz is the most commonly used carrier frequency. By employing WPT, scientists try to avoid bulky batteries, which is critical demand in volume-limited applications where form factor plays significant role. Since, our design targets a fully-implantable, miniaturized NM platform, WPT is an important task.

The Active Rectifier (AR) for the WPT is the most critical block regarding the power efficiency. AR is designed to operate in two different modes: 1) Regular Mode (1X) provides 2.2V

rectified voltage which is sufficient for further voltage regulation and 2) Charging Mode (Doubling Mode-2X) which provides 4.1V output; this voltage is used during the rechargeable battery charging. During 1X Mode, AR architecture is configured as a full-bridge rectifier, while during the 2X Mode it is configured as a voltage doubler – two half-wave rectifiers connected in series. Figure 3.19 shows enabled units in IPMU during the WPT in Charging Mode. The battery charger (BC) receives 4.1V at the input which is necessary for the operation. BC charges the battery with 5-10mA constant DC current. Parallelly, integrated buck DC-DC converter provides 2.2V that is needed for multiple LDOs and normal implant operation. Most of the circuitry that was active during the Wired Mode is disabled and the reverse current flow into tAC-DC rectifier is prevented.

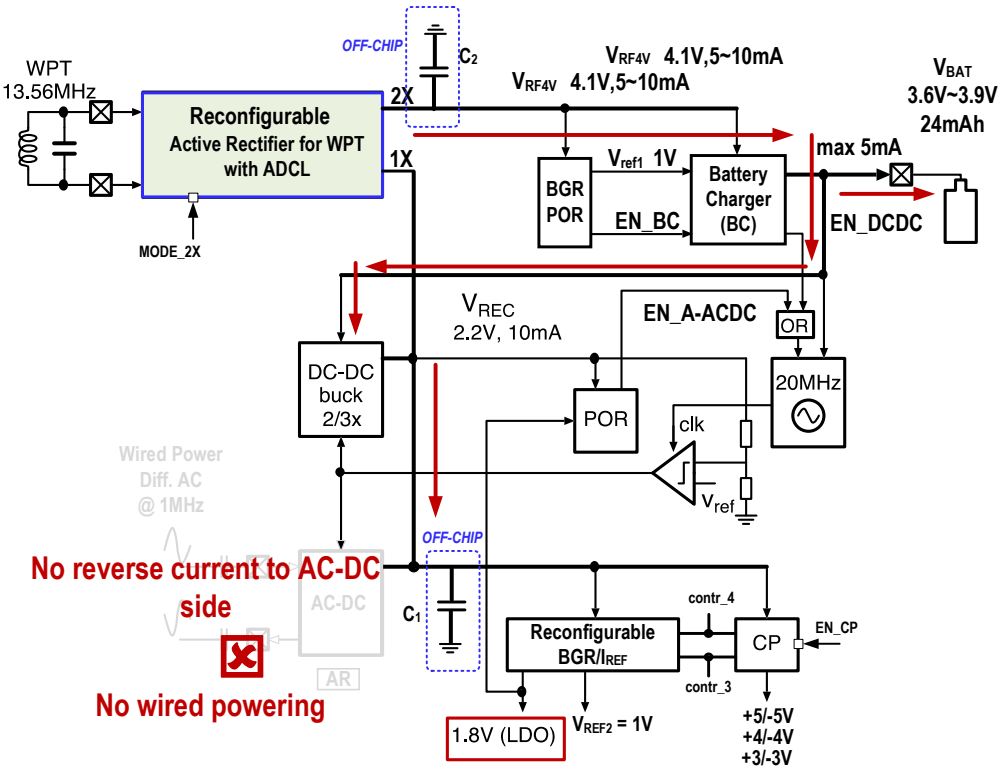


Fig. 3.19: Full-Fledged Power Management Unit in Charging Mode.

Active realization of the AR-WPT requires high power efficiency and Load Adaptation ability. During the implant functioning, the load requirement changes in time –from very light to very high. Also coupling variations significantly mitigate efficiency and make the output voltage

unstable. Most previous designs, [46]-[50], do not consider the excessive power dumping from the input (wireless link) to the output. Excessive power is either dumped to the DC-Limiter or absorbed by the body tissue. Usually, the simple DC-limiter circuit or clamping shunt regulator is employed to bound the V_{REC} value. This will cause significant current leakage and it will mitigate the overall end-to-end efficiency. Since the load requirement varies in the time, power efficient system would need a dedicated adaptive load control unit that will accommodate power flow in regards to the implant requirements.

We proposed a reconfigurable, PVT invariant and power efficient AR-WPT which includes Adaptive Load Control (ALC) unit that accommodates the power delivery. With the ALC unit, input power is controlled and excessive power at the output is significantly reduced. The efficiency of the rectifier is improved due to the new real-time offset controlled schemes that are implemented. With these two techniques, our system is able to perform >10x longer (battery life) compared to the state-of-the-art and has improved efficiency for a wide range of load currents.

During design of active rectifiers for WPT, that use 10's MHz as a carrier frequency, an important drawback has to be considered related to the propagation delays which are introduced by comparators (drivers). These drivers are driving the gates of the power transistors within the active diodes. To have small voltage drops across the active diodes, these power transistors have to be wide. The wider the transistors, their gate capacitance is bigger. To drive these capacitances at high speeds, the comparators require a buffer chain in the output stage. Naturally, there is a delay between changing the state at the comparator input and the buffer chain output. This delay causes power transistors to turn-on/turn-off either too late or too early. Both effects are detrimental

and affect the performance of the rectifier. Either they result in the reverse current flow that causes efficiency drop or the conduction time of diodes is reduced.

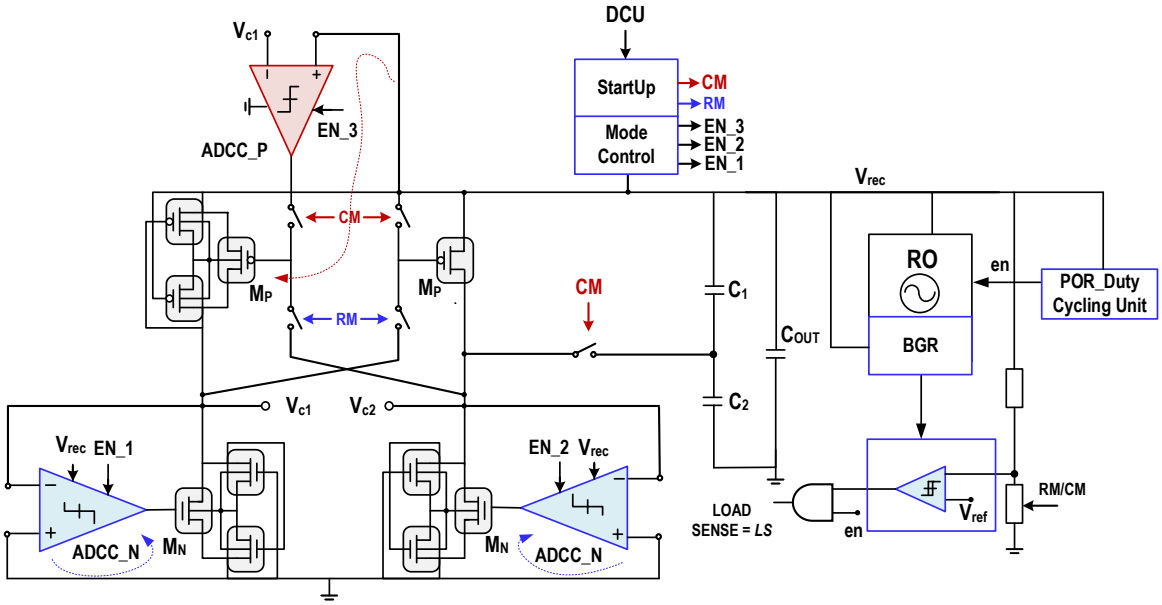


Fig. 3.20: Active Rectifier for WPT.

To keep power conversion efficiency high, several previous works proposed different techniques to compensate for the propagation delays, [46]-[48], [50]-[53]. Some of them introduced a constant offset at the comparator input using the unbalanced-bias scheme (asymmetrical input transistors) to compensate for the OFF delay, [46]. This just partially solve the problem, since the compensation of ON delay is skipped. Some require off-chip offset calibration. A switched offset biasing scheme, [48], was proposed to explicitly control the reverse bias current. Ghovanloo in [47] used an off-chip calibration method. Problems with these approaches are that they are not flexible due to the various reasons (PVT variations, transistor mismatch, offset, etc..). These schemes are usually optimized for the particular operational condition, and their design procedure is complicated. Recently, in [52], the authors explained a near-optimum approach, that does not incorporate ALC unit and PMOS active diode calibration. Without ALC – reaching a steady state and having near-optimum condition is a real challenge.

We proposed the simple architecture that incorporates the adaptive, real-time ON/OFF calibration scheme for both types of active diodes (PMOS, NMOS) that autonomously generates the offset currents for the comparators and is immune to PVT and circuit mismatch. Inspired by the work in [51], Fig. 3.20 shows the overall AR-WPT architecture capable of working in Regular (V_{REC}=2.2V) and Charging (V_{REC}=4.1V) Mode with ALC Unit with complete Calibration Schemes that do not need any tuning.

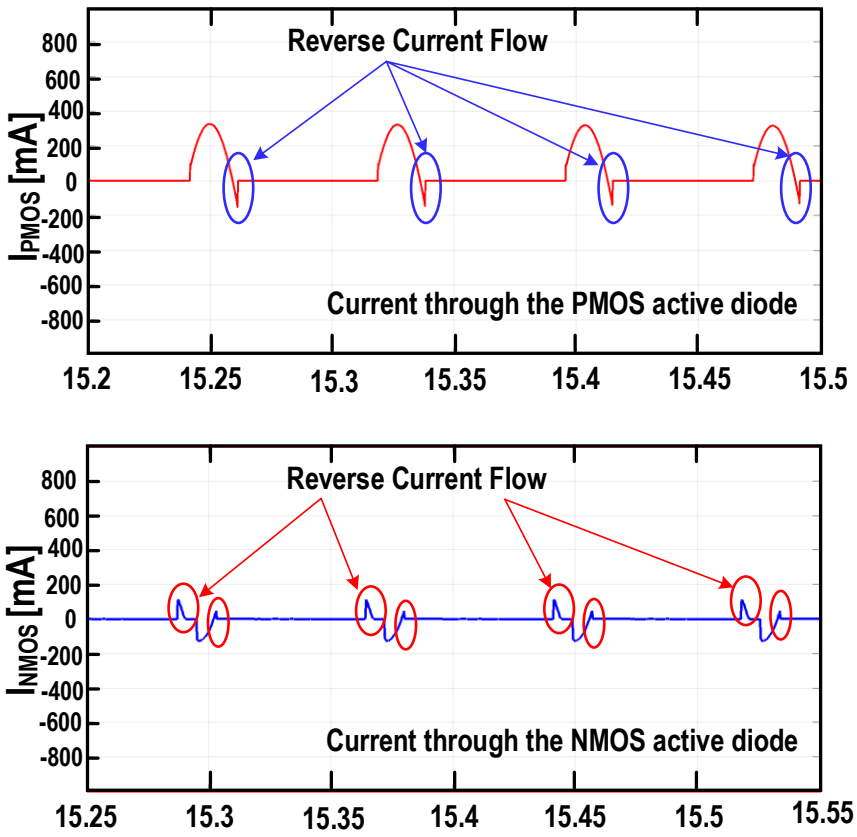


Fig. 3.21: Current through the active diodes without calibration schemes implemented.

AR-WPT consists of five power switches, three adaptive delay compensated comparators (two of them for driving the N-type diode and one for driving the P-type diode), duty-cycling control unit for output regulation along with startup and mode control units. Depending on the states of these 5 switches, the AR-WPT can be configured to work in:

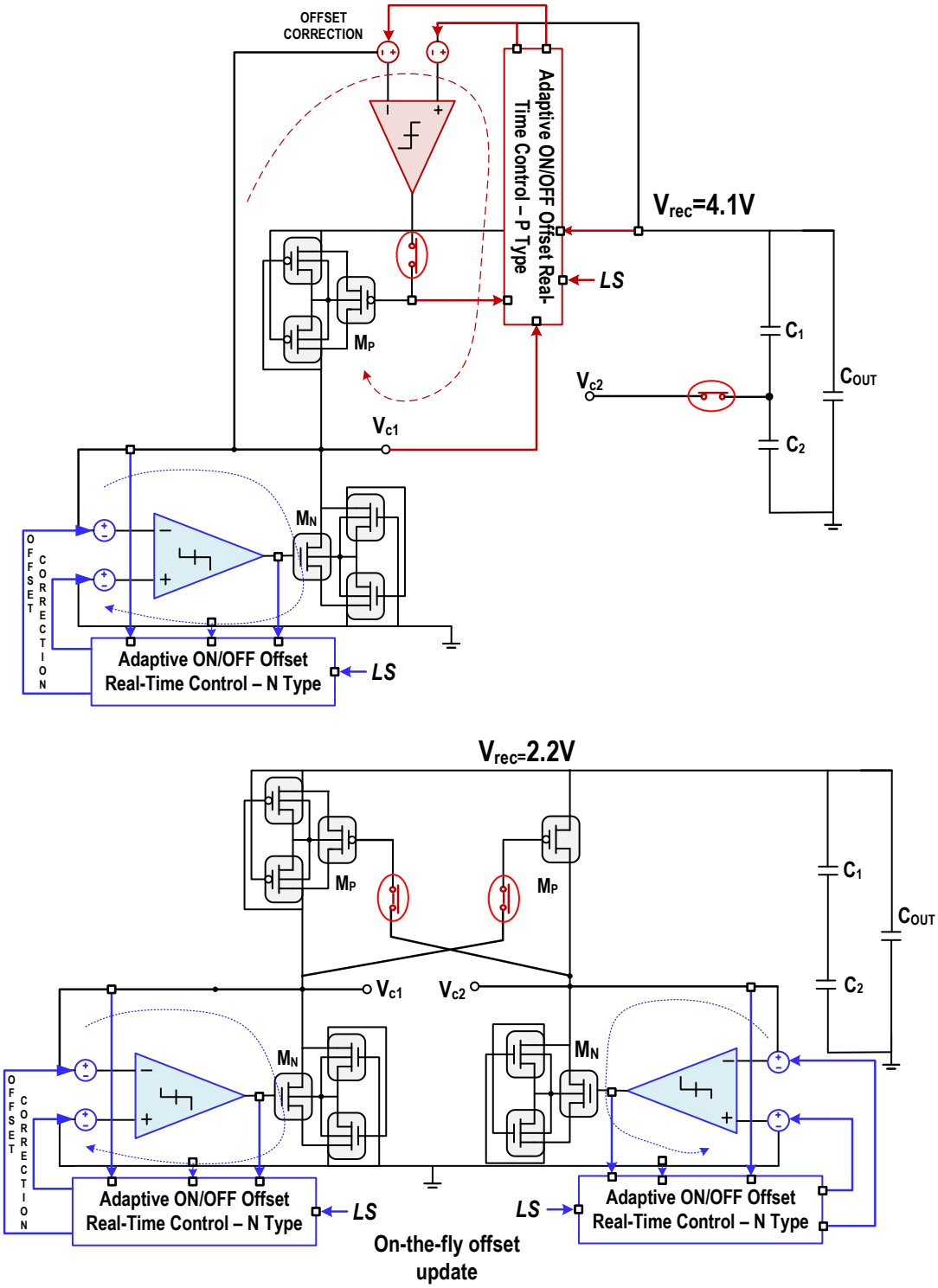


Fig. 3.22: Calibration Feedback Loops for 2X/1X Mode.

Regular Mode, where RM switches are turned-on and MP transistors are cross-connected with the gate of one connected to the drain of other. N-type Active Diodes are enabled, while P-type

diode is disabled;

Charging Mode, where RM switches are turned-off and CM-switches are turned-on. CMP1 and CMP3 are enabled. In the steady state, voltage V_{ac2} (one side of secondary coil) is clamped at $V_{REC}/2$, so MP2 is reversed bias and consequently turned-off.

Red and blue lines show the paths where the delays are introduced by the comparators. The impact of these delays is multifold; Fig. 3.21 depicts the impacts of ON and OFF delays for both type of active diodes (AD) without calibration schemes implemented.

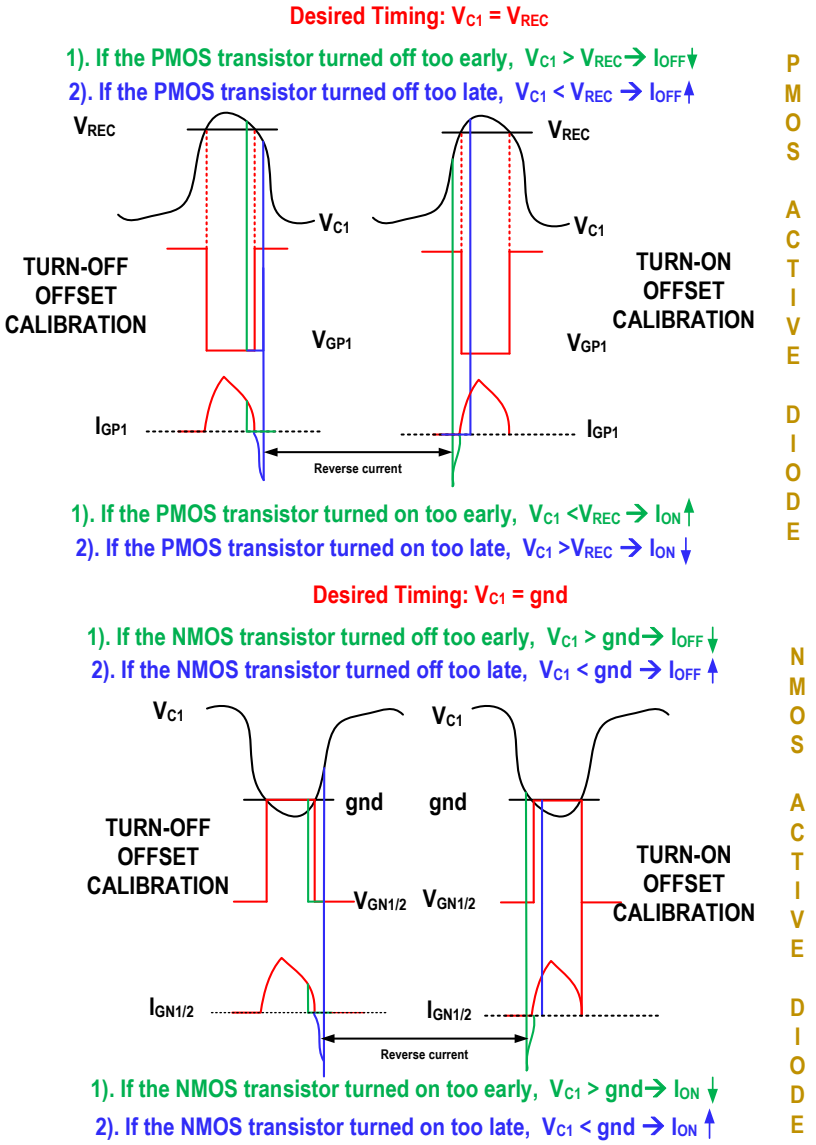


Fig. 3.23: Calibration criteria for active diodes.

None of the previous works, [46]-[53], simultaneously dealt with the real-time offset calibration for reconfigurable AR-WPT and implemented ALC unit that eliminates a lossy voltage limiter. Figure 3.22 shows the block diagrams of the proposed, near-optimum active rectifier in 1X/2X mode with negative feedback loops for a real-time delay calibration for both N-type and P-type turn-on & turn-off delay compensation. These feedbacks are responsible for adaptive generation of the ON/OFF offset currents to compensate the switch delays. The signals V_{C1}, V_{REC} and V_{GP} are used as an input for the P-type calibration scheme, since they contain the information whether the P-type active diode turned-on/off too early/late or if it is close to the optimum timing. Similarly, the signals gnd, V_{C1} and V_{GN} are used for the N-type calibration scheme and derivation of the calibration criteria.

Calibration criteria for both type of active diodes is depicted in Fig. 3.23. Let's consider, optimum timing for the P-type active diode. A similar analogy can be made for the N-type of active diode with different desired timing criteria. If PMOS power transistor is turned-off too early (green line), conduction time is reduced, which means that $V_{C1}>V_{REC}$. To fix this, in the next cycle, more offset current through the off-branches in comparator has to be added. In the analog manner, if PMOS power transistor is turned-off too late (blue line), a reverse current flow will be the result. To reduce this in the next cycle, the offset current through the off-branches should be decreased. Deriving the conclusions for the turn-on offset calibration is done in a similar way. So, the offset is updated in every cycle and within several cycles the desired timing condition is reached. In the steady state, if the input signals V_{C1} and V_{REC} are sampled on the rising and the falling edge of the V_{GP}, the sampled values should be equal. That implies that delays are fully-compensated. For the N-type of active diode, if the input signal V_{C1} is sampled on the rising and the falling edge of V_{GN}, the sampled value should be equal to gnd.

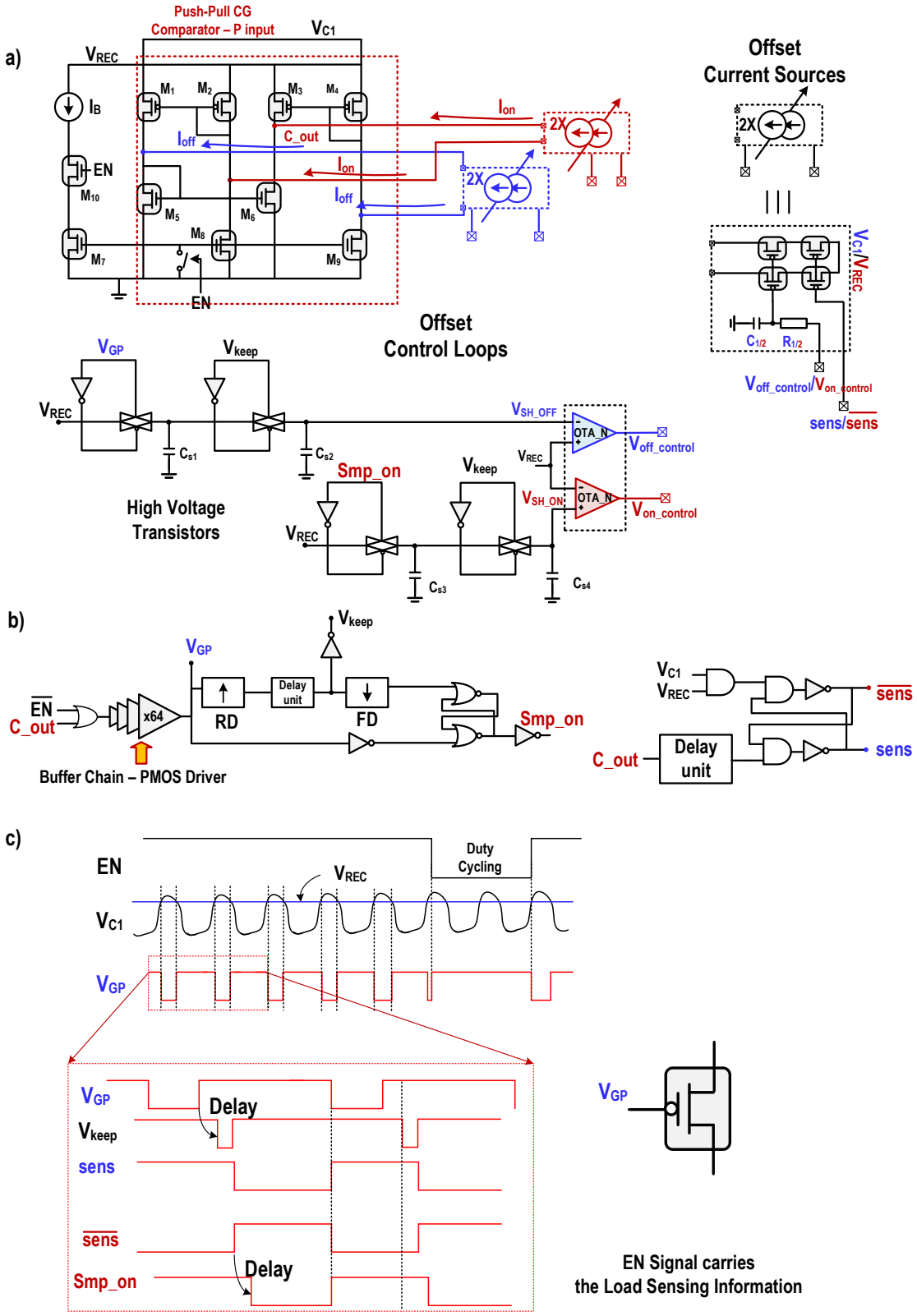


Fig. 3.24: a) Compensation scheme for P-type active diode b) Gate driving circuits-control signals c) Timing diagrams for the control signals.

