Ultra-High-Density In-Vivo Neural Probes

Firat Yazicioglu, Carolina Mora Lopez, Srinjoy Mitra, Bogdan Raducanu, Silke Musa, Fabian Kloosterman

Abstract— **The past decade has witnessed an explosive growth in our ability to observe and measure brain activity. Among different functional brain imaging techniques, the electrical measurement of neural activity using neural probes provides highest temporal resolution. Yet, the electrode density and the observability of currently available neural probe technologies fall short of the density of neurons in the brain by several orders of magnitude. This paper presents opportunities for neural probes to utilize advances in CMOS technology for increasing electrode density and observability of neural activity, while minimizing the tissue damage. The authors present opportunities for neural probes to adapt advanced CMOS technologies and discuss challenges in terms of maintaining the signal integrity and implementing data communication.**

I. Introduction

Great effort is being invested into deciphering the functioning of brain and understanding the specific neural interactions and circuits. Several studies have demonstrated that the understanding of brain functions can only be achieved by monitoring the electrical activity of large numbers of individual neurons in multiple brain areas at the same time. Neural probes comprising multiple biopotential electrodes have proven to be an effective tool for recording activity from large neuronal ensembles. Unlike any other imaging method, implantable neural probes can access virtually any depth of brain and record signals from individual neurons with high temporal resolution. Nevertheless, given the density of neurons in the brain, the spatial resolution and area coverage of neural probes is limited (see **Fig. 1**). This is a widely recognized limitation of present day neural probes [1] and it is known that increasing the density of electrodes, i.e. the number of electrodes per volume, significantly enhances the success of identifying individual neurons and creates possibilities to identify more neurons from larger neuron populations [2][3].

With the emergence of microfabrication technologies in the 1960s, silicon-based probes have started to be fabricated, providing advantages such as higher electrode densities and compatibility with microelectronic processes. The ever-improving routing density of semiconductor technology in combination with advances in materials science and process technologies has enabled silicon multi-electrode recording arrays with increasing number of recording sites [4]-[12][13] even with electrode diameters as small as 1µm-5µm [13]-[15]. Both the density and the observability of the electrodes can be significantly improved by using a more advanced lithography process, which is readily available in CMOS technology. **Fig.1** shows that density of recording electrodes and their observability - the number of electrodes that can be monitored simultaneously - is increased by an order of magnitude in a neural probe utilizing 0.18µm CMOS technologies compared to a neural probe realized in 0.5µm technology.

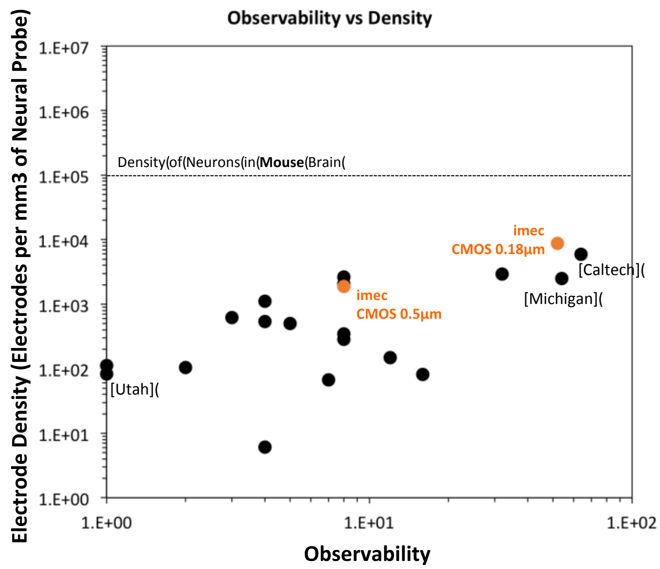

Figure 1. Comparison of different Neural Probe implementations. "Density" indicates the number of electrodes per mm^3 of a neural probe which is inversely related to the invasiveness per electrode of the neural probe. "Observability" indicates the maximum number of electrodes that can be monitored simultenously from a single neural probe shank.

In this paper, we will discuss the opportunities for neural probes to adapt advances in nanometer CMOS technology. We will discuss the advantages of scaling CMOS technologies for neural probes in terms of higher routing density and observability, and analyze signal integrity challenges arising along with high density routing. We will show that SNR requirements puts an upper limit to the density of recording electrodes, regardless of the advances in the routing density. Later, we will discuss emerging analog and digital signal processing techniques that can be used to reduce the data rate prior to data communication. We will show that the capabilities of todays communication technologies falls short compared to the amount of data that can be extracted from high-density neural probes. Emerging analog and digital signal processing techniques offers high power efficiency data reduction techniques, enabling the complete utilization of the capabilities of high-density neural probes.