Detail circuitry of the proposed real-time offset compensation scheme for the P-type active diode is shown in Fig. 3.24. N-type scheme is represented by the dual circuit and analysis is similar. High voltage transistors are used in the implementation, since AR-WPT supports doubling mode and the range of voltages goes up to 5V.

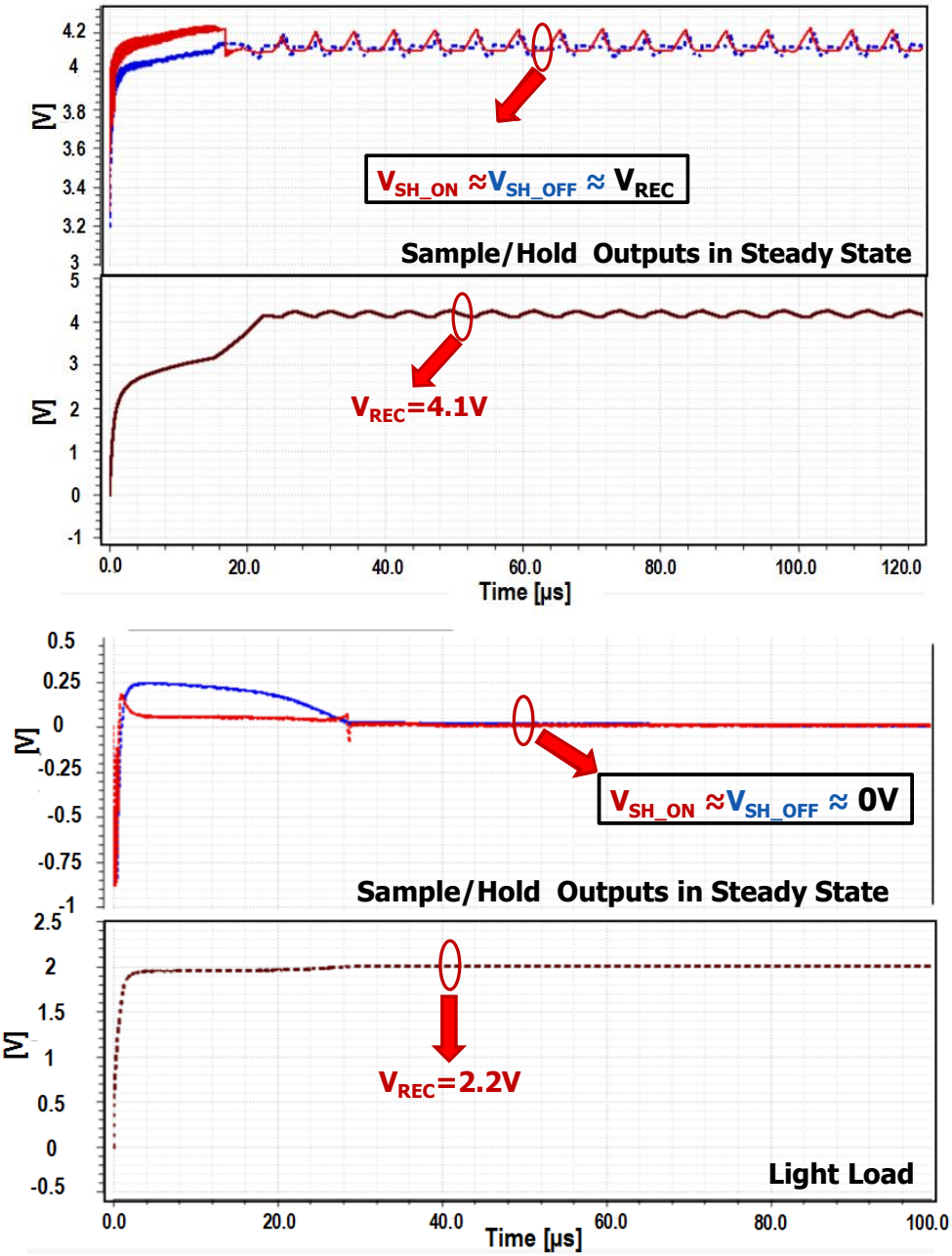


Fig. 3.25: Near- Optimum Steady-state for Charging (2X) and Regular Mode(1X).

In the core of the calibration scheme is the push-pull common gate comparator with the P-input transistors (M_1-M_{10}). Two negative feedback loops are added to adaptively generate ON/OFF offset currents. Every feedback loop comprises of an offset current source, feedback amplifier and the sampling circuitry that plays the role in the ON/OFF timing adjustment. Let us consider the ON-delay compensation path: The control logic generates signals sens, $\overline{\text{sens}}$, V_{KEEP} and Smp_on. On the rising edge of Smp_on, input voltage V_{C1} is sampled on C_{s1}. During the V_{keep} that voltage value is passed onto C_{s2} and the feedback amplifier OTA_N compares the sampled value with V_{REC} until the next falling edge of Smp_on, [53]. We have two possible scenarios: 1) If sampled voltage is smaller than V_{REC}, OTA_N will drive $V_{on\_control}$ to the lower value and more offset current I_{ON} is pushed through the stacked PMOS current source. Consequently, the PMOS diode (switch) will turn on later compared to the previous cycle. 2) If sampled voltage is higher than V_{REC}, OTA_N will drive $V_{on\_control}$ to the higher value and less offset current I_{ON} is pushed through the stacked PMOS current source. In this scenario, the PMOS diode (switch) will turn on earlier compared to the previous cycle, and as a result, after several 10's of cycles the system would reach a steady state; V_{SH_ON} should be equal or close to V_{REC} indicating the desired optimal timing. OFF-compensation path is realized and analyzed in the similar manner – in steady state $V_{SH\_OFF} \approx V_{REC}$. Feedback amplifiers, OTA_N are realized as the low power folded cascode amplifiers with N-type input transistors and GBW <0.5MHz.

To ensure no oscillation and smooth transition between transistor ON/OFF states, RC time delays are added, [48]-[52]. These delays behave also as a low-pass filters; they remove high-frequency components in the offset currents.

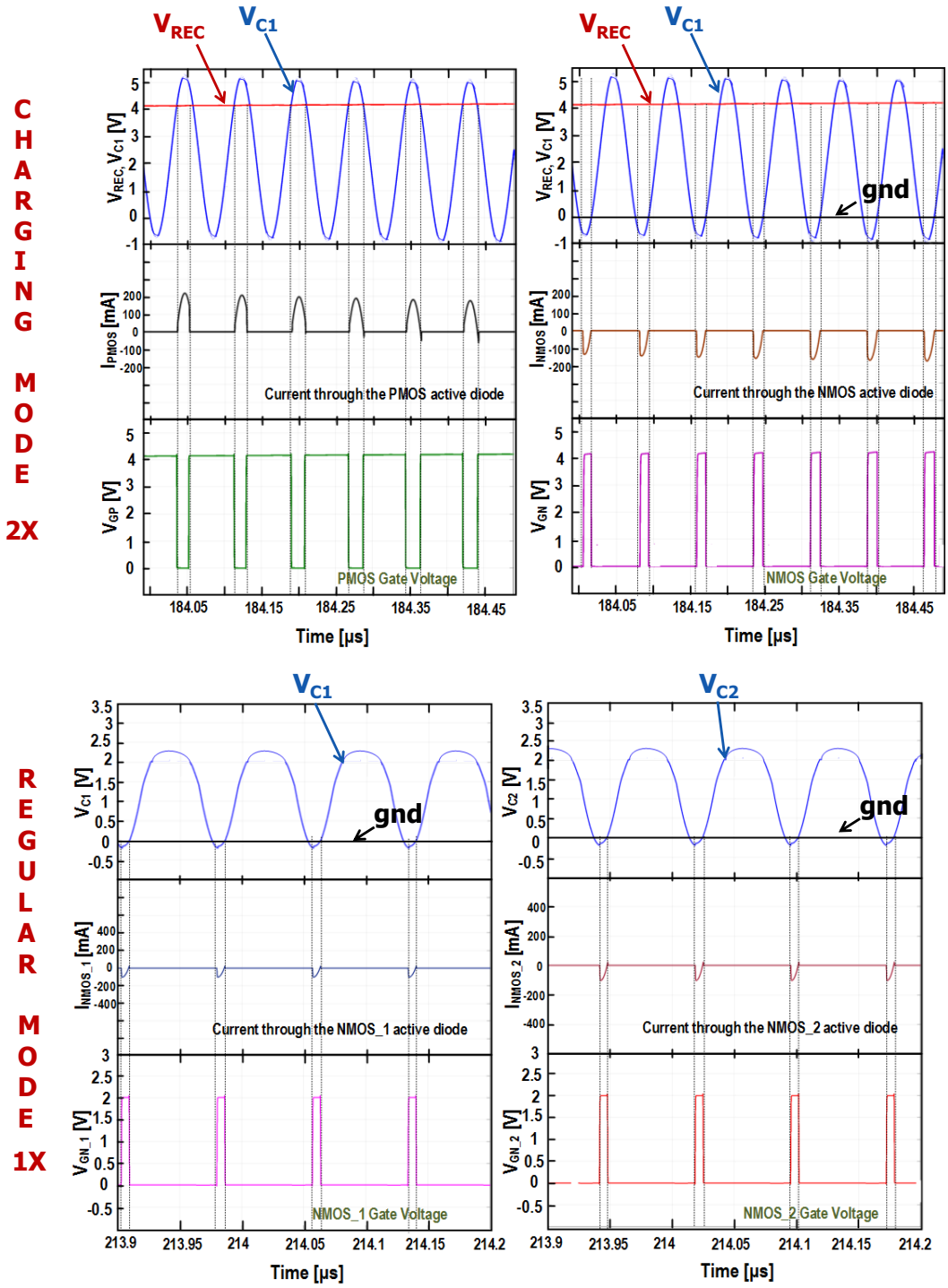


Fig. 3.26: Relevant Waveforms for Active Diodes with delay compensation implemented.

Big advantage of the real-time adaptive compensation scheme is its immunity to the process mismatch and PVT variations. Figure 3.25 verifies the near-optimum steady-state operation for

both PMOS and NMOS type of diodes. The outputs of sampling circuit V_{SH_ON} and V_{SH_OFF} follow the rectified voltage V_{REC} and gnd for 2X and 1X mode respectively. The relevant waveforms for both modes of operations are shown in Fig. 3.26. This demonstrates the effectiveness of the proposed technique – with adaptive ON/OFF compensation scheme implemented, the system reaches the desired optimum timing and the effect of reverse current and reduced conduction time (which affect the efficiency) are eliminated or significantly mitigated.

Since the load requirement varies over time, implementation of ALC unit is necessary. Our ALC unit with Hysteretic Comparator (HC) is shown in Fig. 3.27. The hysteresis is added to the two-stage amplifier by employing a resistor of fixed value together with the steering (current) circuit. This results in the amplifier's negative input terminal shift by the value proportional to the product of the resistor and hysteresis bias current. Hysteresis bias current controls the hysteresis properties (window, slope, etc.). When the output voltage V_{REC} reaches the desired value, the comparator in the feedback will change the value of the control signal LS. As a result, all diodes in the AR-WPT would be turned-off and power transfer from input to output is suspended. If we keep the hysteresis window at 100mV, the output voltage V_{REC} will fluctuate within 100mV window around the desired value. If we used a regular comparator, we would introduce the hard switching and observe a sharp voltage ringing at the output. This can make circuit intrinsically unstable. Consequently, the calibration mechanism would not establish the steady state in the rectifier, since it needs dozens of cycles. With hysteretic comparator, the circuit enters periodically into shut-down (duty cycling) modes, and still have time to calibrate ON/OFF delays in AR-WPT (during LS =0 periods). Load Adaptive signal LS is coupled into comparator enable signals (EN_1-EN_3). HC dynamically keeps the V_{REC} at the desired level by toggling LS which consequently

leads to energy preservation, improving the AR-WPT efficiency and reducing current leakage through the ALC unit.

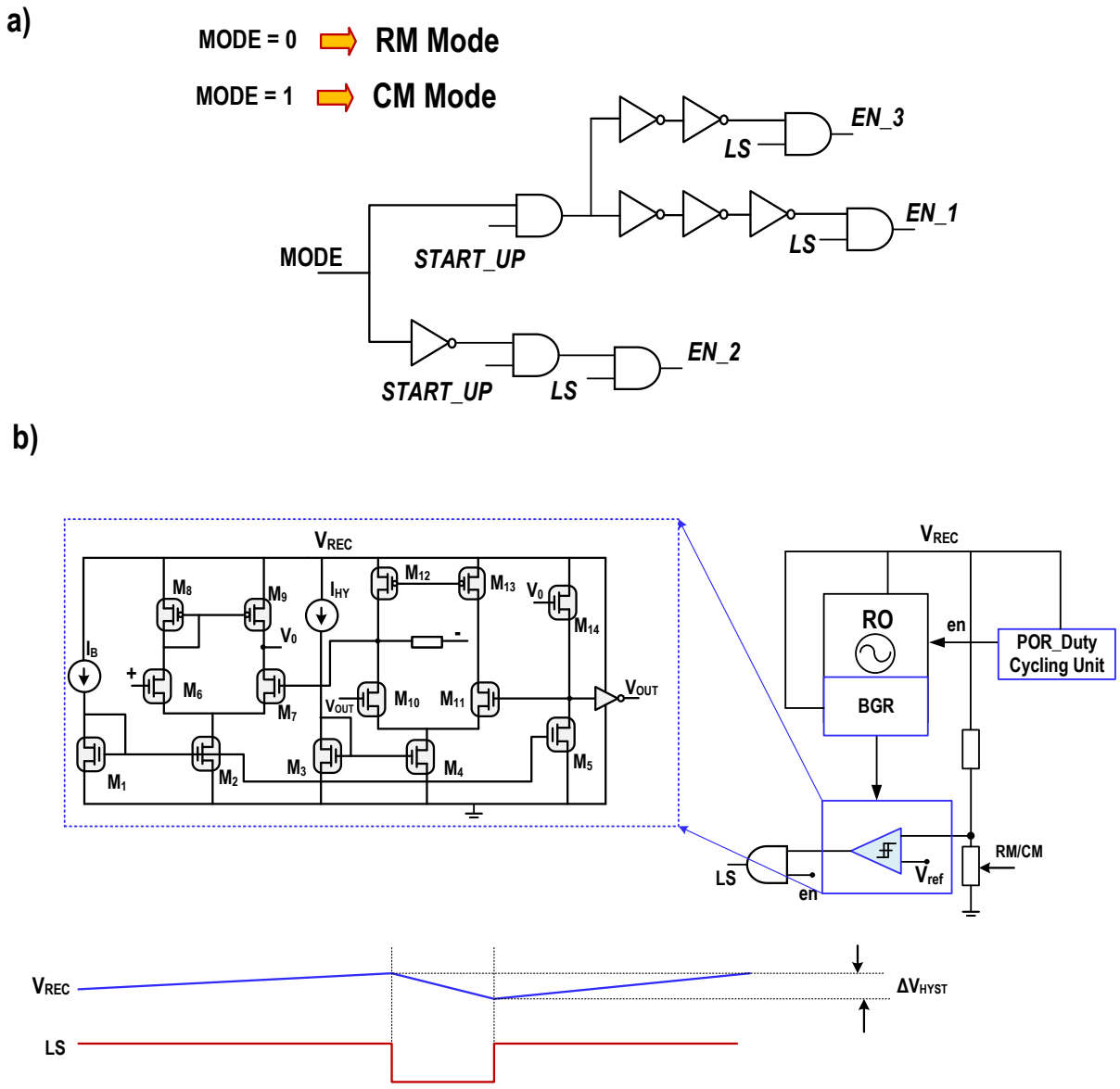


Fig. 3.27: a) Control Logic b) Adaptive Load Control – Shunt Regulator with Hysteretic Comparator.

So far, we have focused on the RX local wireless voltage rectification and regulation. We have shown the circuitry that reduces complexity, requires minimal number of off-chip components, and leads to the improved efficiency. However, the complete wireless power system also requires TX independent IC. There are several works done so far that demonstrated TX-RX wireless inductive link for biomedical applications. In [51], [54], authors proposed backscattering, where TX is driven by the RX as the impedance changes on the receiver side during implant operation. This design requires an extra off-chip coil. Also, [55]-[56] requires TX-RX data link, so that TX can receive feedback information from the RX unit, that contains the sensed loading at the implant side. These systems usually need microcontrollers, pulse generators and other off-chip units that are power hungry. Most previous works, [49]-[51], [53]-[56], use the class D/E power amplifiers on the TX side, that are switching at the carrier frequency and driving the inductive link. These architectures are not suitable for the implant-scale biomedical applications.

We have recently proposed in [57], a new wireless power link architecture that is immune to distance variation and can sense the implant "needs" without explicit feedback from the RX unit. The TX unit together with the link, self-regulates the power delivery to meet implant requirements.

The basic idea is that by employing simple cross-coupled oscillator architecture with automatic amplitude control (AAC), the system can self-tune to one of two stable frequencies, [57]. It can be shown that operation in one of these two frequencies would lead to a constant ratio between the source and load voltages $V_L/V_S=\sqrt{\frac{L_2}{L_1}}$, thus making it independent of coupling coefficient and load. This means that a wireless power system explained in [57] will hold the voltage amplitude at remote load constant as load resistance varies.

3.4.3.3. Battery Charging Unit

Battery charging (BC) unit requires 4.1V at the input and charges (5-10mA loading current) a Li-ion battery-pack system with a constant current. Li-ion battery requires 3.6V-3.9V for normal operation. Integrated buck dc-dc converter steps down the output voltage from charging unit to 2.2V and is able to provide up to 10mA of output current. Motivated by the work in [58], we have implemented a built-in resistance compensator technique that improves the speed of battery charging. This technique dynamically estimates the external resistance of the battery system and extends the phase of the constant-current stage. As shown in [58], a smooth transition method ensures stable transition from the Constant-Current to the Constant-Voltage stage for the BC. In the core of the BC, we have LDO-based circuit accompanied with the built-in resistance compensator and the Smooth Control Circuit and that includes Reference Shift Circuit, External Resistance Detector and Reference Voltage Switch.

3.4.3.4. Adaptive High Voltage Generator

As we stated before, power efficient stimulation in energy limited applications is an imperative. Stimulators require high voltage and high power dual supplies, to support a wide range of stimulations currents and differential stimulations. Fully integrated High Voltage Generators (HVG) in multi-voltage system design, with high power efficiency is targeted. Most prior works, [59-61], rely on bulky external passives (capacitors, inductors, etc.).

We have designed an adaptive closed-loop 4-stage charge pumps, Fig. 3.28, with leakage reduction scheme, that can provide ±7.5V (±5V-v2) supply rails (V_{DD}/V_{SS}) - max 3.5x voltage

conversion ratio. Integrated charge pump efficiency is improved as compared to [62-63] by using the modified Pellicone's cross-coupled cell, [65], for the negative and Favrat's cell, Fig.3.29a, for the positive pumping stage. Each Favrat's cell uses a small auxiliary charge pump structure for biasing PMOS devices. The generation of negative voltages on IC is possible due to the triple-well process, Fig.3.29b. The structure is very similar to the one demonstrated in [66], except the PMOS and NMOS switches exchanged positions and the bulk of PMOS is connected to gnd in order to

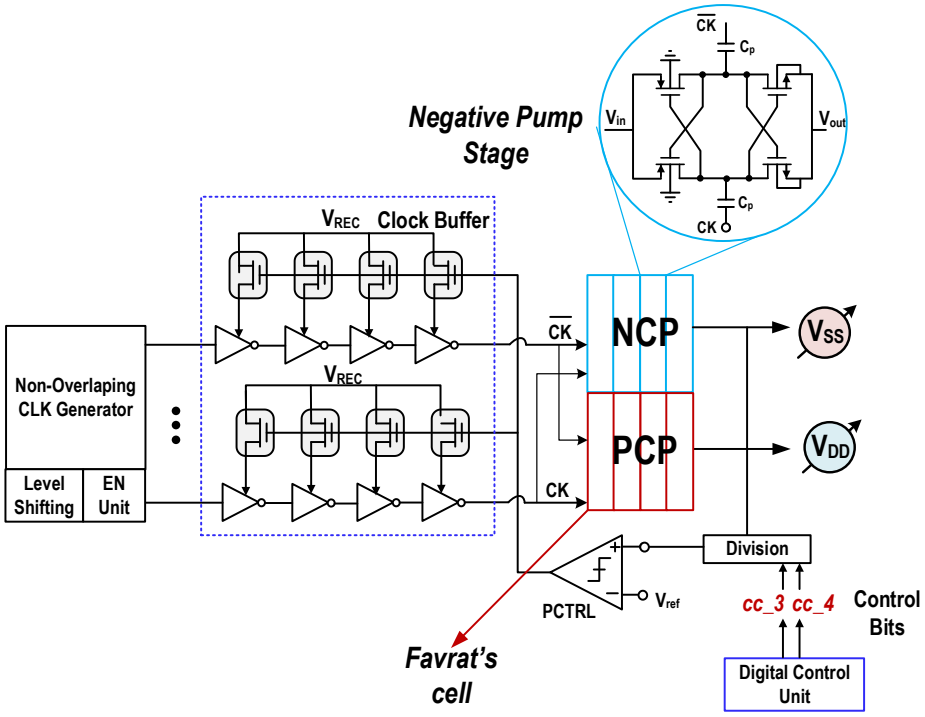


Fig. 3.28. High voltage Generator for STIM Chip.

prevent latch-up. The settling time of NVG is very fast due to the high frequency of operation (fs=20MHz) and full integration of charge pumps. Their performance is optimized to provide up to 1mA of constant DC load current while maintaining a high efficiency.

To further increase energy savings, an efficient high voltage V_{DD}/V_{SS} scaling is employed. DCU can configure control bits cc_3, cc_4 to accommodate the outputs V_{DD} and V_{SS} of HVG at the optimal value which is sufficient for power efficient stimulation.

In the core of the HVG scheme there are multiple pumping stages with non-overlapping clock generators and feedback loop as shown in Fig. 3.28. The feedback loop consists of the clock buffer and comparator that provides adaptive control signal. This feedback decides if the output voltage (V_{DD}/V_{SS}) of the HVG reaches the desired value. When that happens, the comparator outputs the high signal (1.8V), and the charge pump will stop pumping by disabling the clock buffer. Until the output voltage does not reach the targeted value, the output of the comparator is kept low. The main sources of efficiency drop lie in the timing mismatch and in overlapped clock signals and would cause the reverse current flow. To prevent the reversion losses, we have designed a dedicated control scheme with HV level shifting unit, that ensure FETs (switches) are not ON at the same time. This will enhance the power efficiency, and by adding the filtered capacitors at the input of the comparator, the output voltage accuracy is improved.

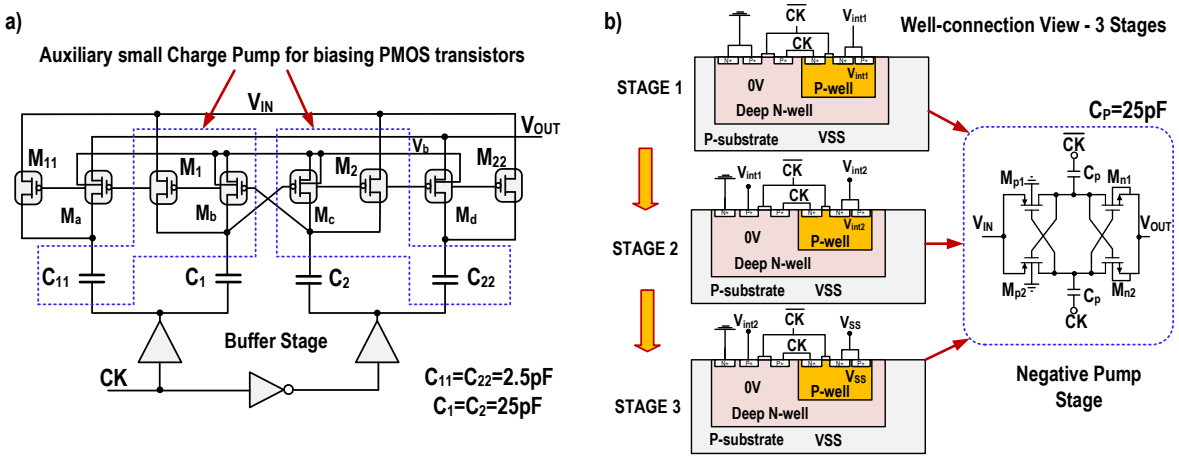


Fig. 3.29: a) Favrat cell – Positive Pump Stage; b) Negative Voltage Generation, [66].

In this topology, the voltage drop across the stage is roughly equal to $2V_{DS}$, while the output voltage can be approximated with

$$V_{OUT} \approx V_{IN} + N\left(\Delta V - \frac{I_{OUT}}{f_s C_f}\right), \tag{3.3}$$

where $\Delta V \approx V_{sup} \frac{C_f}{C_f+C_{par}}$ and C_f is flying capacitor while f_s denotes the switching frequency. For integrated implementation (flying capacitors on-chip) and mA output current capability to achieve a high voltage gain and high power efficiency, the switching frequency has to be in 10's of MHz range. Power efficiencies for both positive and negative closed-loop charge pumps are shown in Fig. 3.30.

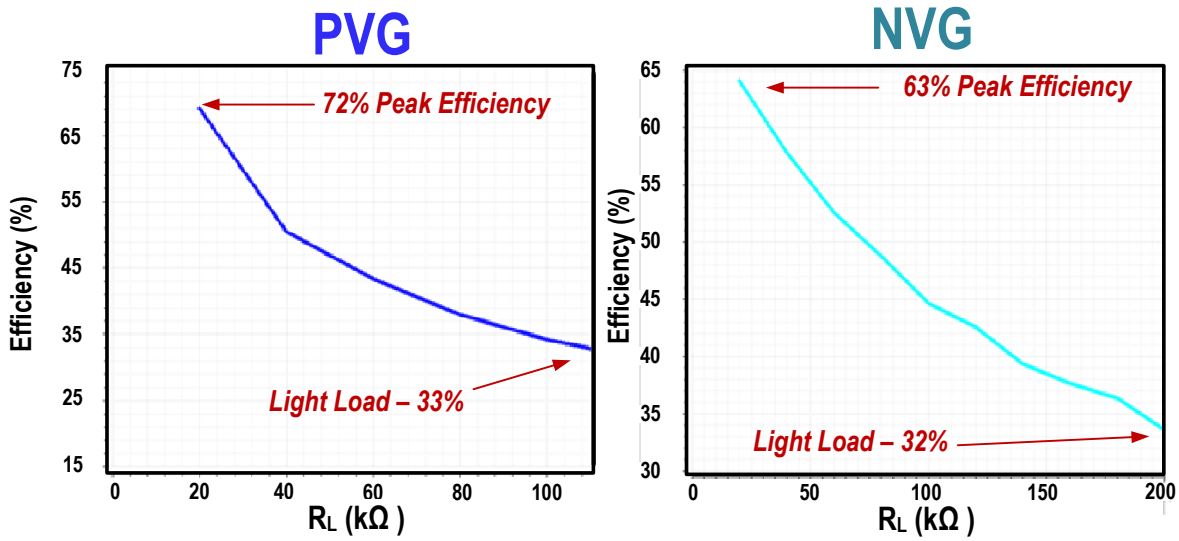


Fig. 3.30: High Voltage Generator – Simulated efficiencies in HV180nm.

Another constraint comes from the stimulator requirement. During the active stimulation, engines can drain 10's of mA of current from the high voltage supplies V_{DD}/V_{SS}. HVG designed on the chip, are not able to provide that amount of current instantly. Logically, the only solution is to have a high value (10µF -20µF) storage capacitances at the output of the HVG. The benefits of using these high value capacitances is twofold. First, the output voltage ripple is proportional to $V_{ripple} = \frac{I_{LOAD}}{f_s C_{OUT}}$, hence it will be mitigated. Secondly, the voltage drop during the stimulation would be in order of several 10's of mV. Otherwise, huge voltage drop would introduce a stimulator malfunction. Also, keeping the V_{DD}/V_{SS} within the safe range is needed for correct BGR and current mirror operation.

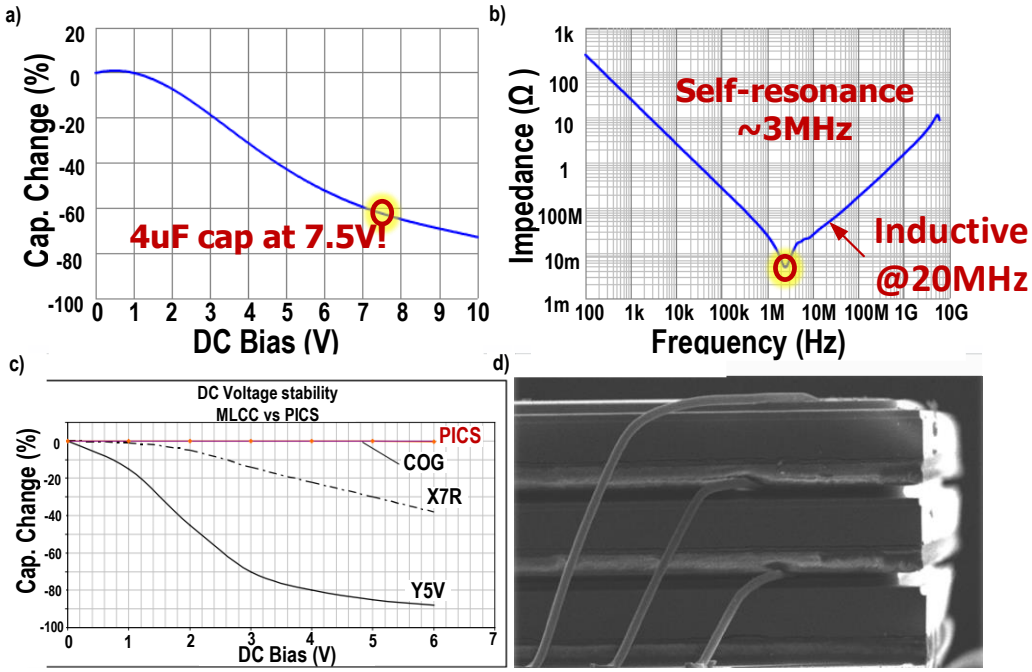


Fig. 3.31: a) Ceramic SMD Capacitance Drop with DC Voltage; b) Self-resonance frequency of ceramic capacitors; c) Comparison between ceramic SMD and Integrated Passive Capacitances; d) Stacked IPDIA capacitors as a compact energy source, [67].

Apart from the necessity and benefits that introduction of external capacitances brings into design, there are a few challenges that have to be considered. Medical-grade, ceramic SMD capacitors show capacitance degradation as the DC voltage across increases, Fig. 3.31a. Also, their self-resonance frequency, for the capacitance values in µF order, is up to a few MHz at the best case scenario. Since switching frequency of our charge pumps is 20MHz, clearly the external ceramic capacitors would clearly show inductive property at that frequency, Fig. 3.31b. As a consequence, the output charge pump ripple, that can be expressed as

$$V_{ripple} \approx L\frac{di}{dt} + ESR * i, \tag{3.4}$$

where ESR = R_{pcb} +R_{chip_wire}+R_{cap} denote the serial accumulated resistance, can reach several 100's of mV. Such a big ripple on supplies, is unacceptable and can cause the stability issues. An elegant solution for this problem is to use Integrated Passive Device (IPD) devices, [67]. These devices

show no capacitance degradation over the DC voltage stress. Also, IPD's negligible serial inductance introduces a very small voltage ripple. Figure 3.31c shows the comparison between the IPD and the ceramic external capacitors. Another factor that plays an important role in volume-limited, miniaturized applications, is the size of external components. IPD is a right choice when it comes to 3D passive integration as a top priority. Thickness of these capacitors can be between 80µm and 100µm while their capacitance density is 4 µF/mm^3. An example of a cubic stack, which can be stacked further on one of the ICs is shown in Figure 3.31d. Stacking of the integrated storage capacitors, will create more space on the assembly board and will reduce the overall size of implant-scale medical device.

3.5. Simulation and Measurement Results

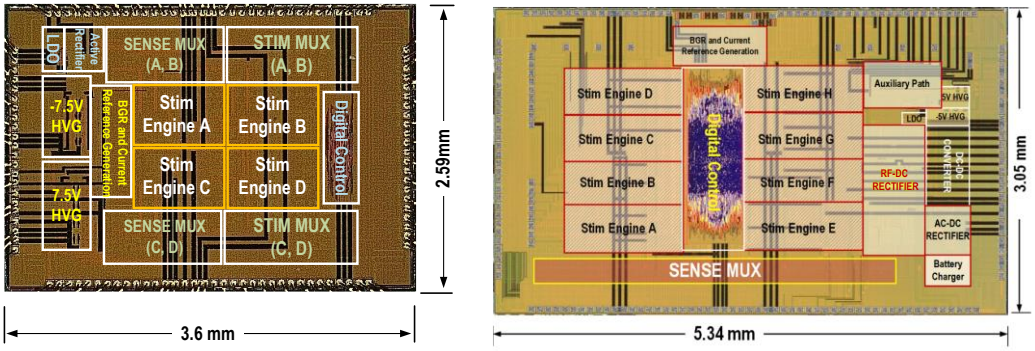


Fig. 3.32: Die Micrographs.

To demonstrate the functionality and performances of our system, we have designed two different STIM/PM ICs (Fig. 3.32) – the first IC has 4 Stimulation Engines (SE) and can drive 32 stimulation cites with V_{DD}/V_{SS} absolute maximum set to 7.5V/-7.5V. In the second version, the stimulator block includes 8 SE that can be individually programmed for monopolar/ differential stimulation. Stimulation current, per engine, covers the range from 20uA to 5.1mA with 20uA step. Programmability includes pulse shape, phase duration, full spatial selection, power control, etc. Engines are designed to be electrode agnostic.

High Voltage STIM/SENSE switching matrices are designed for 64 electrodes and STIM matrix provides a complex spatial resolution. The *stim* IC is integrated in HV-180nm CMOS to support a large voltage, while the *sense* IC is implemented in 40nm CMOS technology for reduced area and power of digital circuits. V_{DD}/V_{SS} are designed to be programmable with the absolute maximum to 5V/-5V.

Measurements are conducted in two phases as depicted in Fig. 3.33. To evaluate the performances of the STIM/PM IC, we have designed the STIM test-bench board, Fig. 3.33a. The measurement setup also includes TX board and wireless inductive link. TX board houses the transmitter IC with AAC that is explained in [57]. The PC is running a control software which sends the STIM and PM control parameters through the FPGA board towards the IC. This setup is primarily used to evaluate the performances of our integrated PM unit, specifically – the reconfigurable ON/OFF delay compensated active rectifier during the operation in 1X/2X mode.

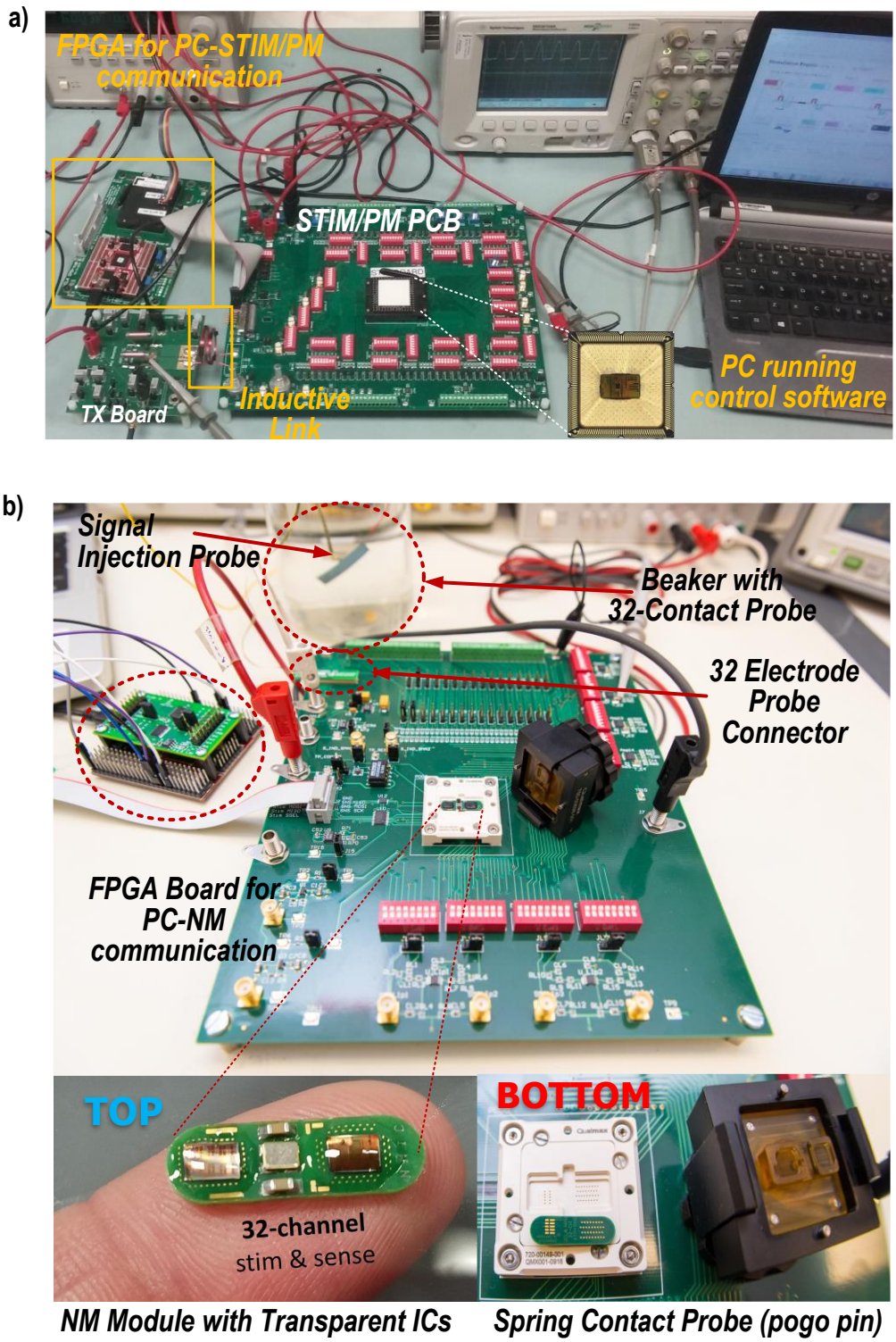


Fig. 3.33: Test Set-Up for NM Assembly In-Vitro Measurement.

The 13.56 MHz signal is used for the power carrier frequency during the rectifier's power conversion efficiency (PCE) evaluation while in the overall measurements, the system self-tunes to a frequency in the range 10.5MHz-13.56MHz.

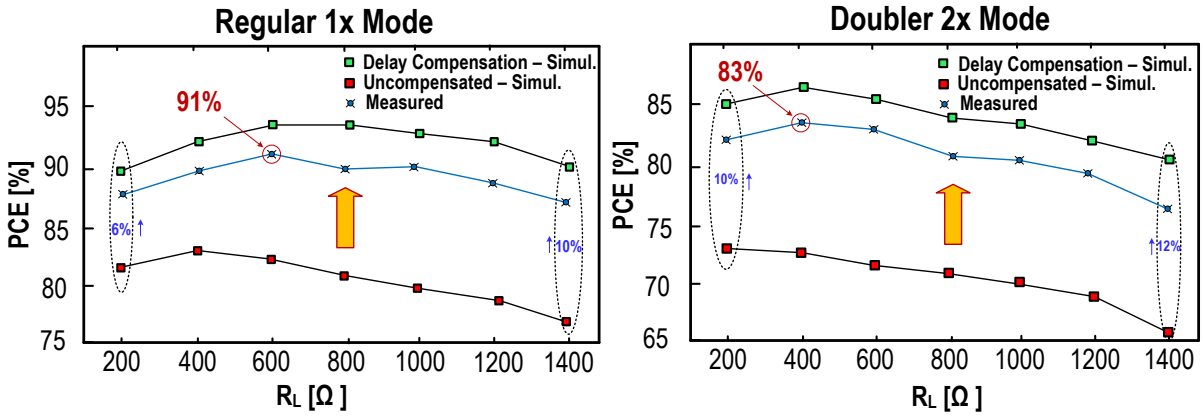


Fig. 3.34: Simulated and measured PCE versus R_L. Impact of delay compensation.

Figure 3.34 shows the PCE performance comparison between the delay compensation technique turned-on and turned-off. Measured results show that high PCE is maintained over a wide range of output powers. During the Regular (1X) Mode, our approach offers, on average, 8% improvements in PCE with 91% peak efficiency and stays above 87% for most of loading conditions. Measured rectifier's PCE, that operates in Charging (2X) Mode, shows up to 12% and 10% PCE improvement during light and heavy load, respectively. Measurements clearly show that implemented adaptive ON/OFF delay compensation technique is more beneficial in eliminating the reverse current flow for lighter loads. This is consisted with our prediction, since the integrated ALC unit is more effective for moderate and small output currents.

To demonstrate the functionality of the NM unit, we have conducted in-vitro measurements with a 32-electrode probe. Figure 3.33b shows the test set-up that is used to evaluate the system integrity under concurrent *stim* and *sense*.

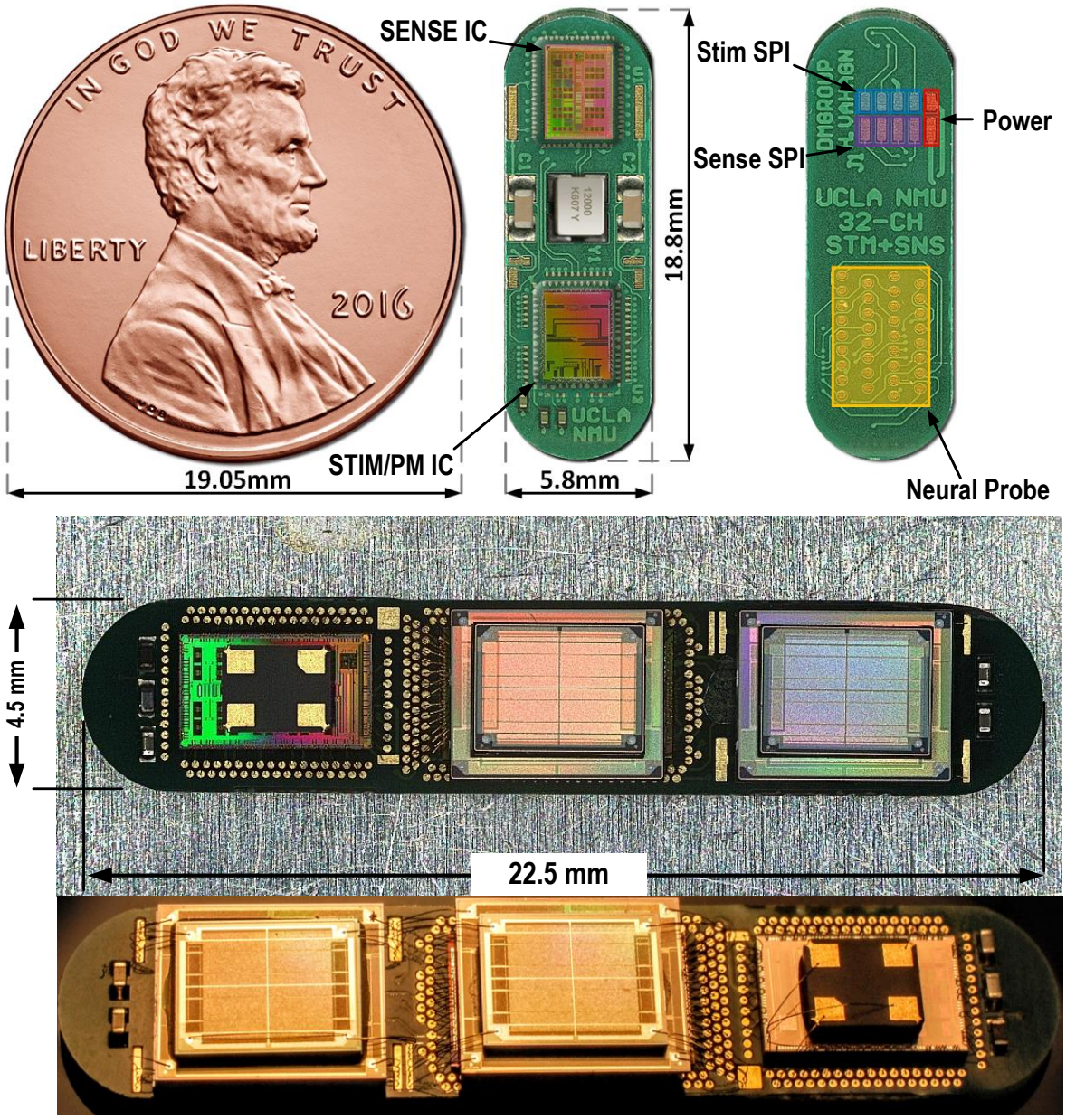


Fig. 3.35: Top) NM PCB Assembly – 32 channel version; Bottom) NM PCB Assembly – 64 channel version

NM units are designed and assembled for both versions of STIM/PM and SENSE ICs. The NM PCB assembly for the first IC version, is smaller than a US penny and occupies 135mm^3 of volume, Fig. 3.35-Top. The functionality of these two ICs is supported by only a few passives - 6 off-chip components are placed on the top side of the PCB. The bottom side is reserved for external connections - 2 anti-phase AC power lines, STIM and SENSE SPI interfaces, and 34 contact neural

electrodes (1 ‘common’, 1 ‘reference’, and 32 targets). This NM unit is small in size, which makes a high-channel count, closed-loop neuromodulation possible. The low-profile NM PCB assembly can be housed in a small package for on-the-skull implantation. This scheme minimizes the recording interference and reduces power in the cables for the NM. Distributed architecture allows a clinician to adjust the number of NM satellites without modifying the system design.