F. Yazicioglu is with the imec, Kapeldreef 75, 3001 LEUVEN (phone: +32 472 470 249; fax: +32 16 28 8100; e-mail: firat@imec.be).

Carolina Mora Lopez, Srinjoy Mitra, Bogdan Raducanu and Silke Musa, are with imec, Kapeldreef 75, 3001 LEUVEN.

Fabian Klostermann is with the Neuro-Electronics Research Flanders, Campus Imec, Kapeldreef 75, 3001 LEUVEN.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

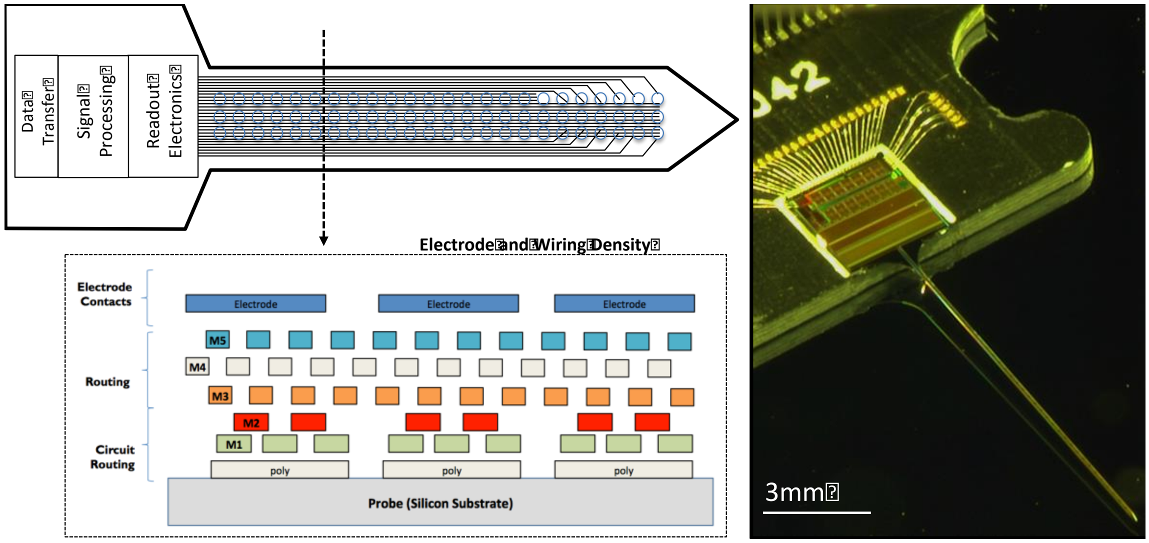

Figure 2. Architecture of a neural probe using CMOS technology. High density electrode arrays is routed to the probe base by using high density metal stack of CMOS technologies. The probe base consists of readout electronics, signal processing, and data communication.

II. High Density Neural Probes

High-density neural probes push the boundaries of technology. Minimizing the electrode area, increasing the routing density, and reducing the probe thickness, the high density neural probes aim for gathering the maximum possible data with minimal tissue damage. **Fig. 2** shows the architecture of such a high-density neural probe realized in 0.18um CMOS process [16]. The shank of the neural probe is covered with recording electrodes, which are connected to the readout electronics at the base through a dense metal routing. By utilizing the metal density of CMOS technology, both the electrode density and the observability of the electrodes can be increased significantly.

Among most well-know high-density neural probes, the Michigan probes have been fabricated in a variety of designs, including single shank, two-dimensional and three-dimensional arrays. The recording sites range from 70 to 4000 μm^2, with up to 54 sites per neural probe shank [6]. The Caltech probe utilizes Silicon-on-Insulator (SOI) wafers and e-beam process to fabricate a silicon neural probe with 64 recording channels [13]. The gold plated recording electrodes are 108μm^2 and e-beam process can produce routing as narrow as 290nm maximizing the observability. On the other hand, imec probes has demonstrated that using CMOS technology (0.18µm), the total number of recording sites can be increased up to 455 with an observability of 52 simultaneous recording electrodes, limited by the routing density of the CMOS 0.18µm CMOS process [16].

A. Signal Integrity and Observability

More advanced CMOS technologies have higher routing density and more number of metal layers. **Fig. 3** shows that the observability, as defined in **Fig. 1,** of a neural probe increases by an order of magnitude in 40nm with respect to 0.18um CMOS technology. Looking at imec's neural probe design in [16] with an observability of 52 out of 455 electrodes, in 40nm CMOS technology, the observability could be increased up to 390, thanks to the availability of higher number of metal layers and smaller pitch routing metals.