To sense biomedical signals of interest, usually only 60dB of dynamic range is required. On the other hand, if the stimulation is enabled during sensing, dynamic range requirement for the sensing front-end goes up to 90dB to capture stimulation artifacts. Wall-powered devices, available in the current market, covers this range by burning more power. However, spending too much power for the implant-scale devices is not suitable because the battery will drain too quickly.

The NM unit is placed in a spring-loaded socket that is used during the testing since it provides a good connection through its feedthrough contacts. Along with the NM test board, a signal generator is used to emulate a neural signal, AC-power supplies to deliver power to NM, oscilloscope to observe stimulation output, and a PC that is running the control software.

A 7Hz sine wave signal, that represents a neural signal, is injected through the large probe into the beaker. The smaller probe is the actual neural probe that contains both stim and sense contacts. During the recording, the sine wave is present, but there are high-frequency components (fuzziness) riding on top of it. These high-frequency components that are contaminating the signal, are stimulation artifacts. The recorded waveform, Fig. 3.36a, shows no front-end saturation with artifacts from periodic 3mA stimulation pulses. The stimulation artifact is $\leq$40mV because of the differential stimulation strategy. The time-domain waveform shows a clear 7Hz envelope and the frequency-domain plot reveals no distortion of neural-signals.

The stimulation artifacts fall inside the LFP band and conventional filtering is not possible. As shown in Fig. 36b, the in-band stimulation artifacts are suppressed by up to 114dB by a custom digital stimulation artifact rejection method, [68]. This stimulation artifact rejection method implemented as digital signal processing unit removes these artifacts before processing neural signals.

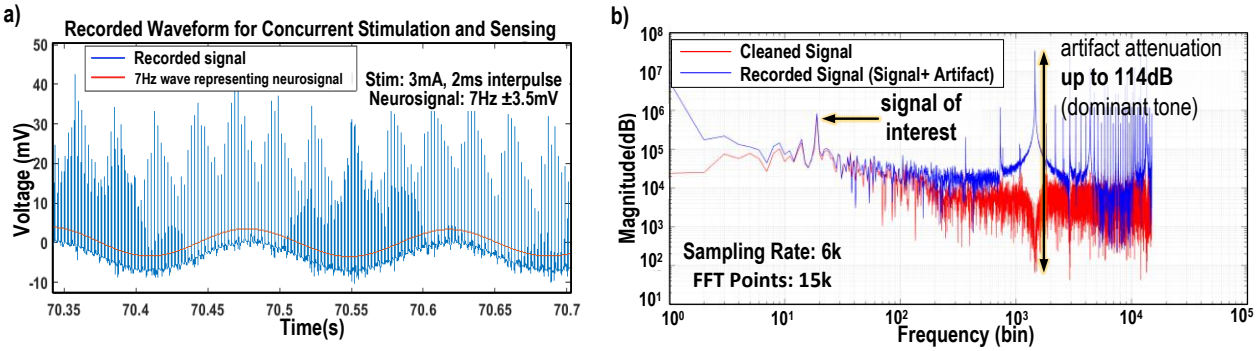


Fig 3.36: a) Time-domain waveform shows that the sensing front-end doesn't saturate under stimulation artifact; b) In-band artifact suppression.

The main advantage of this implantation scheme is the proximity of the sensors and stimulators to the electrode arrays, compared to the traditional approach where the pulse generator is in the chest area.

In the similar manner, we have designed NM PCB unit for 64 channels that houses version 2 of the STIM/PM and SENSE ICs. NM supports different power delivery options and flexibility - wise outperforms version 1 (32channels). This unit includes stacked integrated IPDIA capacitors to further downsize the overall NM module. Similarly, the bottom side has the contacts for external connections - 2 anti-phase AC power lines, 3-wire SPI interface, and 66 contacts neural electrodes (1 'common', 1 'reference', and 64 targets). Overall the NM capsule occupies 552mm^3, while the

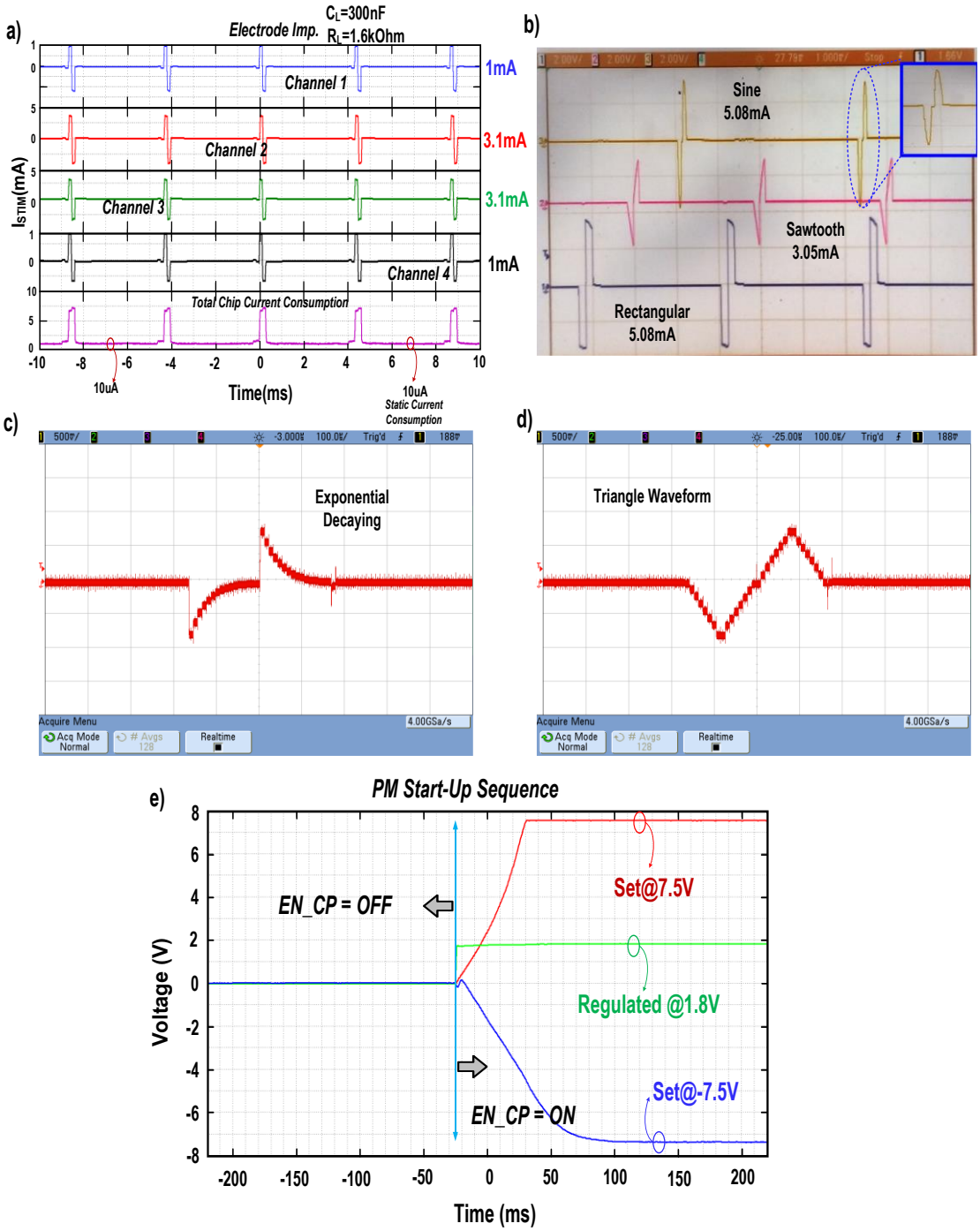


Fig. 3.37: a) Measured simultaneous current waveforms with active duty-cycling; b-d) Arbitrary Waveform Shape – Concurrent Stimulation; e) Power Management Start-Up Sequence.

inner volume where active electronics is placed takes 338mm^3 - W=4.5mm and L=22.5mm, Fig. 3.34-Bottom.

Figure 3.37a-3.37d shows the measured simultaneous current stimulation waveforms.

Waveform can be configured to have rectangular, exponential, sawtooth, triangle and sine pulse shape. Power management start-up sequence is shown in Fig. 3.37e, where control signal EN_CP

Table 3.1: Comparison with the NM state-of-the-art.

Reference		[34]	[35]	[36]	[37]	[38]	This work
Application		Spinal Cord	Cortical	Cortical/ Sub-cortical	Cortical	Cortical	**Cortical/ Sub-cortical**
Process		HV 180nm	HV 180nm	180nm	HV 180nm	HV 250nm /90nm	**40nm / HV 180nm**
# of STIM Channel / # Engines		160/40	8/1	64/4	8/1	32/1	**64/8**
STIM Mode		Current	Current	Current	Voltage	Current	**Current**
Max Current / resolution		0.5mA/7	30µA(fixed)	5.04mA/8	0.23mA/5	12mA/6	**5.1mA/8**
STIM freq/pulsewidth/ resolution		20k/8m/10u	N/A	225/500u/15u	220/440u/40u	1k/12m/100u	**20k/1.26m/10u**
# recording channels		16	8	64	8	8	**64**
AFE	Linear Input Range (V_{pp})	36m[b]	10m	4.6m	1m	16m[b]	**100m**
	ENOB (bit)	8.5[a]	9.57[a]	10.2	6.5	12[a]	**12.8**
	Integrated Noise(μV_{rms})	7.68	5.23	1.6	1.97	2.3	**4**
	Area(mm^2/Ch)	N/A	0.38	0.025	0.35	N/A	**0.12**
	Signal BW (Hz)	1-7k	1-7k	1-500	500-3k	1-250	**1-250**
	SFDR	N/A	N/A	63dB[c]	N/A	N/A	**>80dB**
	Input Impedance	d	d	30MΩ	d/e	d	**∞**
Sensing Under STIM Artifact		NO	NO	NO	NO	NO	**YES**
Supply(V)		±1.8/± 12/± 6	1.8/10	1/3/6/9/12	1/4.5	±1.8/± 12/± 6	**1.8/1.2/0.6/±7.5 Prog. ±5**
Peak Rectifier Eff.		N/A	84.86%	80%	N/A	N/A	**82% (wired) 91%(wireless)**
# of passives		6+2	N/A	N/A	N/A	>66	**5+1**
Power transfer		Wireless	Wireless	Wired	Wireless	N/A	**Diff. Wired/Wireless/ Rech. Bat.**
Implant Size		0.5cm^3	>3cm^3	N/A	N/A	Not Implantable	**0.135cm^3/0.338cm^3**
Chip size(mmxmm)		4.4x5.7	2.76x4.88	2.4x4.8	3.06x2.53	N/A	**2.59x3.6 / 3.05x5.3(STIM) 2.6x3.8(SENSE)**

a) ENOB only for ADC. The amplifier nonlinearity not reported

b) Inferred from reported supply rail and minimum amplifier gain

c) Insufficient Linear range and large common mode swings

d) Requires off-chip ac-coupling caps

e) Parallel recording from all channels not possible

is used to enable/disable HVG.

Our NM implant shows superior form factor and performance compared to the prior art, Table 3.1. The improved performance is a result of highly optimized designs and a modular approach that combines the best of different CMOS technologies. It also provides a unique ability to rapidly

customize systems with varying channel counts via heterogeneous “chiplet” system-in-a-package assembly.

3.6 Conclusion

We proposed the next generation neural interface that is minimally invasive, and addresses the demands for limited area and power. It provides a real-time, full duplex communication during concurrent stimulation and recording of neural signals. Further, our modular approach and scalable architecture allows gathering data from a grid of NM implants. This multi-channel interface meets the requirements of human-quality implants at unprecedented level of electronic miniaturization as compared to the prior art. It offers major new perspectives that translate into significant clinical benefits: always-on sensing for enhanced speed and accuracy of closed-loop therapy, multi-channel arbitrary stimulation waveforms with user-friendly programming, high-spatial-resolution neural interface for more precise target localization.

Also, we have presented the first integrated full-fledged MIMO power management unit that supports different power delivery options, such as wired, wireless and rechargeable battery. This flexibility extends the application range for our NM implant. An adaptive, real-time ON/OFF delay compensation schemes for both N-type and P-type active diodes in an active rectifier, are implemented. The active rectifier can operate in 1X and 2X mode as a part of a 13.56 MHz wireless power transfer link. Due to the calibration schemes, the circuit delays (propagation delays of gate drivers and comparators) are well compensated across PVT corners and mismatches. Proposed circuit techniques improved the PCE (>90%) across a wide loading range, while ensuring that the wireless power link delivers a stable voltage to the implant across load and coupling variations.

CHAPTER 4

Hardware accelerator for simultaneous, real-time recording of large neuronal ensembles for brain imaging

4.1 Introduction

One of the goals of neuroscience is to explain how neurons process, store and encode information. In the brain, information is conveyed through neuron spikes - neurons fire different numbers of spikes in certain time steps. To observe the activity of the brain, currently there are two technologies that offer the prospect of dense, high-fidelity recordings from neurons located within a local volume of brain tissue. The first is a classical (electrophysiological) technology that is based on densely spaced electrodes for extracellular recording. This technology can offer good temporal resolution, but there are a few problems present in this approach. Neuroscientists are probing the brain with electrodes for the past 40 years, but the resolution with which they can probe has progressed slowly. In the best case (theoretically) they can pick up signals from a few hundreds of neurons (10's-100 electrodes), Fig. 4.1. Also, the use of this technology for the spatial localization and cell-type classification of recorded cells is limited. A further drawback is the tissue damage and neuro-inflammation from large electrodes placed in the brain.

Apart from this approach there is an alternative technology that has gained a lot of attention in recent years, since it is able to provide large-scale recordings of brain activity, [71] - [74]. The development of genetically encoded fluorescent indicators and optical microscopy has allowed progress in visualization of neural activity in mammalian and nonmammalian nervous systems. In this approach, the brain is genetically encoded with indicators. When neurons fire, the indicators come into the cells (neurons) and the fluorescence activity is recorded by image sensor. This technology offers a fine spatial granularity (one to two orders of magnitude higher brain coverage),

enabling the observation of thousands of neurons and individual cell activity in real-time. Furthermore, this technology is relatively non-invasive.

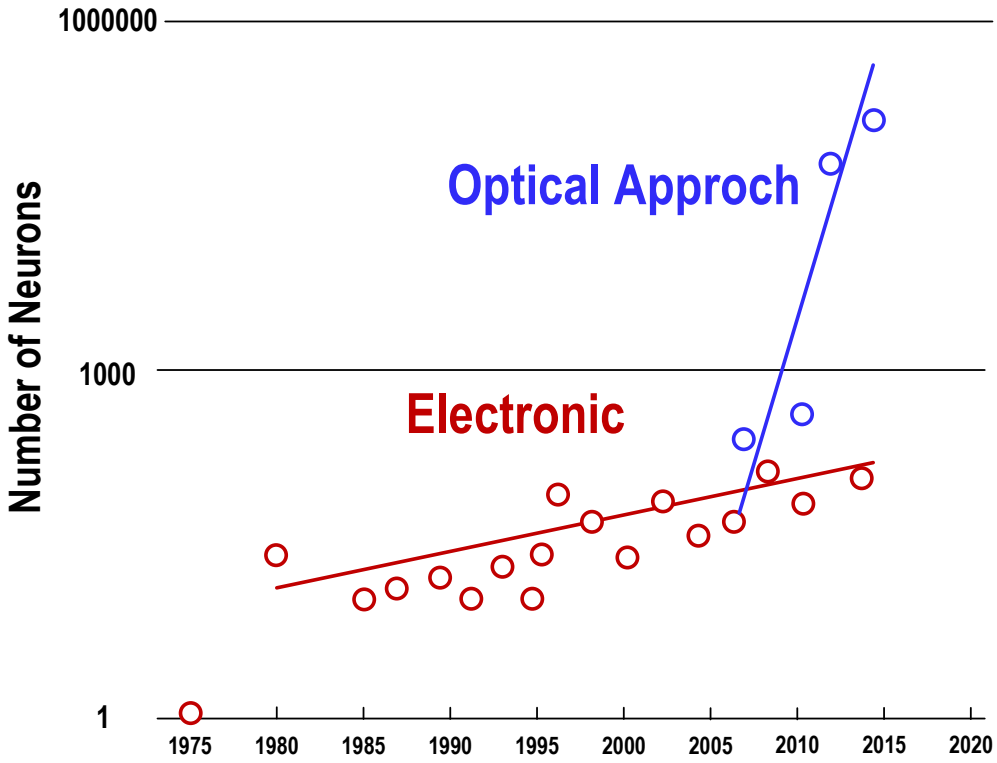


Fig 4.1: Electrophysiological vs. Optical Approach (DARPA-NESD).

Currently, two types of indicators are used:

1) Ca^{+2} based indicators are commonly used and preferable for picking up slower signals. Single/Two-photon imaging of calcium indicators enables simultaneous recording from thousands of neurons over long periods of time, and it provides an indirect measure of neuronal spiking activity. Often, this type of imaging is combined with the optogenetic method to enable closed loop (recording and stimulation) feedback experiments for a specific type of cells. However, this is not a straightforward task since Ca^{+2} imaging only provides an indirect measure of spiking activity. The imaging is based on estimating spike trains from the raw calcium fluorescence, often contaminated with high, time-varying baseline noise. This procedure is commonly referred to as deconvolution.

2) Genetically Encoded Voltage Indicators (GEVI) are promising since they enable the fluorescence readout of fast (~ms) neural dynamics, allowing the capture of action potentials (AP). AP detection needs fast sampling (200Hz-6kHz). GEVI has also progressed to the point where response amplitudes (single AP detection) can be similar to those used in Ca^{+2} indicators, [75] - [77]. Apart from the fact that sampling rates are up to 20 times higher than the one used for the calcium imaging, so far GEVI indicators showed limiting experiment duration, [75].

At the core of this imaging technology is fluorescence microscopy which is miniaturized down to the size of a few mm^2 of cross-section, [78]. In most common animal experiments, the fluorescence camera is attached to the skull of the mouse, and neuroscientists can observe individual neural dynamics while the mouse is freely moving, Fig. 4.2a. So far, the neuroscience community has resolved design problems related to the ultra-fast miniscope – 100's of frames per second (fps) are now possible, allowing the capture of LFP signals. The next generation of the complete system would require real-time neural deconvolution and de-mixing, a closed-loop

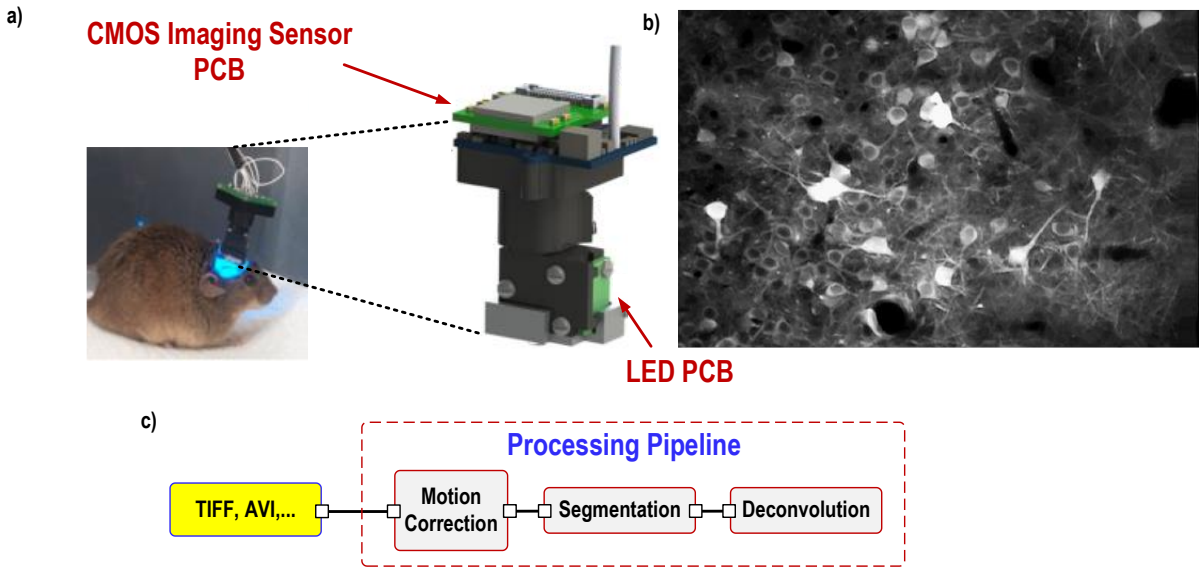


Fig 4.2: a) Wired mini-scope, [77]; b) Calcium Imaging - Single Frame; c) Video Processing Pipeline.

feedback system and a low-power wireless data/power link. These requirements have to be addressed on the hardware side. But there are a few obstacles that prevent a broader usage of this technology: 1) The sensor is not stably fixed, causing motion jitter in the received frames; 2) The position of neurons in a frame is unknown – neurons are visible only during the short period of time when they fire; 3) Even when the position and shape of neurons are known, extraction of the spiking signal (valuable information) from the raw fluorescence traces is necessary.

Currently, the state-of-the-art signal processing, in the neuroscience community relies on offline processing of captured videos. In the best case, up to 15 mins of recording is feasible before the battery powering the miniature camera and its electronics runs out. Even these short-duration recordings yield 10's of GB of data. The custom-developed toolboxes (usually MATLAB) need hours to process this amount of data, and often the videos need to be split into several smaller segments and then processed separately.

Our primary goal was to develop online low-power hardware to support the processing steps in Fig. 4.2c in real time. Enabling hardware-friendly processing and drastic data reduction can speed up and automate the analysis of calcium imaging videos and provide the extraction of biologically relevant information in real-time. Motion Correction and Blind Neuron Detection are solved by using a distributed and power efficient implementation of various computer vision algorithms. By exploiting the sparse nature of the spiking signals and the neurons, both in time and spatial domains, a dedicated processing unit that maps the sparse-approximated (SA) algorithm, into hardware, extracts the valuable information in real-time and in a highly parallel fashion.

4.2 Hardware architecture for real-time frame alignment

So far, all steps in the processing pipeline, Fig. 4c, are solved in software. Motion Correction (MC) is the first task in the pipeline that corrects frame-to-frame motion, which is a big issue in freely moving subjects. The MC task is particularly important during closed-loop experiments (that involves stimulation), where precise cells detection is necessary. Otherwise, some parts of neurons (Region Of Interest - ROI) will fall in or out of ROIs of other neurons during motion. Hence, it would corrupt the fluorescence levels across different ROIs. There are different types of a motion jitter – slow drifts and faster motions. Fast motions are a result of grooming, since during acquisition, the field of view (scanned from top to bottom) moves significantly. This causes non-rigid deformations, [79]. Fortunately, in the most practical cases, translation motion is dominant (rigid motion) and there are several motion-correction techniques that can compensate for this type of motion. For motion correction, neuroscientists usually use a Fourier-transform approach (FFT), [80], [81] which includes two-dimensional FFT-accelerated convolutions, combination of down-sampling, dynamic programming, etc. These FFT-based frame-to-frame rigid (or non-rigid) alignment methods are very computationally and time execution expensive.

Essentially, the purpose of the Motion Correction block is to align every upcoming frame with the reference frame, which is usually the first frame in the video sequence. This can also be achieved through various techniques that rely on intensity-based methods, [82], which compare the video frames (frequency or spatial domain) based on pixel intensity. Also, feature-based methods, [83], apply different transformations to the frame in order to detect and extract features. Later, based on the relation between the features (e.g. edges), motion jitter can be corrected.

Inspired by the work on video stabilization, [84], we decided to exploit template matching techniques that essentially search the space of parameters to maximize the resemblance between

the reference and the transformed version of the current frame. A general transformation would assume affine model that would include image translation, rotation and scaling. However, for the most practical cases, simple translation is sufficient, which is preferable since the hardware would require only arithmetic operations and it would avoid interpolation for more complex transformations.

The template matching technique requires a similarity metric definition. Similarity function describes a metric that compares the candidate image (created by a transformation) to the reference one. By employing the similarity metric, the search function examines the translational space to derive the motion vector $(\Delta x, \Delta y) \in [-P, P]$, where P determines the maximum coordinate shift. As a rule of thumb, the maximum coordinate shift P is always taken as 10% of the smaller frame dimension (P=min (L, W), where LxW determines the resolution of the frame).

There are many functions that are used to compare 2-D images, such as sum of squared differences, normalized cross-correlation (NCC), sum of absolute differences, etc. Even though these functions offer similar results, their computational complexities vary a lot. It is known that the 2-D NCC gives the most precise similarity estimate between two images, but from the embedded hardware point of view, this choice is inadequate due to high hardware complexity. Therefore, to have an efficient implementation in dedicated hardware, we decided to use the Sum of Absolute Differences (SAD) metric, (4.1), which requires only arithmetic addition and absolute value as operations,

$$\mathrm{SAD}(\Delta x, \Delta y) = \sum_{m=0}^{L-1} \sum_{n=0}^{W-1} |F_{cand}(m + \Delta x, n + \Delta y) - F_{ref}(m, n)| \quad (4.1)$$

where F_{can} and F_{ref} denote translated version of the current frame and reference frame, respectively. Here, $(\Delta x, \Delta y)$ depicts the translation vector.

Further, we have introduced two modifications that additionally relax the memory requirements and reduce the number of computations. By observing the general definition of SAD, we can see that one SAD requires 2 complete images pixelwise. Instead, by using the SAD definition accompanied with the Integral Projection (IP) approach, suggested in [84], [85], we can collapse the image columns and rows onto projection vectors, (4.2), and consequently reduce the dimensionality of the task (4.1).

$$F_{px(m)} = \sum_{n=0}^{W-1} f(m,n)\,, F\_py(n) = \sum_{m=0}^{L-1} f(m,n) \tag{4.2}$$

$$SAD_X(\Delta x) = \sum_{m=0}^{L-1} |F_{can\_}px(m+\Delta x) - F_{ref\_}px(m)| \tag{4.3}$$

$$SAD_Y(\Delta y) = \sum_{n=0}^{W-1} |F_{can\_}py(n+\Delta y) - F_{ref\_}py(n)| \tag{4.4}$$

As a result, the IP can be calculated in parallel and on the line as the processing unit receives the pixels from the camera. By using IP, the computation of SAD is reduced to 2 one-dimension processes, (4.3) - (4.4). The true benefit of the IP approach is noticeable during the exhaustive search procedure, in which we are trying to find an optimal displacement vector (Δx, Δy) that will maximize the similarity between the reference and the translated current frame. Instead of implementing a searching algorithm on 2D images, we can perform Split-Half Search along every projection axis and thus parallelize the computation and additionally reduce the memory and execution time. The final parameter values denote the Global Motion Vector (GMV), that represents a motion estimation between the two frames, (4.5).

$$(\Delta x_{gmv}, \Delta y_{gmv}) = \text{argmin}\,[SAD_X(\Delta x), SAD_Y(\Delta y)] \tag{4.5}$$

Fig. 4.3. shows the simplified diagram for the motion estimation unit. The camera is sending the frames line-by-line, while the local memory buffer stores them in the hardware. IPs for the reference frame are calculated during acquisition and the values are stored in local RAM memory.

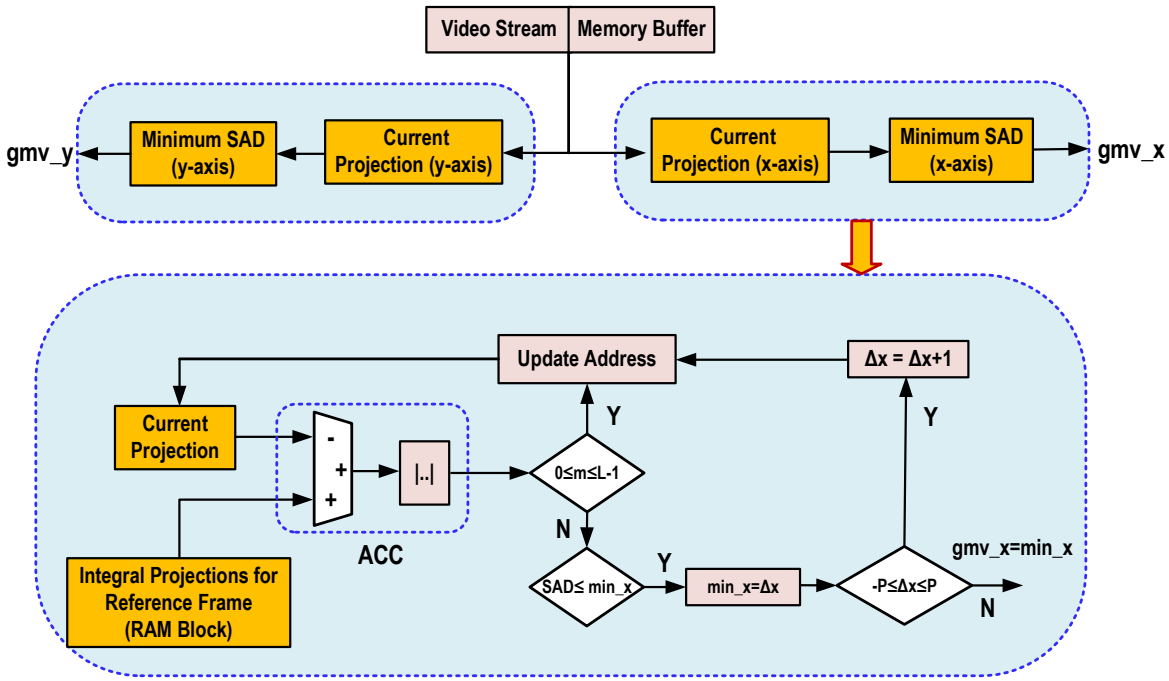


Fig 4.3: Integral Projection Approach - Motion Estimation Unit.

There are two IP blocks, that are concurrently reading the pixels values from the line buffer and computing the IP values along x and y axis. The IP blocks are followed by the two minimum-SAD units that simultaneously calculate the SAD between the reference and current frame projections (for each axis), and scan for the optimal (Δx, Δy) vector that minimizes the SAD. These units use a simple adder, an absolute value circuit and the accumulator, while the Update Address block produces the projections locations in the current frame that are shifted by an offset value. The offset values are updated sequentially (from -P to P), in order to perform the exhaustive search. Minimum SAD units output the Global Motion Vector.

This procedure does not have to be performed on the whole frame. By calculating the GMV vectors on the block level, Fig. 4.4, we can still get sufficiently good results. By simply loading the strip of image lines into the buffer, calculating the optimal translation parameters for several blocks and averaging the values, (4.6), we can get good performances while reducing the memory requirements and parallelizing the computations. With this distributed approach, the execution time of the motion estimation unit is reduced by a factor BxS (B=number of blocks within the strip, S= number of strips within the frame).

After the motion vector is found, the candidate image is a displaced version of the current image. This implies that the dimension of the processed image is actually smaller in size. By subtracting the GMV vector from the pixel's original addresses, we can obtain the addresses for the new pixels. The pixels will show up at the output with a latency that is equal to the time needed for loading the initial strip of lines into the memory buffer.

$$\overline{\Delta x_{gmv}} = \frac{1}{B}\sum_{k=1}^{B} \operatorname{argmin} SAD_{X_k}(\Delta x), \qquad (4.6)$$

$$\overline{\Delta y_{gmv}} = \frac{1}{B}\sum_{k=1}^{B} \operatorname{argmin} SAD_{Y_k}(\Delta y)$$

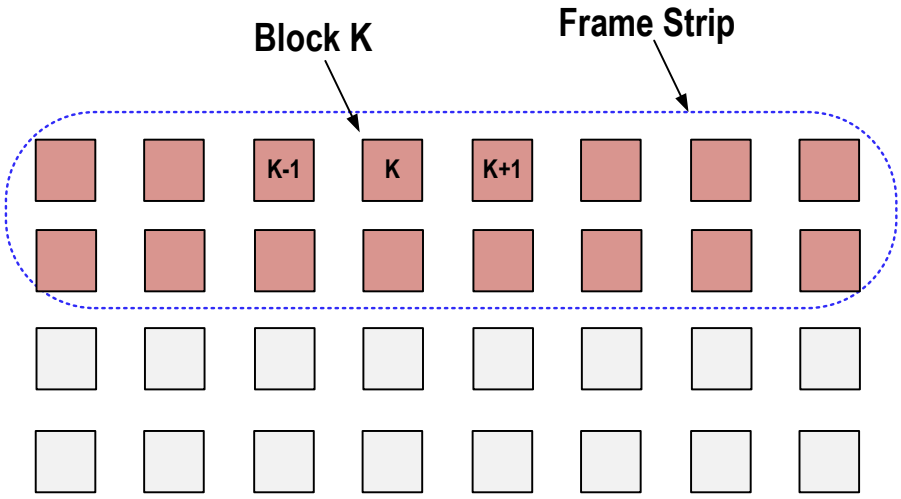


Fig 4.4: Block-Level Motion Estimation.

4.3 Neuron Detection

Since they are major challenges, segmentation and deconvolution of neural activity from the calcium images have attracted a lot of attention in the previous years. Many methods have been proposed, [73], [87] - [90], but most of them require the whole video to be loaded into memory, while the processing involves various computationally and memory expensive techniques. Even though there is no systematic approach to solve these problems, two paths can be found in the literature that offer solutions for these tasks. In the first approach, segmentation and deconvolution are treated separately. Such methods include greedy algorithms, [88], supervised learning, [89], etc. Recently, authors in [73], [87], proposed methods that are based on the hypothesis that spatio-temporal Ca^{+2} activity can be expressed as a product of two matrices: a temporal matrix that conveys the information about the evolution of Ca^{+2} concentration for every neuron, and a spatial matrix that contains the information about position and shape of each neuron. Mukamel in [73] employs PCA/ICA, while Pnevmatikakis in [87] proposed a constrained nonnegative matrix factorization (NMF) method for simultaneous denoising, factorization and spike deconvolution.

These methods are often manual or semi-manual since they require tuning of various regularization parameters, the number of principal components, the number of ROIs that needs to be found, threshold setup, etc. Previous approaches also employ some kind of greedy algorithm or spatial filtering techniques for the initial estimate of neuron footprints. All these approaches are impractical for the implementation in hardware, since they assume offline video processing and hence the detection approach has to be revised.

Neuron localization and identification of the spatial footprints pose a difficult task due to the presence of noise and surrounding neuropil activity. Also, since the camera is projecting a 3D volume onto a 2D plane, neuron footprints are often spatially overlapped.

We proposed a real-time processing technique that detects and de-mixes neuron footprints on-the-fly as the frames are received and does not require any assistance ("blind" detection).

The detection step is called the "learning phase", since during this phase we are trying to obtain the shapes and the position of neurons based on incoming frames. To detect neurons, we have used a modified version of a computer vision algorithm called Maximally Stable Extremal Regions (MSER), [91]. The MSER algorithm can be described as follows. Let's consider the MxN grid that corresponds to the MxN intensity grey image and let's start thresholding with t=255 down to 0 with step Δ (i = number of threshold levels | i=1:1:mod (255/ Δ)). During thresholding, all pixels that have value equal or above the current threshold t, start showing up as a white pixel in the frame. In other words, as t is reducing from 255 towards 0, some white areas will start to become visible in the frame. By further decreasing the threshold, new white regions will appear and the previous regions will continue growing and possibly merging. Ultimately, the whole frame would become white. During this procedure, we keep track of the size (cardinality) for each white region. We say that the MSER is detected if the stability function q(t), defined in (4.7), has a local minimum.

$$q(t) = |Q(t + \Delta) - Q(t - \Delta)|/Q(t) \tag{4.7}$$

Q(t) denotes cardinality of the region at the threshold t.

For automatically assigning labels to the above-determined white regions at all t values, an efficient Union-Find (UF) algorithm is employed, [92] - [93]. The UF algorithm, also returns the seeds (reference points and sizes) and the seed list at each threshold value t. For convenience, the first pixel location in every region is assigned to be the reference point. After merging of two or more regions, the reference point of the largest region becomes the new reference point.

The detected MSER regions match the positions and shapes of the neurons in the current frame. Since the neuron visibility changes with time, it is necessary to conduct the MSER and UF procedures on a train of frames, and to update the detected MSER base after every frame.

To implement MSER and UF algorithm in a hardware friendly way, motivated by the work in [94] – [95], we have divided this phase into 4 basic steps:

1.) Preprocessing – this step outputs the vector that contains pixel positions sorted by the number of pixels at each intensity level and by the intensity level. First, the intensity histogram is calculated and then it is sorted by using the radix-bin-sort.

2.) Clustering – during this step, UF algorithm is employed to keep track of the connected pixels. UF can check whether two pixels appear within the same region; if not, the algorithm can group them and add them to that region. Also, it keeps track of the cardinality of the region. This step uses an auxiliary memory bank (Region Map - RM). RM is the same size as the image; each location within the RM corresponds to the pixel position in the original image. As suggested in [95], each position in RM contains one number (FL) whose role is to determine the status of the pixel. If the pixel is not connected to any other pixel - (FL=0); if the pixel belongs to the same region as the pixel at position FL – (FL>0); if the pixel is the reference point for the region (FL<0). To keep track of whether the pixel is placed in the RM, a single-bit is assigned to every pixel-position in the RM.

3.) Detection – In parallel with changing the threshold, it is necessary to keep track of all regions and their cardinality for three consecutive thresholds because of the definition of the stability function, (4.7). That explicitly helps in calculating the stability function q and detecting the local minimum. To decide whether the function $q(t_0)$ has a local extremum in t_0, we also store the derivative value dq/dt and the value q. Simple hash-tag memory

structure is used for this particular task. This hash-tag memory has 5 rows for every seed(region), as shown in Fig. 4.5. Note that in order to avoid division by zero, during the addition of the new seed, $q(t_0)$ and $q(t_0 - \Delta)$ are filled with ones, [96]. To simplify calculations, instead of keeping dq/dt in the memory, a one-bit flag $sign(q(t_0) - q(t_0 - \Delta))$ is sufficient, [94]. When this flag is negative and the stability function is increasing ($q(t_0+\Delta)>q(t_0)$), the function has a local minimum $q(t_0)$ at t_0.

@ t :

Seeds_ID	#1	#2	#3	#4
Q(t-Δ)	40	24	15	x
Q(t)	67	44	49	x
Q(t+Δ)	128	105	150	x
q(t)	1.31	1.84	2.76	x
sign(q(t)-q(t-Δ))	x	x	x	x

@ t+Δ :

Seeds_ID	#1	#2	#3	#4
Q(t-Δ)	67	44	49	1
Q(t)	128	105	150	1
Q(t+Δ)	145	310	190	24
q(t)	0.61	2.53	0.94	23
sign(q(t)-q(t-Δ))	-1	1	-1	1

Fig 4.5: Example of Hash-Tag Memory Update.

4.) Display – when the MSER is detected, it is important to save the MSER position (size and the pixels positions) into a separate memory bank. The size of the detected MSER region is equal to (1-FL). But, with RM-based approach, there is no efficient way to perform this transfer, unless we scan the whole RM structure and check if the pixels have the same reference point. Also, to determine if the function has a local minimum at $q(t_0)$, all pixels with the intensity level $t_0 + \Delta$, have to be placed into RM. Otherwise, the readout will be wrong. To overcome these shortcomings, the authors in [95], proposed a simple extension of the UF algorithm by introducing the linked region concept. The basic idea is that every new pixel placed into the RM should refer to the next pixel within the same region. Consequently, a circular chain that connects all pixels in the region is formed. An elegant way to create this chain is to assign another field for each position in RM, that carries a

pointer to the next pixel within the same region. Initially, the pointer value is the position of the pixel.

Furthermore, two modifications were adopted as compared to the original MSER algorithm flow. Since we are detecting the neurons across a train of frames, not all neurons are going to be visible in one frame. It is important to neglect all MSER regions that are the result of merging of previously detected (smaller) MSERs, since it is obvious that we have multiple neurons in that region, just visible at different periods of time. Also, if multiple MSERs are

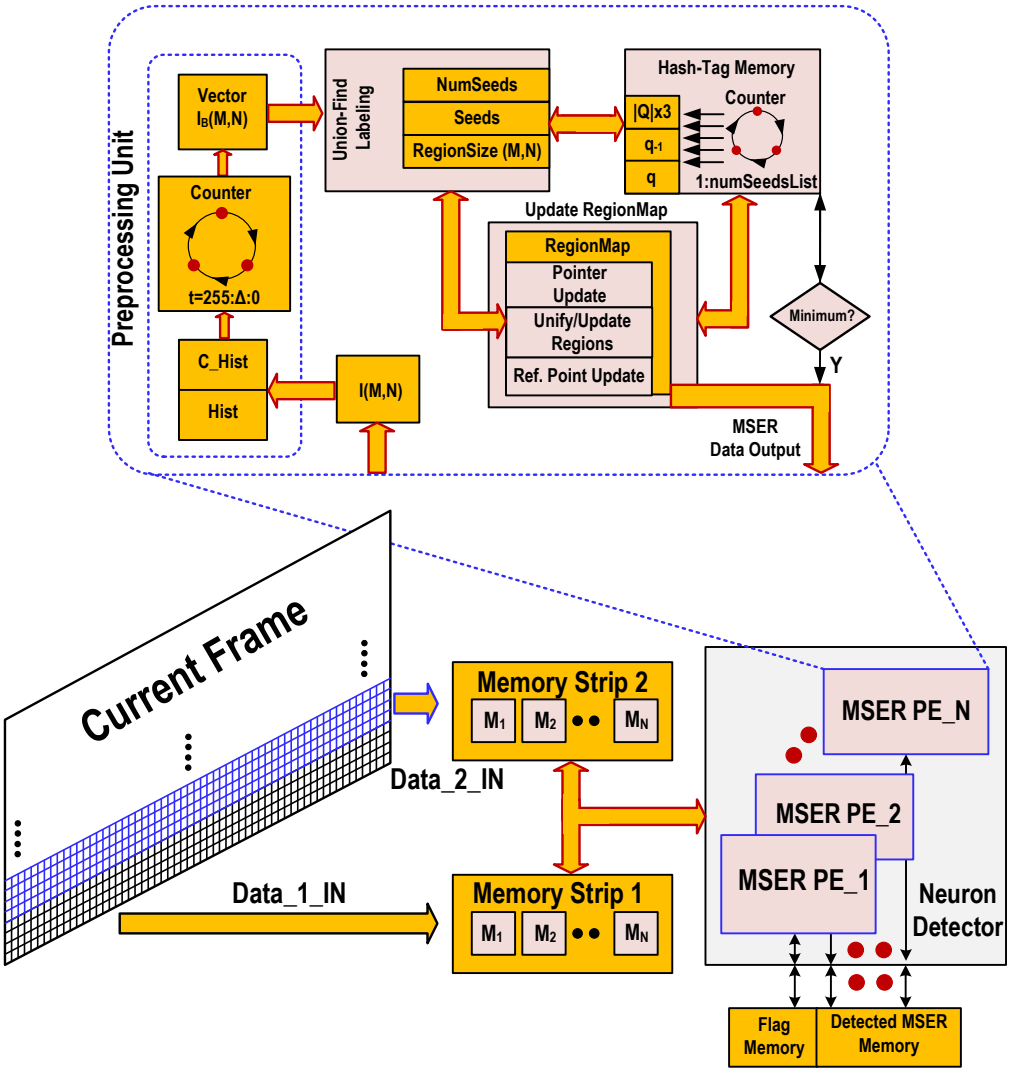


Fig 4.6: Distributive Approach – Architecture (MSER&UF) for the Neuron Detection.

detected at the same place, we should keep the one whose stability function q has a global minimum, [92].

Implementation of the MSER and the UF algorithms on the whole frame is very memory and computationally expensive. Hence, we decided to adopt a distributive approach, [94]. Since pixels are coming line by line, we can divide image into strips and every strip into several smaller blocks. While we are processing one strip with multiple MSER processing elements (PEs), we can simultaneously load the new line strip into the memory buffer. However, by employing a distributive approach, certain neurons could be split into several blocks during parallel processing. This issue can be elegantly circumvented by employing a single flag memory; whenever a new neuron is detected, we can check if its pixels are "touching" the neighboring blocks, and accordingly conduct merging, [94]. Figure 4.6 shows the block-level architecture for the neuron detection in the calcium imaging.

4.4 Real-Time Deconvolution of the Spiking Signals from Ca^{+2} Imaging

After the "learning phase", when the neuron positions are detected, it is necessary to infer the valuable information from the neuron's fluorescence time series – extracting the raw fluorescence traces is not the primary scientific goal. However, determining the underlying neural activity from large neural ensembles, and the specific timesteps in which neurons fire, is a difficult and an open problem in neuroscience. Current methods, [87], [90], [98], [99] are typically applied on large imaging data offline, after the recording experiments are done. Since there is a need for the causal exploration of neural activity (dynamics, connectivity), real-time processing that will enable closed-loop experiments, is an imperative. Closed-loop experiments in brain-machine interfaces are usually driven by electrical recording; enabling the optical method for that purpose would

provide more details into how neurons (individual cell resolution) behave during the learning (comprehension) phase.