Another consequence of increasing routing density is the increasing shunt capacitance between the routing lines. Considerable crosstalk between neural recording channels (> 1% or 40 dB) leads to spuriously detected neural spikes and increased noise in recordings [4],[6]. Our calculations in [16] show that in 0.18µm CMOS technology, two adjacent routing lines along the probe length inflict 2.5% crosstalk. If higher routing density of 40nm technology is used, the crosstalk between two adjacent channels increases by a factor of 4, reaching up to 10% (see **Fig. 3**). Hence, it can be concluded that only using high density routing does not yield high-density neural probes.

B. Recording Electrode Density and Flicker Noise

To prevent crosstalk, an active electrode concept can be utilized [16][17]. In this concept, an active circuitry, a buffer or an amplifier, is located at the electrode site driving the routing between the electrode and readout electronics. This active circuit can reduce the crosstalk by 14dB in 0.18um CMOS process. However, the presence of active electronics

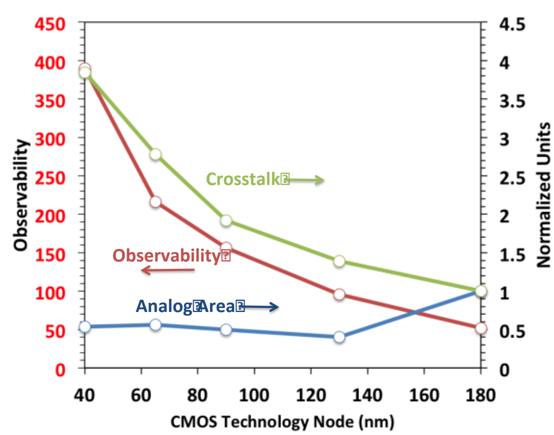

Figure 3. Affect of technology on neural probe observability (red), signal integrity (green), and electrode density (blue).

under the recording electrodes limits the density of recording sites.

Unlike analog design for RF applications, the low-frequency content (0.5Hz – 10kHz) and weak amplitudes (10uV – 10mV) of neural signals make flicker noise of CMOS transistors the dominant noise source. As the parameters that define the flicker noise do not significantly improve with CMOS technology scaling (gate-oxide capacitance, Cox, and flicker noise constant K), the area of =analog electronics cannot scale in advanced CMOS technologies as seen in the equation below;

$$\boldsymbol{S}(f) = \frac{K}{\boldsymbol{C_{OX}}} \times \frac{\mathbf{1}}{\boldsymbol{AREA}} \times \frac{1}{f}$$

For a given flicker noise specification, the relative area of an analog integrated circuit can only be reduced by 50% when moving from 0.18um to 40nm technology (**see Fig. 3 – blue curve**) providing only minor benefits on the density of recording electrodes.

As a conclusion, while on one side scaling in CMOS technology helps in increasing observability, on the other side, it has only minor benefits for increasing the electrode density. This is mainly due to the fact that increasing routing density leads to crosstalk and requires the use of analog electronics, or so called active electrodes. These analog electronics suffer from flicker noise and occupy considerable area even in scaled CMOS technologies.

III. Analog and Digital Signal Processing Techniques to Reduce Data Rate

In neural recording experiments, the digitized neural data needs to be transferred from the neural probe to storage or processing system. The increased data rates in high-density neural signal recording require the availability of large bandwidth communication. One of the major challenges of neural probe technologies is to implement such large bandwidth data communication with minimal possible size and weight impact such that the freely moving animal experiments are not affected.

A. High Density Neural Recording and Data Communication

Higher integration has opened the path to more complex and miniaturized neural probes and systems. In these neural probes, it is typical that the neural signal is captured by using instrumentation amplifier, filters, and analog-to-digital converter inside the probe itself or in nearby circuits[16]. This digitized data can be communicated by using either wireless or wireline communication technologies.

While wireless systems allow for maximum movement freedom during the experiment, their main limitations come from high power consumption per transmitted bit and the low amount of available bandwidth. This forces current wireless systems to compromise in regard to the number of channels, the amount of captured data, or the bandwidth of the transferred data when applied to miniature neural recording systems [20][21]. As wireless systems require a local power source another tradeoff that needs to be made is in the size of the system and the length of continuous operating times. A typical commercially available wireless system with 16 recording channels transmitting has an operating time of approximately 2 hours [18]. Should such a system be scaled linearly to 1000 channels, the operating time from the same battery will be reduced to 2 minutes only making in-vivo recordings impractical. Future wireless technologies and future battery technologies may alleviate the situation, but it is clear that a great improvement is required to facilitate the wireless transmission of data from high density neural probes.