For this purpose, after identifying neuron locations and de-mixing spatially overlapped sources (footprints), deconvolving and denoising the neuron spiking activity from the slower dynamics of the Ca^{+2} indicator is the next step. Motivated by the works in [87], [99], we have adopted the proposed algorithms and offer simplifications that can lead to a hardware-friendly solution for the sparse, non-negative deconvolution problem.

4.4.1 Mathematical Model for Ca^{+2} Dynamics

Let's consider a stable auto-regressive (AR) model for the Ca^{+2} dynamics proposed by Pnevmatikakis, [87]. Let x_t denote the raw fluorescence trace at the timestep t, and c_t denote the underlying Ca^{+2} concentration at the timestep t. The Ca^{+2} dynamics can be approximated by a stable auto-regressive (AR(p)) process of order p, where p is a small positive number, usually equal to 1 or 2, as

$$x_t = c_t + b + \epsilon_t, t = 1, \dots, T; \tag{4.8}$$

$$s_t = c_t - \sum_{i=1}^{p} \gamma_i c_{t-i}, t = p + 1, \dots, T. \tag{4.9}$$

The variable s_t in (4,8), represents the spiking signal - the influence of a spike on the Ca^{+2} level at the t-th timestep, while $\varepsilon_t \sim N(0, \sigma^2)$ denotes the noise term with variance σ^2. The term b is the time-varying baseline noise, which for simplicity is assumed to be zero. The parameter p for all practical reasons never takes a value higher than 2. The quantities γ_i denote the AR model parameters. The only accessible (observed) quantity is x_t, while all others are unobserved.

The purpose of the deconvolution step is to extract the neural activity **s** from the observation vector **x**. If we assume that the vector **s** is sparse (in most time-bins there would not be any spikes ($s_t = c_t - \sum_{i=1}^{p} \gamma_i c_{t-i} = 0$)) and that the individual values s_t are nonnegative since the spiking

would only boost the Ca^{+2} concentration level, the first order AR(p=1) model would lead to the following optimization problem, [98]:

$$\min_{\mathbf{c}} \left\{ \frac{1}{2} \sum_{t=1}^{T} (x_t - c_t)^2 + \theta \sum_{t=2}^{T} 1_{(s_t \neq 0)} \right\} \quad \text{subject to } s_t \geq 0. \tag{4.10}$$

The last term in the objective function $1_{(s_t \neq 0)}$ is equal to 1 if $s_t \neq 0$ is true. Also, θ denotes a tuning parameter that dictates the number of timesteps at which spiking happens. This form of optimization problem that incorporates l_0-minimization penalty is known as l_0-minimization task, it is highly non-convex and there are no efficacious algorithms able to find the global optimum.

$$\min_{\mathbf{c}} \left\{ \frac{1}{2} \sum_{t=1}^{T} (x_t - c_t)^2 + \theta \sum_{t=2}^{T} |s_t| \right\} \quad \text{subject to } s_t \geq 0 \tag{4.11}$$

Fortunately, it has been shown that in order to avoid computational challenges, the l_0-minimization penalty can be translated to l_1-minimization norm, (4.11), and the implied solution would represent a proper approximation of the original problem, [102], [104]. The same logic was followed in [87], [90]. Now, the non-convex problem is approximated with the corresponding convex problem for which optimization algorithms are available. If we extend the order of AR process to 2, the final optimization problem takes the non-negative LASSO form, [109]:

$$\min_{\mathbf{c}} \frac{1}{2} \|x - c\|^2 + \theta \|s\|_1 \quad \text{subject to } s = Gc \geq 0, \tag{4.11}$$

where G is defined as a lower triangular matrix of autoregression parameters,

$$G = \begin{pmatrix} 1 & 0 & 0 & . & . & . & 0 \\ -\gamma_1 & 1 & . & . & . & . & 0 \\ -\gamma_2 & -\gamma_1 & 1 & . & . & . & 0 \\ . & -\gamma_2 & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & . & . & . & -\gamma_2 & -\gamma_1 & 1 \end{pmatrix}. \tag{4.12}$$

By observing (4.11), we can conclude that authors in [87], [90], [99] introduced a hard constraint on the energy of the residual signal between the underlying Ca^{+2} dynamics and the raw fluorescence data (traces) by penalizing the sparsity of each neuron (detected ROI). Also, since

the spiking vector contains non-negative values, the last term in the objective function in (4.11) can be expressed as a function of matrix G elements and Ca^{+2} concentration vectors as:

$$\theta\|s\|_1 = \theta\sum_{i=1}^{T}\sum_{j=1}^{T} G_{ji}c_i = \theta\sum_{i=1}^{T}\left(1-\gamma_1-\gamma_2+(\gamma_1+\gamma_2)\delta_{iT}+\gamma_2\delta_{i(T-1)}\right)c_i = \sum_{i=1}^{T}\alpha_i c_i \tag{4.13}$$

where "δ" refers to the Kronecker's delta function.

Before solving the minimization norm, it is necessary to evaluate the unknown parameters – AR parameters γ_i and noise variance σ^2. As suggested in [90] and [99], AR parameters can be inferred from the time series analysis method. If the AR order p is known, the autocovariance function of the observed fluorescence **x**, ARR_x, satisfies the following equality:

$$ARR_x(t) = \begin{cases} \sum_{k=1}^{p}\gamma_k ARR_x(t-k) - \sigma^2\gamma_t, 1 \le k \le p \\ \sum_{k=1}^{p}\gamma_k ARR_x(t-k), k > p\,. \end{cases} \tag{4.14}$$

The AR parameters can be evaluated by inserting the autocovariance samples into (4.14). Later, when the AR coefficients are known, the noise variance σ^2 can be found. The autocovariance method surmises that the spiking signal **s** have a uniform Poisson distribution and effectively gives very crude estimations for the AR parameters. Since the rising time of the indicator is much faster than the duration of the timestep, the order p is assumed to be equal to 2, [87].

4.4.2 Extension to the Spatio-temporal Case

As discussed in [87], the goal of the constrained NMF method is to decompose the spatio-temporal neural activity into temporal and spatial parts. In this way, the Ca^{+2} dynamics can be modeled and the individual neuron structure (ROI) can be preserved. After the one-dimensional, single neuron deconvolution formalism is explained in 4.4.1, let's extend analysis to a full spatio-temporal case.

Assume that the field of view captures N neurons and that the time series has a total number of T timesteps. If the frame consists of d pixels (single column vector), the overall observation can be seen as matrix $\mathbf{F}_{dxT}$. The Ca^{+2} activity $\mathbf{c_i}$ for each neuron i can be depicted with AR dynamics. If $\mathbf{a_i} \in R^d$ and $B_t \in R^d$ denote the spatial footprint of the neuron i and the baseline concentration at the t-th timestep, the overall spatial Ca^{+2} concentration (at time t) can be expressed as

$$X_t = \sum_{i=1}^{N} a_i c_{it} + B_t + E_t, \quad t = 1, \ldots, T, \tag{4.15}$$

where E_t depicts the diagonal matrix with additive noise terms.

From (4.9) and (4.15) we can derive the matrix form for the spatio-temporal spike inference:

$$S = CG^T, X = AC + B + E, \tag{4.16}$$

$$S = [\mathbf{s_1}, \ldots \mathbf{s_N}]^T, A = [\mathbf{a_1}, \ldots \mathbf{a_N}], C = [\mathbf{c_1}, \ldots \mathbf{c_N}]^T, \tag{4.17}$$

$$B = [B_1, \ldots B_T]^T, X = [X_1, \ldots X_T]. \tag{4.18}$$

Solving (4.11) in the spatio-temporal case is not a trivial task. Authors in [87] proposed a method that alternatively updates the temporal matrix C, by solving (4.11) in parallel for every pixel, and estimates the spatial representation by dividing the problem into d separate programs for each pixel. For a single pixel deconvolution, they used a non-negative LARS (least angle regression) algorithm, which is convenient due to the structure of the dual representation of (4.11). On the other hand, Friedrich in [90] used the pool adjacent violators algorithm (PAVA) for isotonic regression to obtain an Online Active Set method to Infer Spikes (OASIS). However, these methods are only suitable for CPU/GPU processing since they exhibit very high computational complexity and extremely large memory requirements. Hence, their efficient hardware implementation, without algorithm modification, is not possible.

Since, the matrix A contains vectors that are very sparse, solving (4.11) for every pixel in the frame is unnecessary. Instead of updating the spatial representation based on the work in [87], we

are proposing solution that will find the initial neural footprints and simultaneously update their shapes, based on the method explained in chapter 4.3. Also, instead of searching for the spiking vector for every pixel, we have solved the deconvolution problem directly in the spike domain for every individual neuron (ROI), by using the sparse approximation formalism (SAF), [110]. SAF can offer efficient, hardware friendly algorithms to solve (4.11).

4.4.3 Homotopy/LASSO/LARS Algorithm Design Consideration

As we mentioned before, the deconvolution problem (4.11) with a baseline offset included ($\|\mathbf{x}-\mathbf{c}\|^2 \rightarrow \|-\mathrm{b}\mathbf{1}^{\mathbf{T}}+\mathbf{x}-\mathbf{c}\|^2$) has a form of non-negative (constrained) LASSO problem. Often, this expression is presented in a modified form in the spike domain as

$$\min_{\mathbf{s}} \frac{1}{2\sigma^2}\|\mathbf{x}-\mathrm{G}^{-1}\mathbf{s}-\mathrm{b}\mathbf{1}^{\mathbf{T}}\|^2+\theta\|\mathbf{s}\|_1 \quad \text{subject to } \mathbf{s}\geq 0, \|\mathbf{x}-\mathrm{G}^{-1}\mathbf{s}-\mathrm{b}\mathbf{1}^{\mathbf{T}}\|^2 \leq \sigma^2\mathrm{T}. \tag{4.19}$$

The optimization problem (4.11), has a very similar structure to the Basis Pursuit Problem (BP) which can be efficiently solved by many linear programming algorithms. On the other hand, the work in [99] proposes a non-negative LARS algorithm for obtaining the solution of (4.19). In the case of very sparse spiking signals, LARS method is particularly efficient and the algorithm

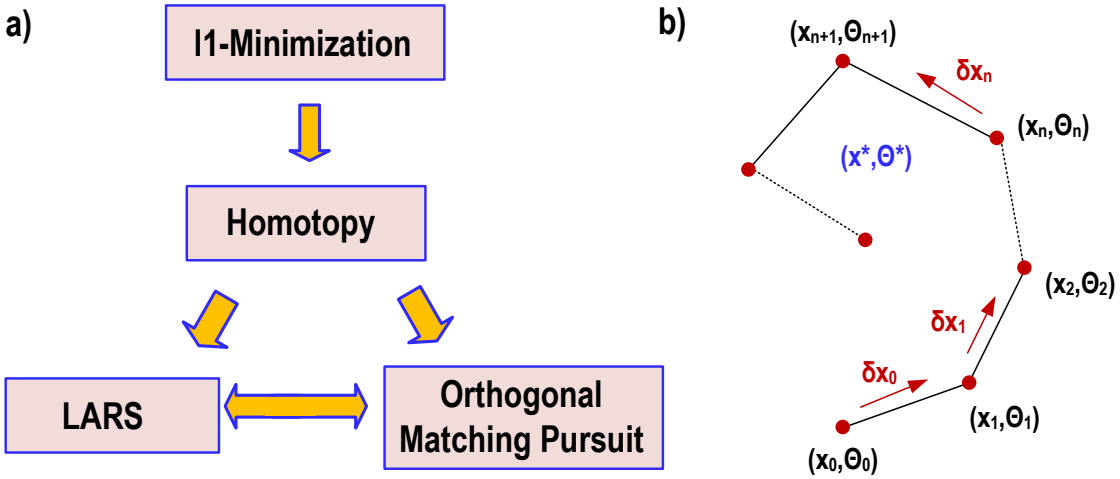


Fig 4.7: a) Relation between different l1-min algorithms b) Homotopy Path.

converges only after a few iterations. Furthermore, positive l1-min problems terminate much earlier and yield a more parsimonious solution than general l1-min problems, [111].

Both non-negative LASSO and non-negative LARS can be efficiently solved by employing a very fast algorithm called homotopy, which was initially proposed for solving noisy overdetermined l1-penalized least squares problems, [104], [113]. The Homotopy algorithm starts from an initial solution and converges along the so-called homotopy path, Fig. 4.7, by controlling a transformation parameter – the homotopy parameter. Below, we will briefly introduce LASSO/LARS homotopy algorithm that solves the problem in (4.20). To be consistent with the notation in literature, a constrained minimization problem is expressed as

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{\Phi x}-\mathbf{y}\|^2+\theta\|\mathbf{x}\|_1 \quad \text{subject to } \mathbf{x}\geq 0, \tag{4.20}$$

where Φ denotes the measurement matrix (dimension M x N), **y** is the set of measurements and **x** depicts the (sparse) unknow signal. Parameter Θ is the Lagrange multiplier that plays the role of the threshold parameter, that controls the tradeoff between the measurement fidelity and the sparsity of the solution.

Pseudo-code (PC) and description for the non-negative homotopy algorithm, [113]-[114], is given below.

1) **Initialize:** $\mathrm{k}=1, \mathrm{x}_0=\vec{0},\ \mathrm{corr}_0=\Phi^{\mathrm{T}}\mathrm{y}, \Theta_0=\|\mathrm{corr}\|_\infty, \Gamma_0=\mathrm{argmax_i}|\mathrm{corr(i)}|, \mathrm{r}_0=\mathrm{y}, \mathrm{d}=\vec{0}$

2) **Repeat:** $\Phi_{\Gamma_{\mathrm{k}-1}}^{\mathrm{T}}\Phi_{\Gamma_{\mathrm{k}-1}}\mathrm{d}(\Gamma_{\mathrm{k}-1})=\mathrm{sign}(\mathrm{corr}(\Gamma_{\mathrm{k}-1})),\ \mathrm{u}=\Phi_{\Gamma_{\mathrm{k}-1}}\mathrm{d}(\Gamma_{\mathrm{k}-1})$

3) $\mathrm{t}=\min_{\mathrm{i}\epsilon\Gamma_{\mathrm{k}-1}}\left(\frac{-\mathrm{x}_{\mathrm{k}-1}(\mathrm{i})}{\mathrm{d(i)}}\right), \delta^-=\mathrm{argmin}_{\mathrm{i}\epsilon\Gamma_{\mathrm{k}-1}}\left(\frac{-\mathrm{x}_{\mathrm{k}-1}(\mathrm{i})}{\mathrm{d(i)}}\right), \gamma^-=\max(\mathrm{t},0)$

4) $\delta^+=\mathrm{argmin}_{\mathrm{j}\epsilon\Gamma_{\mathrm{k}-1}^{\mathrm{c}}}\left(\frac{\theta_{\mathrm{k}-1}-\mathrm{corr(j)}}{1-\boldsymbol{\varphi}_{\mathrm{j}}^{\mathrm{T}}\mathrm{u}}\right), \gamma^+=\min_{\mathrm{j}\epsilon\Gamma_{\mathrm{k}-1}^{\mathrm{c}}}\left(\frac{\theta_{\mathrm{k}-1}-\mathrm{corr(j)}}{1-\boldsymbol{\varphi}_{\mathrm{j}}^{\mathrm{T}}\mathrm{u}}\right)$

5) **If** $(\gamma^-<\gamma^+)$

$$\gamma=\gamma^-, \Gamma_{\mathrm{k}}=\Gamma_{\mathrm{k}-1}\backslash\delta^-$$

else $\gamma = \gamma^{+}, \Gamma_k = \Gamma_{k-1} \cup \delta^{+}$

6) $\Theta_k = \Theta_{k-1} - \gamma, x_k = x_{k-1} - \gamma d, r_k = r_{k-1} - \gamma u, corr = \Phi^T r_k, d = \vec{0}$

7) Until: $\|r_k\|_2 \leq \epsilon \vee \Theta_k \leq \epsilon$.

Specifically, terms Γ and corr denote the support (active set - the index set of nonzero coefficients) of the vector x_k and the residual correlation vector, [104], [113]. The algorithm estimates solutions x_k in an iterative way starting with the initial solution $x_0 = \vec{0}$. It starts with a large value of Θ and decreases Θ down to the final value in a simple sequence of steps. As Θ shrinks, the current value x_k follows a piecewise-linear and polygonal path. The sign sequence and the active set of the solution dictate the direction and the length of every segment within the path. While jumping to the new vertex in the homotopy path, the algorithm either removes existing elements or adds new one to the active set Γ. Also, as explained in [113], direction update and the step size cause a one-element change in the active set Γ. These parameters can be estimated by employing optimality conditions, which are derived from the subdifferential of the objective function (4.20). Note that since the positivity constraint is added in (4.20), step 4 in the PC is slightly different than the one explained in [104], [113]. Since, we are always dealing with noisy data, the algorithm terminates as soon as the residual satisfies $\|r_k\|_2 \leq \epsilon$ or Θ has been lowered to its desired value. As Θ converges to 0, it is clear that PC provides the solution to the BPDN problem for all the values of ϵ, [114].

It has been shown that if the underlying solution of (4.20) has only k non-zeros elements, where k<<M, N, the homotopy method reaches the solution in only k iterations. Also, as it is pointed in [104], the LARS procedure is very similar to the homotopy method except the LARS omits the step that removes the variables from the active set and limits its procedure to the new-element addition only.

When the k-step solution property holds, the homotopy method converges to the solution for (4.20) much faster than general LP solvers. The bulk of the computational cost comes from computing $\Phi_{\Gamma_k}^T \Phi_{\Gamma_k}$ (solving the Q x Q linear system, where Q is equal to the size of the current active Γ in iteration k) and from computing vectors **d** and **u** in step 2 of PC, which are used to calculate the step-size.

To have an efficient implementation of the homotopy algorithm, Cholesky factorization is employed during the computation of $\Phi_{\Gamma_k}^T \Phi_{\Gamma_k}$ term, and during the active set update (addition/removal of the new element). If M~N, it can be shown (from [104]) that in the case of dense data (no sparsity constraints), k Homotopy steps need $\frac{4kM^2}{3} + kMN + \mathcal{O}(kN)$ flops (floating-point operations), which is significantly better than $\mathcal{O}(M^3)$ flops that is required for regular LP solvers. Furthermore, if the sparsity constraint holds, and k<<M~N, the homotopy gives more favorable estimates and roughly terminates in $k^3 + kMN$ flops. For comparison, using the least-square to solve the system $\Phi x = y$, we would require $2M^2N - 2M^3/3$ flops, [104]. It is important to mention that computational complexity does not translate linearly into hardware complexity since the VLSI implementation depends on the memory organization, data flow, scheduling, choice of architecture, etc.

4.4.4 Homotopy Algorithm Reformulation

To reduce the number of operations needed for the homotopy execution, we have adopted several mathematical transformations that directly reduce the algorithm complexity.

Since $A = \Phi_{\Gamma_k}^T \Phi_{\Gamma_k} \in R^{kxk}$ is a symmetrical, positive-definite matrix, as pointed in [114], [116], instead of using conventional Cholskey factorization method that involves square-root

operations in calculation of diagonal elements of L in (4.21), an alternative method is employed that avoids costly square-root operations by simple reformulation, [114].

$$A = LL^T = (LD^{-1})(DD)(D^{-1}L^T) = L'D'L'^T \tag{4.21}$$

Diagonal matrix $D \in R^{kxk}$ has all the square-rooted factors, while $L \in R^{kxk}$ is a lower-triangular matrix. It is easy to show that matrix D and L have the same main diagonals (diag(D)=diag(L)). Matrix $L' \in R^{kxk}$ is a lower-triangular matrix that has all diagonal elements equal to one, that satisfies $L' = LD^T$ and the matrix $D' \in R^{kxk}$ ($D' = D^2$) is a diagonal matrix that is free of square-roots.

Furthermore, when comes to removing/adding a new element from/into an active set, the general procedure (in terms of number of flops) can also be simplified. First, if we want to add a new element into the active set in iteration k, the matrix $A_k = \Phi_{\Gamma_k}^T \Phi_{\Gamma_k}$ can be constructed by simply adding a new column and row as pointed in (4.22) - (4.23). By employing the Cholesky decomposition of matrix A_k, it can be shown that the factorization elements in (4.21) can be updated in an incremental way, [114], with only a few simple, recursive operations per iteration.

$$\Gamma_k = \Gamma_{k-1} \cup \delta^+, \ \Phi_{\Gamma_k} = [\Phi_{\Gamma_{k-1}}, \varphi_\delta] \tag{4.22}$$

$$A_k = \begin{bmatrix} A_{k-1} & \Phi_{\Gamma_{k-1}}^T \varphi_\delta \\ \varphi_\delta{}^T \Phi_{\Gamma_{k-1}} & \varphi_\delta{}^T \varphi_\delta \end{bmatrix} \tag{4.23}$$

On the other hand, removing a bad atom from the active set is slightly more complicated procedure which will be explained in detail. By deleting a row i from a matrix Φ_Γ, we are effectively deleting the row and column i from the matrix A. Let

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & A_{22} & A_{23} \\ A_{13}^T & A_{23}^T & A_{33} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & l_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix} \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & l_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix}^T \tag{4.24}$$

be the Cholesky factorization of matrix A and let

$$A' = \begin{bmatrix} A_{11} & 0 & A_{13} \\ 0 & 0 & 0 \\ A_{13}^T & 0 & A_{33} \end{bmatrix} = \begin{bmatrix} L_{11}' & 0 & 0 \\ L_{21}' & l_{22}' & 0 \\ L_{31}' & L_{32}' & L_{33}' \end{bmatrix} \begin{bmatrix} L_{11}' & 0 & 0 \\ L_{21}' & l_{22}' & 0 \\ L_{31}' & L_{32}' & L_{33}' \end{bmatrix}^T \tag{4.25}$$

be the Cholesky factorization of matrix A' obtained from matrix A by zeroing the i-th column and row. Then it is easy to prove that $L_{11}' = L_{11}$, $L_{21}' = 0$, $L_{31}' = L_{31}$, $l_{22}' = 0$, $L_{32}' = 0$ and

$$A_{33} = L_{31}L_{31}^T + L_{32}L_{32}^T + L_{33}L_{33}^T = L_{31}L_{31}^T + L_{33}'L_{33}'^T \tag{4.26}$$

$$L_{33}'L_{33}'^T = L_{32}L_{32}^T + L_{33}L_{33}^T. \tag{4.27}$$

Since L_{32} is a vector, (4.27) gives a rank-1 update of the Cholesky factorization $L_{33}L_{33}^T$. Removing the i-th row and column of A' is trivial now. Rank-1 update of the Cholesky factorization is a standard task, [117], and can be done in $\mathcal{O}(d^2)$ operations, where d is the size of L_{33}. It can be shown that if $A_k = \Phi_{\Gamma_k}^T \Phi_{\Gamma_k}$ is dense and we want to remove the i-th row from Φ_{Γ_k}, the total number of flops is equal to $2(k-i)^2 + 5(k-i) + 1$ (roughly k^2 if i=1 and $k^2/2$ if i=n/2). In the case of sparse matrix Φ_{Γ_k} (our case), the total number of flops is a small fraction of the flop count needed for the dense matrix row removal.

Following the procedure explained in [114], after new element addition/removal, a residual update can be also done in an incremental way; first, the direction d is computed by calculating the normal equation from step 2 of PC, and then the residual r_k is updated based on the previous values r_{k-1} and u_{k-1}.

Going back to our case, we can conclude that in order to infer the spiking vector **s**, matrix G^{-1} in (4.19) plays the role of matrix Φ in (4.20) and PC for homotopy. Note that matrix G is sparse, and has only 3 non-zero diagonals (4.12). If the order of AR process is 1, it is trivial to show that

$$G^{-1} = \begin{pmatrix} 1 & 0 & 0 & . & . & . & 0 \\ \gamma_1 & 1 & . & . & . & . & 0 \\ \gamma_1^2 & \gamma_1 & 1 & . & . & . & 0 \\ . & \gamma_1^2 & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ \gamma_1^{T-1} & . & . & . & \gamma_1^2 & \gamma_1 & 1 \end{pmatrix}, \tag{4.28}$$

where T is the number of timesteps. If p=2, we can use simple polynomial factorization. If $G = (I + Tr)$, where Tr is the strictly lower triangular matrix (the main diagonal contains all zeros), we can get a compact form for G^{-1} by using the following identity,

$$G^{-1} = (I + Tr)^{-1} = I + \sum_{j=1}^{T-1}(-1)^j Tr^j. \tag{4.29}$$

Since matrix Tr has a trivial form, by employing the binomial theorem, it is easy to show that matrix G^{-1} can be expressed as

$$G^{-1} = \begin{pmatrix} 1 & 0 & 0 & . & . & . & 0 \\ \gamma_1 & 1 & . & . & . & . & 0 \\ \gamma_1^2 + \gamma_2 & \gamma_1 & 1 & . & . & . & 0 \\ \gamma_1^3 + 2\gamma_2\gamma_1 & \gamma_1^2 + \gamma_2 & . & . & . & . & . \\ . & \gamma_1^3 + 2\gamma_2\gamma_1 & . & . & . & . & . \\ . & . & . & . & . & . & . \\ \gamma_1^{T-1} & . & . & \gamma_1^3 + 2\gamma_2\gamma_1 & \gamma_1^2 + \gamma_2 & \gamma_1 & 1 \end{pmatrix}, \tag{4.30}$$

where all sub-diagonal elements are calculated by combining the coefficients in binomial expansion.

Note that the optimization problem (4.19) is not underdetermined - the vector of observation and the spiking vector are of the same length T (number of timesteps). Hence, since the sampling rate is equal to one, the homotopy algorithm achieves very good spiking signal recovery performance. Its reconstruction signal-to-noise ratio ($RNSR = 20\log\left(\frac{\|s\|_2}{\|s-\bar{s}\|_2}\right)$) goes above 90%, [102].

Simplifications introduced in the PC steps of the Homotopy algorithm will result in significant reduction in the number of operations needed per iteration. These modifications on the algorithm level are necessary to relax the overall computational requirements and to improve the system throughput. Hence, only after the algorithm is decomposed into hardware-friendly tasks, we can go into the architectural design. However, homotopy algorithm mapping into hardware was not discussed here and it is left for the future work.

Flexible and efficient hardware design of the homotopy engine imposes different types of challenges – reconfigurability and high parallelism of the processing elements, efficient memory control schemes, resource sharing, etc. The goal of the future work is to map the homotopy algorithm into a dedicated hardware unit so that we can parallelize the deconvolution method to a large extent. Basically, every detected neuron and its fluorescence trace will allocate a specialized unit (homotopy/LARS) that will perform spiking signal extraction in real-time.

4.5 Simulations results

To show the performance of the proposed processing technique, we have simulated our approach with real data (single-photon Ca^{+2}-based imaging, [78]) and compared the results with "ground-truth" spiking activities that were obtained by employing Paninski/ Pnevmatikakis online-toolboxes (MATLAB) available online, [87]. In this particular example, the dimension of the frame is 512x768 (W x L) and the video sequence contains 1000 frames while it runs at a speed of f = 20fps.

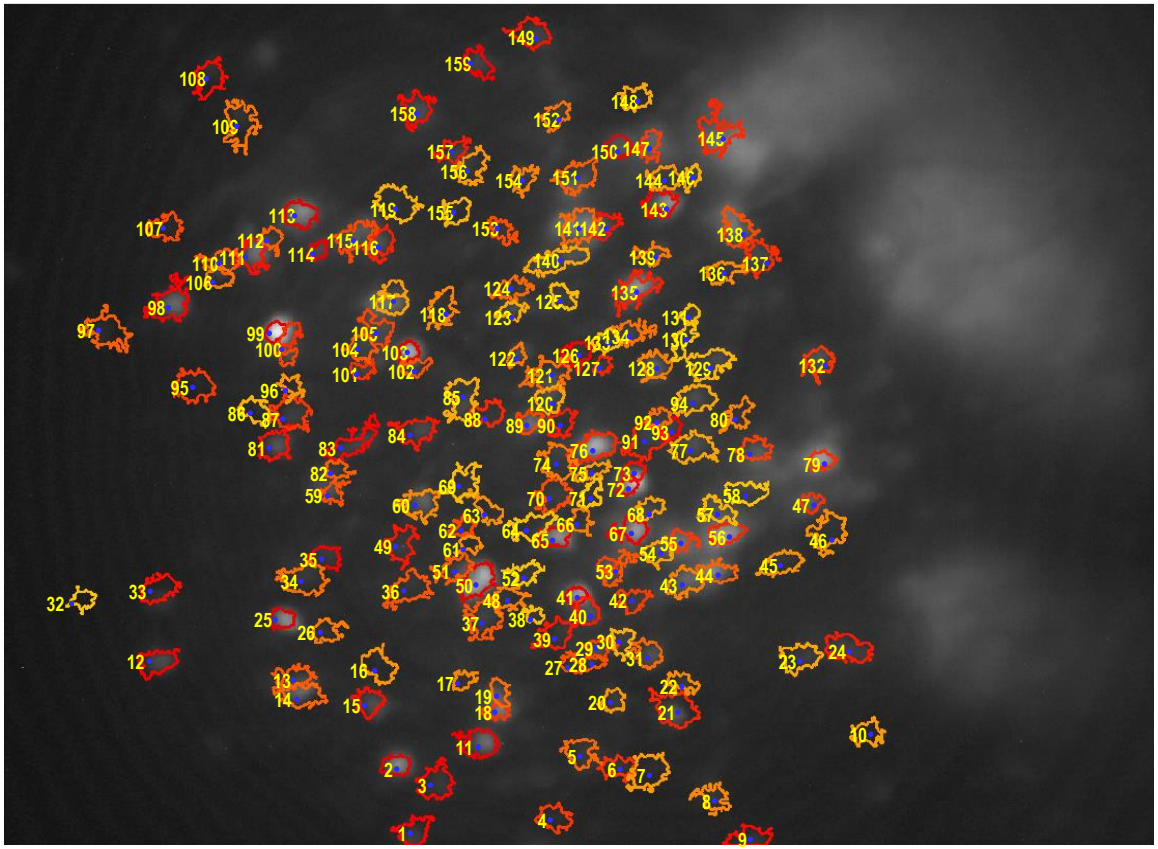


Fig 4.8: Detected Neurons – their spatial footprints.

Motion Correction Block is enabled during the whole video sequence – the frames are received, line-by-line and the unit automatically performs alignment as shown in section 4.2. In the initial phase, the frame sequence that contains T =256 frames, is used for the Neuron Detection. Figure 4.8 shows the detected neurons in a single frame as the output result of this phase. As we can see, our method provides compact spatial footprint estimates and separates neurons ROIs even in the case of significant spatial overlap. The video used in this simulation had a focused field of view, which resulted in a smaller number of detected neurons, while the neuron's spatial footprint occupies more pixels on average. The same memory budget can be used for videos that contain tens of thousands of cells. The number of pixels per neuron, in that case, would be much smaller.

As a comparison, an offline-method was able to find 161 ROIs for the thousand-frame long sequence.

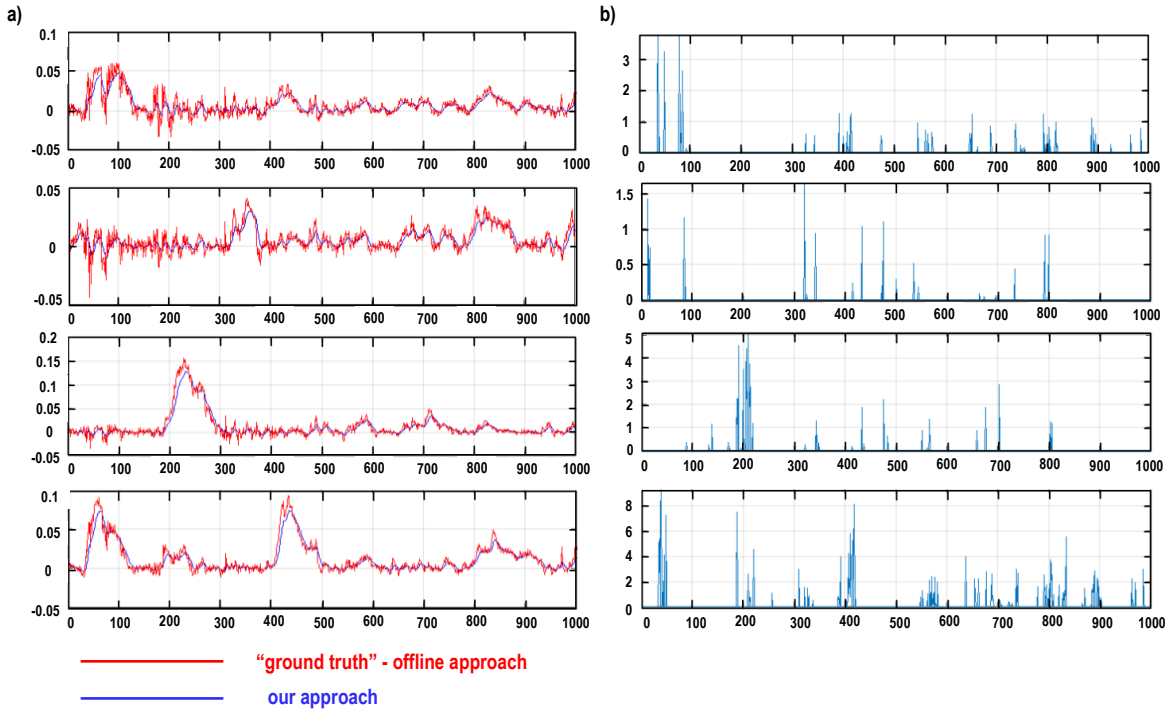


Fig 4.9: a) Extracted Temporal Traces (**ΔF/F**) b) Extracted Spiking Signal (**s**).

As the most memory expensive steps, we have evaluated the resources needed for (MSER&UF)-based neuron detection, while the frame alignment is active. Every new frame is scanned from top to bottom and (MSER&UF) simultaneously works with two strips of lines – while we are processing the first one, the other is simultaneously loaded into the internal memory. Processing includes calculation of integral projections, motion estimation, motion correction and MSER&UF operations. Following the distributive approach, if the frame is partitioned into N_{STRIP} strips and every strip is divided into N_{BL}, it is not difficult to show that the memory (in bytes) needed for these steps can be approximated with

$$MEM_{Req} = 2(N_{BL}(1.25 + 0.5(\log_2 R_{BL}) + 64\log_2 R_{BL} + R_{BL}) + IP_{MEM} + CORR_{mem} \quad (4.31)$$

where R_{BL} is the resolution of the individual block, IP_{MEM} is the memory needed for IP calculation and $CORR_{mem}$ defines the memory requirement for the correction operation. Details for (4.31) are

omitted, but one can refer to [94], [95] and section 4.2 for more thorough explanations. Note that the number of cycles needed for the completion of these tasks is proportional to the number of strips N_{STRIP}. If N_{STRIP} = 16 and N_{BL}= 12, total memory needed for Motion Correction and Neuron Detection is estimated to be 400kB, while the video can run as fast as 200fps.

After the Detection step is completed, most of the memory space (Region Map, Hash-tag memory, auxiliary memory banks) that was employed, is freed and reused in the deconvolution method. The regions of interest are stored in a separate memory bank together with the corresponding fluorescence traces. The Neuron Detection Block is disabled and the Decovolution Unit that analyses the fluorescence traces and takes advantage of the spatio-temporal data structure, is enabled till the end of the sequence. We demonstrated the effectiveness of our estimation method on real calcium imaging data based on the deconvolution procedure explained in section 4.4.

The Deconvolution method extracts the spiking signal, (4.19), based on the fluorescence traces that are packetized in groups of 256. Since the number of timesteps in one packet is equal to T=256, the Deconvolution Unit sends the results at the output with latency Lat =T/f.

The proposed method extracted the spiking signals by employing highly parallelized homotopy solvers (multiple homotopy engines), while achieving 100x data reduction and real-time frame processing. Figure 4.9 shows the extracted temporal traces and corresponding spiking signals for 4 randomly chosen neurons (ROI). A commonly used metric for representation of the temporal traces is $\Delta F/F$ that is defined as

$$\Delta F/F(t) = \frac{\int_0^t R(t-\tau)w(\tau)d\tau}{\int_0^t w(\tau)d\tau}, \tag{4.32}$$

where $w(\tau) = e^{-|\tau|/\tau_0}$ and R(t) captures the relative change in fluorescence from F(t) and $F_0(t)$ and can be expressed as

$$R(t) = \frac{F(t) - F_0(t)}{F_0(t)}. \tag{4.33}$$

The term $F_0(t)$ denotes baseline noise and $F(t)$ is the mean fluorescence of a neuron's ROI at the t-th timestep. From Fig. 4.9, we see close match of extracted signals, while the accuracy of deconvolved spiking signals is at the satisfying level. However, the offline method would need about 45 minutes for data processing, while our approach extracts the results in real-time with latency Lat. Table 4.1 summarizes the benefits of the proposed method and compare it with the state-of-the-art. The processing resources that were used are very modest, while we achieved drastic data reduction and reduction in the computational complexity. Most importantly, we have shown that the modified nonnegative matrix factorization formalism that efficiently distinguishes the overlapping neural sources and directly models the Ca^{+2}-indicator dynamics, can be implemented in real-time, and is a promising and suitable tool for large brain-data processing.

Table 4.1: Comparison between Electrophysiological and Optical Approach.

Reference	Electrode-based	Optical	
		state-of-the-art	This work
Coverage Nº of Neurons	Up to 200	10^3-10^5	
Processing	X	Offline	Real-Time
Memory Requirement	X	10-100GB	0.4MB
Fully invasive	YES	NO	NO
Data Reduction	X	1x	100x

CHAPTER 5

Contributions and Future Work

5.1. Summary of Research Contributions

The goal of this research is to tackle several different problems that will enhance the field of biomedical applications. By employing low-power, flexible energy processing techniques we pave the road to the fully integrated self-powered sensors. Furthermore, we have developed an implant-scale, closed-loop neuromodulation interface that offers superior performance, power efficiency and unmatched level of electronic miniaturization. This dissertation also presents a new recording paradigm that allows real-time data processing from large neural ensembles at the resolution of individual neurons. Several main contributions presented in this research are:

• Design and demonstration of a thin-film, array-based Thermo-Electric Harvesting platform with a surface area of 0.83cm^2 that is made of biocompatible materials, mitigates the heat leakage and can autonomously supply energy processing IC. The platform meets the stringent anatomical and biophysical confinements of living subjects.

• Proposed and demonstrated the inductive-load ring oscillator (ILRO) architecture that can be triggered with very low input voltages. The ILRO was employed as a startup circuit - the minimum voltage needed to start the oscillation was measured to be 65 mV. That was the first fully-integrated solution for the cold startup in thermal harvesting applications.

• Developed a single-inductor topology with integrated 2-phase, analog maximum power point tracking (MPPT) unit. Integration of startup circuit and MPPT unit was done in 65nm technology

and our solution resulted in a 68% peak end-to-end efficiency (92% converter efficiency) and less than 20ms MPP tracking time.

• As a proof of concept, an in-vivo test was conducted - a 645µW regulated output power (effective 3.5K of temperature gradient) was harvested from a rat implanted with our harvesting system, demonstrating true energy independence in a real environment while showing a 7.9x improvement in regulated power density compared to the state-of-the-art.

• Design of two HV 180nm ICs (4-channel and 64-channel respectively) as a part of Restoring Active Memory (RAM) project. The IC includes 4 Macro and 4 Micro stimulation engines for macro and micro types of electrodes, integrated power management unit (PMU), multiplexers for spatial selection and access to sensing IC, etc. PMU is designed for the wireless power transfer and features active rectifier, high-voltage generators (HVG), LDOs etc.

• Development of a highly programmable implantable power management unit that can process multiple input power deliveries on-the-chip. Unit is able to process wireless power, power delivered through wires and power from/to rechargeable battery. This MIMO Management System significantly extends the range of biomedical applications for the implant.

• Development of reconfigurable active rectifier (AR) for wireless power transfer (WPT), wherein the AR operates in a Regular Mode and a Charging Mode, wherein the AR-WPT includes an adaptive load control (ALC) unit that accommodates power delivery with load requirements, wherein the ALC unit keeps the AR voltage at a desired value. As a part of AR-WPT, we proposed an adaptive ON/OFF delay compensation schemes for both types of active diodes (P and N) that by employing feedback generates in real-time offset currents to compensate switch delays.

Proposed circuit schemes showed improvements in PCE (PCE > 90% - 12% and 10% improvement at light and heavy load, respectively) across a wide loading range, while ensuring that the wireless power link delivers a stable voltage to the implant across load and coupling variations. Also, the ALC unit implementation allowed static current reduction.

• Development of a programmable electrode agnostic stimulation engine (SE) for the implantable neuromodulation systems. The SE features a high output impedance current source and current sink in order to support different types of electrodes and a wide range of stimulation currents. In the core of a stimulation engine (SE) is a precise, high-compliance and ultra-high output impedance current mirror for the source/sink part of the SE. Furthermore, high-voltage adaptive generators (V_{dd}/V_{ss}) are provided to accommodate voltage drops across high electrode impedances and to additionally save the power during the stimulation. The SE is designed primarily to enable simultaneous, multichannel, differential stimulation that is necessary to achieve concurrent stimulation and sensing in the neuromodulation systems.

• Prototyped two different STIM/PM ICs in HV 180 nm technology – the first IC has 4 Stimulation Engines (SE), can drive 32 stimulation cites and V_{DD}/V_{SS} absolute maximum is set to 7.5V/-7.5V. The second IC houses 8 SE that can drive 64 stimulation cites and can be individually programmed for monopolar/ differential stimulation. The SE current covers the range from 20uA to 5.1mA with 20uA step, while V_{DD}/V_{SS} are programmable with the absolute maximum set to 5V/-5V. HV STIM multiplexers provide a complex spatial resolution. Further, we have designed two low-profile NM units (for each version of IC) that occupy 135mm^3 and 338mm^3 of volume. These high-channel count, closed-loop neuromodulation units present the next generation neural interfaces, that is minimally invasive, and address the demands for limited area and power. We also, have

demonstrated a real-time, full duplex communication during concurrent stimulation and recording of neural signals.

• Developed methodology for a new neural recording paradigm based on the fast calcium imaging. Proposed hardware-friendly approach allows analysis of large neural ensembles in a single pipeline and in real-time, while relaxing the memory and computational requirements.

• Developed a Matlab model that implements the Motion Correction and Blind Neuron Detection steps for the fast calcium imaging, by employing modified computer vision algorithms such as Maximally Stable Extremal Regions and Template Matching. The model abandoned frame-level processing and adopted the distributed approach.

• Introduced algorithm modifications into the deconvolution step, that exploit the sparse nature of neurons and spiking signals both in spatial and time domain. The proposed simplifications allow the design of specialized dedicated units that map the Sparse Approximation algorithm into hardware. This would lead to an extraction of spikes at the resolution of individual neurons at real-time and 100x data reduction.

5.2. Looking to the Future

The work presented in this dissertation has provided solutions for a variety of problems related to the biomedical applications. Further research is going to be continued and upgraded in many ways. The next step is related to the work explained in the chapter 3 - verification and validation of the closed-loop, implant-scale NM interface in humans. Apart from the hardware characterization during the in-vivo tests, we would also follow the impact that our miniaturized implant would have in diagnostics and therapy of neurological disorders in the upcoming years.

The research presented in the chapter 4 requires a lot of evaluation of hardware feasibility and it is going to be continued. Building application-specific dedicated units/kernels for big data analysis is an open research area. Our work will enable high performance processing and lay the foundation for real-time brain decoding on a large scale. Efficient hardware mapping of the sophisticated algorithms is necessary to allow complete system deployment onto wearable platforms and its integration with the fluorescent sensor.

Although we have proposed some algorithm simplifications and simulated for their impact, more careful analysis and flexible VLSI implementation are needed to verify the functionality of the system, to evaluate the accuracy of signal recovery and to estimate the overall power consumption. The first step is development of a dedicated accelerator for the homotopy algorithm. Implementation of a such unit needs joint algorithm-architecture consideration – their evaluation and optimization in a separate manner is inefficient and deteriorate performance and flexibility-efficiency trade-off. The Homotopy accelerator engine should feature high parallelism and configurability so that the algorithm can reuse the hardware resources and have efficient access to the memory banks. This would improve the area efficiency and the throughput.

Complete system-level integration of the processing chain is the next step to demonstrate the benefits of real-time processing for ultra-fast Ca^{+2} imaging. Such a system shall embed a specialized controller for image alignment that was described in section 4.2, and an on-chip MSER detector for simultaneous ROI sensing and updates - section 4.3. Also, hardware implementation includes deployment of the processing threads for the manipulation of results and synchronization. Furthermore, since many tasks down the processing chain are done in sequential order, we can employ massive computing and memory resource sharing. Lastly, flexible hardware implementation has to be followed with MAC layer software development, so that after data acquisition/processing flow, the information is available to the user.

Enabling real-time recording from large neural ensembles would significantly improve our understanding of brain dynamics and allow closed-loop experiments for calcium imaging. Collecting the signals from thousands and tens of thousands of neurons at the resolution of individual neurons, and their simultaneous decoding will enhance the research capabilities of brain-computer interfaces.

REFERENCES

[1] I. Doms, P. Merken, C. V. Hoof, R. P. Mertens, "Capacitive Power Management Circuit for Micropower Thermoelectric Generators with a 1.4μA Controller," *IEEE J. Solid-State Circuits*, vol. 44, no. 10, pp. 2824–2833, Oct. 2009.

[2] E. J. Carlos, K. Strunz, B. P. Otis, et al., "A 20mV Input Boost Converter with Efficient Digital Control for Thermoelectric Energy Harvesting" *IEEE J. Solid-State Circuits*, vol. 45, no. 4, Apr. 2010.

[3] Y.K. Ramadass, A. P. Chandrakasan, et al., "A Battery-Less Thermoelectric Energy Harvesting Interface Circuit with 35mV Startup Voltage," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, Jan. 2010.