The alternative to wireless systems is wired systems, which connect the animal, and recording system through a tether, supplying power and facilitating the communication between the implant and recording equipment (**Fig. 5)**. While such a system constrains the freedom of movements during the experiment, it has the advantages of being able to operate for an unlimited time and sustain a high data rate required to transmit the complete captured data. While at low data rates CMOS signal levels may be used, its power consumption will rise as throughput and cable length increases due to the requirement to charge the line capacitance to the full power supply. The typical alternative for data rates starting at 100's of Mbps is to use low voltage differential systems (LVDS) or derivatives. Current low voltage differential systems are capable of realizing a point to point connection at multi Gbps data rate over a single twisted pair [19] of wires consuming mW of power only. This opens up the possibility of connecting a highly integrated neural probe containing signal chain, conversion, processing, transmission and power supply conditioning through as little as 4 wires which can insure a rather flexible tether.

B. Analog and Digital Signal Processing for Data Reduction

Data transmission from high density neural probes accounts for the significant portion of the power budget. It is also quite often the case that the Analog Front End (AFE) for a neural probe has a small footprint while a separate PCB (head-stage) is built to drive a long cable at hundreds of Mbps (or even higher) [16]. This power-hungry head-stage has to be larger compared to the probe itself and has to be placed at a distance such that there is minimal reverse thermal flow to the brain. The additional system constrains

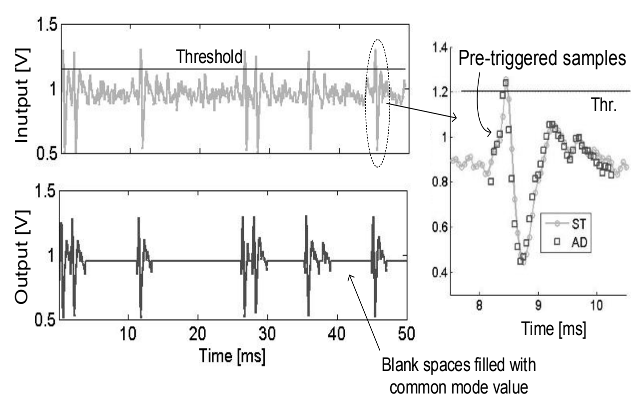

Figure 4. Data compression using analog buffers. The input and and output from the aFIFO (including buffer, ADC etc.) is shown above. A zoomed image shows the samples before threshold crossing (pre-triggers) with no significant distortion.

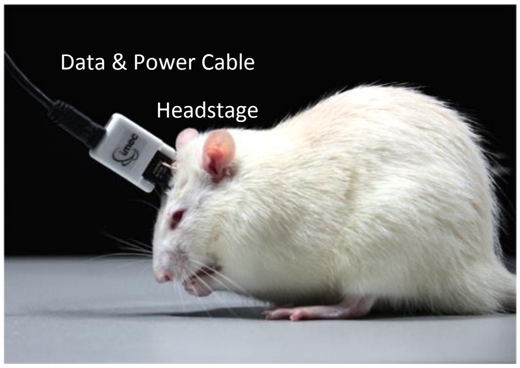

Figure 5. A neural recording headstage of the neural probe shown in **Fig. 2,** measuring in-vivo neural potentials from a rat brain.

become even more evident for probes with wireless communication modules. Here the battery volume and weight dominates the form factor of the entire system and restricts long time usage on freely moving animals. The handling of the data deluge on the receiver side is also turning out to be problematic once the number of electrodes reaches few hundreds.

Given these limitations, it is obvious that data compression techniques are necessary, if these neural probes are intended to be used on small animals for an extended period of time. It is widely considered that the primary information in neural signals is confined within the infrequent action potentials (AP or spike), hence the recorded data can be considerably reduced by using a system based on spike activity. The simplest and most compact data reduction technique is to transmit only the time instance when a spike is detected [22]. A spike is normally detected by comparing the raw value of the action potential or some transform (e.g. NEO [23]) with constant or an adaptive threshold. However, this compression method loses all information on the shape of the action potential and is unsuitable when spike-classification from multiple neurons is necessary. Hence other methods are devised to retain some of the spike shape with a varying degree of accuracy. This includes information on spike width, spike amplitude, zero crossing instances and piece-wise approximation [24]-[26]. Nevertheless, these lossy compression methods severely limit the usability of the system when a single electrode captures spikes from multiple neurons. Transmitting a window of uncompressed data around a spike is an alternative approach that leads to better classification. This can be easily done by continuously digitizing and buffering the data, and transmitting it only when a spike is detected. A more energy saving approach is to use an analog buffers (aFIFO) to store a small fraction of the data and only digitize and transmit it after a spike is detected [27].