[4] J.-P. Im, S.-W. Wang, S.-T. Ryu, G.-H. Cho, "A 40 mV Transformer-Reuse Self-Startup Boost Converter with MPPT Control for Thermoelectric Energy Harvesting," *IEEE J. Solid-State Circuits*, vol. 47, no. 12, Dec. 2012.

[5] H.-Y Tang, P.-S. Weng, P.-C. Ku, L.-H. Lu, "A Fully Electrical Startup Batteryless Boost Converter with 50mV Input Voltage for Thermoelectric Energy Harvesting," *IEEE Symp. VLSI Circuits Dig. Tech. Papers*, pp. 196-197, 2012.

[6] P.-H. Chen, K. Ishida, K. Ikeuchi, X. Zhang, K. Honda, Y. Okuma, Y. Ryu, M. Takamiya, T. Sakurai, et al., "Startup Techniques for 95 mV Step-Up Converter by Capacitor Pass-On Scheme and V_{TH}-Tuned Oscillator with Fixed Charge Programming," *IEEE J. Solid-State Circuits*, vol. 47, no. 5, May. 2012.

[7] A. Shrivastava, N. E. Roberts, O. U. Khan, D. D. Wentzloff, B. H. Calhoun, "A 10 mV-Input Boost Converter with Inductor Peak Current Control and Zero Detection for

Thermoelectric and Solar Energy Harvesting with 220 mV Cold-Start and 14.5 dBm, 915 MHz RF Kick-Start," *IEEE J. Solid-State Circuits*, vol. 50, no. 8, May. 2015.

[8] T. Torfs, V. Leonov, R. F. Yazicioglu, P. Merken, C. V. Hoof, R. J. M. Vullers, B. Gyselinckx, "Wearable autonomous wireless electroencephalography system fully powered by human body heat," *IEEE Sensors*, pp. 1269–1272, Oct. 2008.

[9] W. Jung, S. Oh, S. Bang, Y. Lee, D. Sylvester, and D. Blaauw, "A 3 nW fully integrated energy harvester based self-oscillating switched capacitor DC-DC converter," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2014, pp. 398–399.

[10] M. B. Machado, M. C. Schneider, C Galup-Montoro, "On the Minimum Supply Voltage for MOSFET Oscillators," *IEEE Transaction on Circuits and Systems I: Regular Papers*, vol. 61, no. 2, pp. 347-357, Feb. 2013.

[11] C. Enz, F. Krummenacher, E. Vittoz, "An analytical MOS transistor model valid in all regions of Operation and dedicated to low-voltage and low-current applications", *Journal on Analog Integrated Circuits and Signal Processsing*, Kluwer Academic Publishers, pp. 83-114, July 1995.

[12] J. D. Meindl and A. J. Davis, "The fundamental limit on binary switching energy for terascale integration (TSI)," *IEEE J. Solid-State Circuits*, vol. 35, no. 10, pp. 1515–1516, Oct. 2000.

[13] H.-M. Lee, M. Ghovanloo, "An Integrated Power-Efficient Active Rectifier with Offset-Controlled High Speed Comparators for Inductively Powered Applications", *IEEE Trans. Circuits and Systems—I: Regular Papers,* vol. 58, no. 8, Aug. 2011.

[14] C.–S. A. Gong," An active-diode-based CMOS rectifier for biomedical power harvesting applications", *International Journal of Circuit Theory and Applications*, 2011; 39(5):439–449.

[15] C. van Liempd, S. Stanzione, Y. Allasasmeh, C. van Hoof, "A 1μA-to-1mA energy-aware interface IC for piezoelectric harvesting with 40nA quiescent current and zero-bias active rectifiers" *IEEE ISSCC Dig. Tech. Papers*, Feb. 2013, pp. 76–77.

[16] E. Rogers," Understanding boost power stages in switch mode power supplies", Application report, Texas Instrument, 1999

[17] D. Rozgić; D. Marković, "A 0.78mW/cm^2 autonomous thermoelectric energy-harvester for biomedical sensors", *Symposim on VLSI Circuits*, June 2015, pp. 278–279.

[18] R.W. Erickson, D. Maksimović, "Fundamentals of Power Electronics," 2nd Ed, Springer, 2001.

[19] Laird Technology, <http://lairdtech.com>, eTEG Series PG37.

[20] A. M. Niknejad, H. Hashemi, "mm-Wave Silicon Technology 60 GHz and Beyond," New York, NY, USA: Springer, 2008.

[21] S. Bandyopadhyay, P.P. Mercier, A.C. Lysaght, K.M. Stankovic, A.P. Chandrakasan, "A 1.1 nW Energy-Harvesting System with 544 pW Quiescent Power for Next-Generation Implants," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 49, no. 12, pp 2812-2824, Dec. 2014.

[22] P.-H. Chen and P.-Y. Fan, "An 83.4% peak efficiency single-inductor multiple-output based adaptive gate biasing DC-DC converter for thermoelectric energy harvesting," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 62, pp. 405–412, Feb 2015.

[23] C. Veri, L. Francioso, M. Pasca, C. De Pascali, P. Siciliano, and S. D'Amico, "An 80 mV Startup Voltage Fully Electrical DC–DC Converter for Flexible Thermoelectric Generators", *IEEE Sensors Journal*, vol. 16, no. 8, April 15, 2016.

[24] S. Carreon-Bautista, L. Huang and E. S. Sinencio, "An Autonomous Energy Harvesting Power Management Unit with Digital Regulation for IoT Applications," *IEEE J. Solid-State Circuits*, vol. 51, no. 6, June 2016.

[25] J. Kim, P. K. T. Mok and C. Kim, "A 0.15 V Input Energy Harvesting Charge Pump with Dynamic Body Biasing and Adaptive Dead-Time for Efficiency Improvement*", IEEE J. Solid-State Circuits*, vol. 50, no. 2, Feb. 2015.

[26] A. Zurbuchen, A. Haeberlin, A. Pfenniger, L. Bereuter, J. Schaerer, F. Jutzi, C. Huber, J. Fuhrer and R. Vogel "Towards Batteryless Cardiac Implantable Electronic Devices—The Swiss Way", *IEEE Trans. Biomedical Circuits and Systems*, to appear.

[27] X. Wang, D.Wu, F. Qiao, P. Zhu, K. Li, L. Pan, R. Zhou, "A High Efficiency CMOS Charge Pump for Low Voltage Operation," ASICON '09. *IEEE 8th International Conference*, 2009.

[28] S. Lineykin, S. Ben-Yaakov, "Modeling and Analysis of Thermoelectric Modules," *IEEE Trans. on Industry Application*, vol. 43, no. 2, March. 2007.

[29] A. Shrivastava, Y. K. Ramadass, S. Khanna, S. Bartling, B. H. Calhoun, "A 1.2 W SIMO energy harvesting and power management unit with constant peak inductor current control achieving 83–92% efficiency across wide input and output voltages," in *Symp. VLSI Technology and Circuits*, 2014, pp. 1–2.

[30] E. E. Aktakka, K. Najafi, "A Micro Inertial Energy Harvesting Platform with Self-Supplied Power Management Circuit for Autonomous Wireless Sensor Nodes," *IEEE Journal of Solid-State Circuits,* vol. 49, no. 9, Sep. 2014.

[31] M. B. Machado, M. C. Schneider, C Galup-Montoro, "Analysis and Design of Ultra-Low-Voltage Inductive Ring Oscillators for Energy-Harvesting Applications," *IEEE Fourth Latin American Symposium on Circuits and Systems,* (LASCAS), March. 2013.

[32] P. Feng, Z. Yiqi, L. Xiaoming, "A high efficiency charge pump circuit for low power applications," *Journal of Semiconductors,* 2010, 31(1): 015009.

[33] A. E. Mendrela et al., "A Bidirectional Neural Interface Circuit with Active Stimulation Artifact Cancellation and Cross-Channel Common-Mode Noise Suppression," *JSSC, Apr.* 2016, pp. 955-965.

[34] Y.-K. Lo, et. al, "A 176-Channel 0.5cm^3 0.7g Wireless Implant for Motor Function Recovery after Spinal Cord Injury," *ISSCC 2016,* pp. 382-383.

[35] W.-M. Chen, et. al., "A Fully Integrated 8-Channel Closed-Loop Neural-Prosthetic SoC for Real-Time Epileptic Seizure Control," *ISSCC 2013,* pp. 286-287.

[36] Rikky Muller, et al., 'A Minimally Invasive 64-Channel Wireless μECoG Implant', *JSSC,* Jan. 2015, 344-359.

[37] Y. P. Lin et al., "A Battery-Less, Implantable Neuro-Electronic Interface for Studying the Mechanisms of Deep Brain Stimulation in Rat Models," *IEEE TBioCAS,* 2015, pp 98-112.

[38] Cong, P. et al, "A 32-Channel Modular Bi-directional Neural Interface System with Embedded DSP for Closed-Loop Operation," *ESSCIRC, 2014.*

[39] B. C. Johnson , et al. ,"An Implantable 700μW 64-Channel Neuromodulation IC for Simultaneous Recording and Stimulation with Rapid Artifact Recovery", VLSI, June 2017.

[40] W. Jiang, *et al.,* "A ±50mV Linear-Input-Range VCO-Based Neural-Recording Front-End with Digital Nonlinearity Correction", *ISSCC 2016,* pp 484-485.

[41] K. Roach, "Electrochemical models for electrode behavior in retinal prostheses," Master's thesis, Massachusetts Institute of Technology, USA, 2003.

[42] S. Kelly et al., "A power-efficient voltage-based neural tissue stimulator with energy recovery," in *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. IEEE, 2004*, pp. 228-524.

[43] H.-M. Lee, K.-Y. Kwon, W. Li, M. Ghovanloo, "A power-efficient switched-capacitor stimulating system for electrical/optical deep-brain stimulation", *IEEE Int. Solid State Circuits Conf. Dig. Tech. Papers*, pp. 414-415, Feb. 2014.

[44] M. H. Maghami, A. M. Sodagar, and M. Sawan, "Analysis and design of a high-compliance ultra-high output resistance current mirror employing positive shunt feedback," Int. J. Circuit Theory Appl., vol. 43, no. 12, pp. 1935–1952, Dec. 2015.

[45] H. Chandrakumar, et. al., "A 2.8μW 80mVpp-linear-input-range 1.6GΩ-input impedance bio-signal chopper amplifier tolerant to common-mode interference up to 650mVpp," *ISSCC 2017*, pp. 448-449.

[46] S. Guo, et al., "An efficiency-enhanced CMOS rectifier with unbalanced-biased comparators for transcutaneous-powered high-current implants," *IEEE J. Solid-State Circuits*, vol. 44, no. 6, pp. 1796–1804, Jun. 2009

[47] S. B. Lee, et al., "An inductively powered scalable 32-channel wireless neural recording system-on-a-chip for neuroscience applications," *IEEE Trans. Biomed. Circuits Syst.*, vol. 4, no. 6, pp. 360–371, Dec. 2010

[48] Y. Lu, et al., "A 13.56 MHz CMOS active rectifier with switched offset and compensated biasing for biomedical wireless power transfer systems," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 3, pp. 334–344, Jun. 2014.

[49] Y. Lu, et al., "A 13.56 MHz fully integrated 1X/2X active rectifier with compensated bias current for inductively powered devices," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2013, pp. 66–67

[50] C.-Y. Wu, et al, "A 13.56 MHz 40mW CMOS high-efficiency inductive link power supply utilizing on-chip delay-compensated voltage doubler rectifier and multiple LDOs for implantable medical devices," *IEEE J. Solid-State Circuits*, vol. 49, no. 11, pp. 2397–2407, Nov. 2014.

[51] X. Li, et al, "A 13.56 MHz Wireless Power Transfer System with Reconfigurable Resonant Regulating Rectifier and Wireless Power Control for Implantable Medical Devices," *IEEE J. Solid-State Circuits*, vol. 50, no. 4, Apr. 2015.

[52] C. Huang, et al, "A Near-Optimum 13.56 MHz CMOS Active Rectifier with Circuit-Delay Real-Time Calibrations for High-Current Biomedical Implants," *IEEE J. Solid-State Circuits*, vol. 51, no. 8, Aug. 2016.

[53] L. Cheng, et al, "Adaptive On/Off Delay-Compensated Active Rectifiers for Wireless Power Transfer Systems," *IEEE J. Solid-State Circuits*, vol. 51, no. 3, Mar. 2016.

[54] G. Wang, et al, "Design and analysis of an adaptive transcutaneous power telemetry for biomedical implants ," *IEEE Trans. On Circuits and Sustems I: Regular Papers.*, vol. 52, no. 10, Oct. , 2005.

[55] K. Chen, et al, " A system Verification Platform for High-Density Epiretinal Prosthesis," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 3, Jun. , 2013.

[56] R. R. Harrison, et al, “A Low Power Integrated Circuit for a Wireless 100-Electrode Neural Record,” *IEEE J. Solid-State Circuits*, vol. 42, no. 1, Jan. 2007.

[57] J. Pan, A. Abidi, D. Rozgić, H. Chandrakumar, D. Marković, “An Inductively-Coupled Wireless Power Transfer System that is Immune to Distance and Load Variations,” in *Proc. IEEE International Solid-State Circuits Conference* (ISSCC'17), San Francisco, CA, USA, 2017, pp. 382-383.

[58] C.-H. Lin, et al, “A Li-Ion Battery Charger With Smooth Control Circuit and Built-In Resistance Compensator for Achieving Stable and Fast Charging,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 57, no. 2, Feb. , 2010.

[59] H. Lee and P. K. T. Mok, “Switching noise and shoot-through current reduction techniques for switched-capacitor voltage doubler,” *IEEE J. Solid-State Circuits*, vol. 40, no. 5, pp. 1136–1146, May 2005.

[60] J. F. Dickson, “On-chip high-voltage generation in MNOS integrated circuits using an improved voltage multiplier technique,” *IEEE J. Solid State Circuits*, vol. SC-11, no. 3, pp. 374–378, Jun. 1976.

[61] J. Wibben and R. Harjani, “A High-Efficiency DC/DC Converter Using 2 nH Integrated Inductors,” *Journal of Solid-State Circuits, IEEE*, vol. 43, no. 4, pp. 844 – 854, 2008.

[62] Y.-C. Huang, M.-D. Ker, C.-Y. Lin, "Design of negative high voltage generator for biphasic stimulator with soc integration consideration", *Proc. IEEE BioCAS*, pp. 29-32, Nov. 2012

[63] L. Bisoni, et. al, “An HV-CMOS Integrated Circuit for Neural Stimulation in Prosthetic Applications,” *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS*, VOL. 62, NO. 2, FEBRUARY 2015.

[64] P. Favrat, et al., “A High-Efficiency CMOS Voltage Doubler,” *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, VOL. 33, NO. 3, MARCH 1998.

[65] R. Pelliconi, et al., “Power Efficient Charge Pump in Deep Submicron Standard CMOS Technology,” *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, VOL. 38, NO. 6, JUNE 2003.

[66] S. Either, et al., “A ±9 V fully integrated CMOS electrode driver for high-impedance microstimulation,” *IEEE International Midwest Symposium on Circuits and Systems*, 2009. MWSCAS '09.

[67] http://www.ipdia.com/index.php?page=our_products&item_id=104

[68] S. Basir-Kazeruni, et al., "A Blind Adaptive Stimulation Artifact Rejection (ASAR) Engine for Closed-Loop Implantable Neuromodulation Systems," in *Proc. IEEE EMBS Conf. on Neural Eng. (NER'17)*, May 2017, Shanghai, China.

[69] P. M. Furth et al., "On the design of low-power CMOS comparators with programmable hysteresis", *53rd IEEE International Midwest Symposium on Circuits and Systems (MWSCAS) 2010.*, pp. 1077-1080, 1–4 Aug 2010.

[70] H. Chun, “Stimulating Circuits for Visual Neuroprostheses,” Doctoral Thesis, The University of New South Wales, 2011.

[71] D. Smetters, A. Majewska, and R. Yuste, " Detecting Action Potentials in Neuronal Populations with Calcium Imaging ", *Proc. ICSSS*, pp. 215-221, 1999.

[72] N. Spruston, et. Al, “Activity-dependent action potential invasion and calcium influx into hippocampal CA1 dendrites”, *Science* **268**, 297–300 (1995).

[73] E. A. Mukamel, A., Nimmerjahn and M. J. Schnitzer, “Automated analysis of cellular signals from large-scale calcium imaging data”, *Neuron* **63**, 747–760 (2009).

[74] K. K. Ghosh, et al. “Miniaturized integration of a fluorescence microscope”, *Nature Methods 8,* Aug. 2011.

[75] M. Z. Lin, M. J. Schnitzer, “Genetically encoded indicators of neuronal activity”, *Nature Neuroscience 19*, pp. 1142–1153, (2016).

[76] Y. Gong, C. Huang, J. Li, B. Grewe, Y. Zhang, M. Eismann, M. Schnitzer,” High-speed recording of neural spikes in awake mice and flies with a fluorescent voltage sensor*”*, *Science,* Vol. 350 No. 6266, pp. 1361-1366, 2015.

[77] T. -W. Chen, et al., “Ultrasensitive fluorescent proteins for imaging neuronal activity”, *Nature 499,* July 2013.

[78] miniscope.org

[79] D.S. Greenberg, J. N. Kerr, “Automated correction of fast motion artifacts for two-photon imaging of awake animals”, Jan. 2009.

[80] A. Dubbs**,** J. Guevara**,** R. Yuste, **“**moco: Fast Motion Correction for Calcium Imaging”, *Frontiers in Neuroinformatics*, Feb. 2016.

[81] E. A. Pnevmatikakis, A. Giovannucci, “NoRMCorre: An online for piecewise rigid motion correction of calcium image data”, biorxiv, https://doi.org/10.1101/108514.

[82] Y. Bin and D. Hui-Chuan, “Image stabilization by combining gray-scale projection and block matching algorithm,” IEEE International Symposium on Medicine Education, Vol. 1, Aug. 2009, pp. 1262 –1266.

[83] T. Sledevic and A. Serackis, “Surf algorithm implementation on FPGA,” in *Electronics Conference (BEC),* 2012 13th Biennial Baltic, 2012, pp. 291–294.

[84] L. Araneda, M. Figueroa, "Real-time video stabilization on an FPGA", *17th Euromicro Conference on Digital System Design*, 2014.

[85] S. Cain, el al., “Projection-based image registration in the presence of fixed pattern noise”, *IEEE Image Processing Transaction,* Vol. 10, No. 12, PP. 1860-1872, Dec. 2001.

[86] L. Xu, X. Lin, “Digital Image Stabilization Based on Circular Block Matching”, *IEEE Transaction on Consumer Electronics,* Vol. 52, No 2, May 2006.

[87] E. A. Pnevmatikakis, et al., “Simultaneous Denoising, Deconvolution and Demixing of Calcium Imaging Data”, *Neuron,* Vol. 89, Issue 2, Jan. 2016.

[88] B. F. Grewe, et al., “High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision”, *Nat. Methods* 7, 399–405, 2010.

[89] L. Theis, et al., “Supervised learning sets benchmark for robust spike detection from calcium imaging signals”, arXiv:1503.00135, 2015.

[90] J. Friedrich, et al., “Fast online deconvolution of calcium imaging data”, *PLOS Computational Biology,* March 2017.

[91] M. Donoser and H. Bischof, “Efficient maximally stable extremal region (mser) tracking,” in In Proc. of *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR),* New York, USA, June 2006, pp. 553–560.

[92] E. Salahat, et al., "Novel Fast and Scalable Parallel Union-Find Implementation for Real-Time Digital Image Segmentation," in *Annual Conference of the IEEE Industrial Electronics Society,* Yokohama, Japan, Nov. 2015.

[93] R. Sedgewick, Algorithms, 2nd ed. Addison-Wesley, 1988.

[94] V. Petrovic, et al., “A method for real-time memory efficient implementation of blob detection in large images”, *Serbian Journal of Electrical Engineering,* Vol. 17, 2017.

[95] F. Kristensen and W. J. MacLean, "Real-Time Extraction of Maximally Stable Extremal Regions on an FPGA," in *IEEE International Symposium on Circuits and System*, New Orleans, LA, May 2007.

[96] E. Salahat, et al., "A Maximally Stable Extremal Regions System-on-Chip for Real-Time Visual Surveillance," in *Annual Conference of the IEEE Industrial Electronics Society*, Yokohama, Japan, Nov. 2015.

[97] www.vlfeat.org

[98] S. Jewell, Daniella Witten, "Exact Spike Train Inference Via l_0 Optimization", https://arxiv.org/abs/1703.08644, March 2017.

[99] E. A. Pnevmatikakis, et al., "A structured matrix factorization framework for large scale calcium imaging data analysis", https://arxiv.org/abs/1409.2903, Sep. 2014.

[100] J. P. Rickgauer, et al., "Simultaneous cellular-resolution optical perturbation and imaging of place cell firing fields", *Nat Neurosci*, 17(12):1816–1824, 2014.

[101] A. M. Packer, et al. "Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo", *Nat Methods*, 12(2):140–146, 2015.

[102] F. Ren, D. Marković, "A configurable 12–237 kS/s 12.8 mW sparse-approximation engine for mobile data aggregation of compressively sampled physiological signals", *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 68-78, Jan. 2016.

[103] H. Huang, H. Yu" Least-squares-solver Based Machine Learning Accelerator for Real-time Data Analytics in Smart Buildings ", *Emerging Technology and Architecture for Big-data Analytics*, pp. 51-76, Springer, April 2017.

[104] D. L. Donoho, Y. Tsaig, "Fast Solution of l1-norm Minimization Problems When the Solution May be Sparse", 2006.

[105] E. Cands, "Compressive Sampling", *Proc. Int'l Congress of Mathematicians*, 2006.

[106] R. Tibshirani, "Regression Shrinkage and Selection via the LASSO", *J. Royal Statistical Soc. B*, vol. 58, no. 1, pp. 267-288, 1996.

[107] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[108] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," IEEE Signal Process. Mag., vol. 25, no. 2, pp. 21–30, Mar. 2008.

[109] B. Efron, et al. "Least Angle Regression." *The Annals of Statistics*, **32**(2), 2004.

[110] B. Natarajan, "Sparse Appr. Solutions to Linear Systems", *SIAM J. Comput.*, May 1995.

[111] C. Shen, et al., "Sparse representation classification and positive L1 minimization", Aug. 2014.

[112] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach, "A new approach to variable selection in least squares problems", *IMA J. Numerical Analysis*, 20:389–403, 2000.

[113] M. S. Asif, "Dynamic compressive sensing: Sparse recovery algorithms for streaming signals and video," Doctoral Thesis, Georgia Institute of Technology, 2013.

[114] F. Ren, "A Scalable VLSI Architecture for Real-Time and Energy-Efficient Sparse Approximation in Compressive Sensing Systems," Doctoral Thesis, University of California Los Angeles, 2015.

[115] G. M. James, C. Paulson and P. Rusmevichientong, "The constrained Lasso", Technical report, University of Southern California, 2013.

[116] D. Yang, "Turbo Bayesian Compressed Sensing", Doctoral Thesis, Dept. Elect. Eng., Univ. of Tennessee, Knoxville, Knoxville, TN, 2011.

[117] T. A. Davis, et al. "Row modifications of a sparse Cholesky factorization", *SIAM J. Matrix Anal. Appl.* 26, 3, 621–639. 2005.