All these data compression mechanisms require very little computational overhead and can be done with fairly simple analog and/or digital circuits. More complex techniques based on delta compression [28], Discrete wavelet transform [29] and Compressed Sensing [30] have been proposed over the last few years. In these methods, apart from the on-chip compression circuitry, the data recovery algorithm (running at the base station or PC) also has to be designed with special consideration. Rather than just the spikes, these methods typically concentrate on transmitting the least amount of data from which the entire waveform can be reconstructed. Such waveforms are often preferred by neuroscientists so as to have the possibility of more sophisticated off-line data processing. A recently proposed technique based on Compressed Sensing [31] can reconstruct the entire waveform but also incorporates the benefits of spike activity dependent processing. In this case, when a spike is detected, the compressed data is reconstructed using pre-learned 'dictionary' while the inter spike intervals are reconstructed using wavelets. This method achieves ~95% spike classification accuracy even after 10x data reduction and yet provide a means to post-process the entire waveform.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

Among different brain imaging technique, the electrical measurements of neural activity using neural probes provide the highest temporal resolution. Nevertheless, the capabilities of current neural probes technologies have limited spatial resolution when compared with the density of neurons in the brain. CMOS technologies are emerging as a key enabler for increasing the electrode density of neural probes yet new and miniature instrumentation architectures are necessary to ensure signal integrity in high-density neural probes. Once the high-density neural probes are a reality, the data transfer, wired or wireless, will be the key bottleneck. Thus, data compression techniques in combination with low-power data transmission circuits will be the key enablers towards understanding the functioning of brain.

REFERENCES

[1] G. Buzsaki, Nat Neurosci 7/5, 446 (2004).
[2] B. L. McNaughton et. al., J. Neurosci. Methods 8, 391 (1983)
[3] M. L. Recce et. al., Soc. Neurosci. Abstr. 15, 1250 (1989)
[4] P. K. Campbell et.al., IEEE Tran Biomedical Eng 38, 758 (1991)
[5] R. Olsson et.al., IEEE Biomedical Engineering 52, 1303 (2005)
[6] T. Blanche et.al, J Neurophysiol 93, 2987 (2005).
[7] R. J. Vetter et.al., IEEE Eng. In Med and Bio Conf, 7341 (2005)
[8] G. E. Perlin, K. Wise, J MicroElectroMech Sys 19, 1409 (2010).
[9] P. Ruther et.al, IEEE Biomedical Engineering, 2284 (2011)
[10] T. Torfs et al., IEEE Trans on Biomed Crts and Sys 5, 403 (2011).
[11] G. U. Hofmann et.al., IEEE Trans on Biomed Eng 53, 1672 (2006)
[12] P. Norlin et.al. J Micromechanics and Microeng 12, 414 (2002).
[13] J. Du et.al., Plus One 6/10, e26204 (2011)
[14] D. B. Suyatin et.al., PLoS One 8/2, e56673 (2013)
[15] J. T. Robinson et.al, Frontiers in Neural Crts 7, 1 (2013)
[16] C. M. Lopez et.al., IEEE J Solid State Crts 49/1, 248 (2014)
[17] J. Xu et. al., IEEE J Solid State Crts 48/7, 1575 (2013)
[18] W-16 System, http://www.multichannelsystems.com/products, [online] Accessed June. 06, 2014
[19] M. Chen et al, IEEE J Solid State Crts 40, 2 (2005)
[20] A Borna et al, IEEE J Solid State Crts 49,2 (2014)
[21] Farshchi, S et al, IEEE Trans 18, 2 (2010)
[22] R. Harrison. IEEE Proceedings 96, 7 (2008)
[23] M. S. Chae et. al., IEEE Trans Neural Sys 17, 4 (2009)
[24] B. Gosselin, et. al., IEEE Trans on Biomed Crts and Sys, 3, 3 (2009)
[25] Awais M. Kamboh et. al, IEEE ISCAS (2010)
[26] A.Rodríguez-Pérez et. al., IEEE Trans on Biomed Crts and Sys 6, 2 (2012)
[27] S. Mitra et. al., IEEE ISSCC, 292 (2013)
[28] J. N. Y . Aziz et. al., IEEE J Solid State Crts 44,3 (2009)
[29] K. G. Oweiss et. al., IEEE Trans on Crts and Sys I 54,6 (2007)
[30] F. Chen et al., IEEE J Solid State Crts 47,3 (2012)
[31] J. Zhang et. al., IEEE Trans on Biomed Crts and Sys, (2014)