Data Compression versus Signal Fidelity Trade-off in Wired-OR ADC Arrays for Neural Recording

Yan, Pumiao ; Shah, Nishal P.; Muratore, Dante; Tandon, Pulkit ; Chichilnisky, E.J. ; Murmann, Boris

DOI
10.1109/BioCAS54905.2022.9948677

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the 2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)

2022 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 978-1-6654-6917-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/BioCAS54905.2022.9948677

Data Compression versus Signal Fidelity Trade-off in Wired-OR ADC Arrays for Neural Recording

Pumiao Yan*, *Student Member, IEEE* Nishal P. Shah*†, Dante G. Muratore§, *Senior Member, IEEE*, Pulkit Tandon*, E.J. Chichilnisky†‡, and Boris Murmann*, *Fellow, IEEE*
*Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA
†Department of Neurosurgery, Stanford University, Stanford, CA 94305 USA
‡Department of Ophthalmology, Stanford University, Stanford, CA 94305 USA
§Department of Microelectronics, Delft University of Technology, Delft, The Netherlands

***Abstract*—This paper investigates the efficacy of a wired-OR compressive readout architecture for neural recording, which enables simultaneous data compression of action potential signals for high channel count electrode arrays. We consider a range of wiring configurations to assess the trade-offs between compression ratio and various task-specific signal fidelity metrics. We consider the fidelity in threshold crossing detection, spike assignment, and waveform estimation, and find that for an event SNR of 7-10 the readout captures at least 80% of the spike waveforms at ∼150x data compression.**

***Index Terms*—A/D conversion, brain-machine interfaces, compression algorithm, neural interfaces.**

I. Introduction

The current trend in brain-machine interfaces is to record from an increasing number of neurons. State-of-the-art sytem examples include Neuropixel [1], Argo [2] and Neuralink [3], [4]. Despite these advances, the number of channels that can be simultaneously recorded within the power and area constraints of an implantable device is still limited to a few thousand. A compromise for larger channel counts is to use an on-chip switch matrix or multiplexer [2], [5]. However, this precludes simultaneous recording and hence limits the capabilities of the end application. Performing lossy data compression as close as possible to the the physical interface is a promising approach to address this issue.

For lossy compression, it is crucial to identify the signal's salient information versus unnecessary data samples. Traditionally, after bandpass filtering and channel noise estimation, spikes are detected using a negative threshold (see Fig. 1(a)). Depending on the application, the spikes can be sorted to separate individual neurons and study their interactions. In sensory and motor applications, it has been shown that the spike waveforms can be reduced to threshold crossings (binary spike trains, see Fig. 1(b)) [3], [6]–[8]. In contrast, the spike waveform shape (see Fig. 1(c)) is essential for applications that require cell type identification [9]–[12]. Regardless of the application, most of the information about extracellular activities are captured in the spike waveforms, which are hence sufficient and desirable for complete information extraction [13].

Given the significance of the spike events, one idea for data compression is to detect the spike times and only record samples in their vicinity (thus eliminating baseline samples between spikes). From a hardware perspective, a key issue with this approach lies in finding the proper threshold and managing the data movement with limited resources in a dense array. These issues are seen in [14], which uses analog memory cells and additional computation to find the thresholds. Our previous work [15] sidesteps this issue using a wired-OR analog-to-digital converter (ADC) array. In this architecture, samples are discarded based on a wired-OR competition between the pixels and no thresholding is needed. While we have shown that this technique works well for retinal cell identification (∼40x compression while missing less than 5% of cells), we wish to assess its suitability for a broader range of applications. Thus, the purpose of this paper is to analyze the wired-OR architecture with a range of commonly used neural signal processing methods to understand the trade-offs between performance and compression ratio. Additionally, we explore generalizations of the wired-OR architecture toward application-specific reconfigurability.

Section II reviews the wired-OR architecture and discusses different wiring schemes to expand its configuration space. Next, Section III summarizes our simulation results. The experiments are based on large-scale, high-density *ex vivo* primate retina recordings, which exhibit a range of event signal-to-noise ratio (SNR) values. This enables a translation of the results to any neural system as long as the event SNR is known. We assess a range of use cases (threshold crossing detection, spike assignment, waveform estimation) and find that the wired-OR topology captures at least 80% of the spike waveforms at ∼150x data compression.

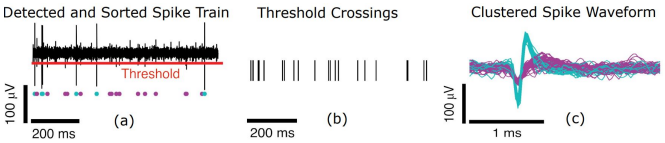

Fig. 1 Neural signal waveforms.

This project was supported in part by Stanford's Wu Tsai Neurosciences Institute. P. Y. was supported by a Stanford Bio-X SIGF fellowship.

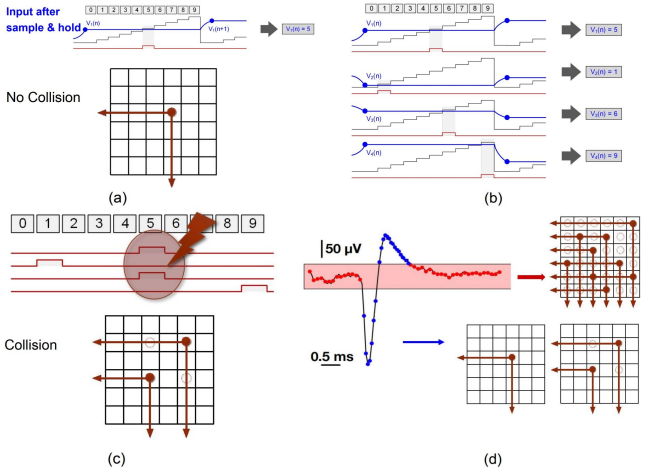

Fig. 2 Wired-OR readout concept. (a) Conversion of voltage to pulse position and collision-free readout of one pixel. (b) Multiple pixels with collision-free pulse timing. (c) Collision between two pixels. (d) Likely outcomes across different waveform levels.

II. Readout Architecture

A. Wired-OR readout concept

In the wired-OR readout architecture [15], each pixel conditions and samples the input as commonly done in neural interfaces. The sampled voltage is then converted into a pulse position, which is achieved by comparing it to a globally-distributed ramp step signal (see Fig. 2(a)). In the most basic implementation, the pulses from pixels in the same row or column are combined onto single wires using wired-OR circuitry. In essence, signal compression occurs by having the pixels compete for these limited wire resources. If only a single pixel produces a pulse at a given time step (i.e., it is the only channel with a quantized voltage corresponding to the time step, see Fig. 2(b)), then the pixel location and its A/D conversion result (ramp counter state) can be uniquely recovered. On the other hand, if multiple pulses from different pixels occur at the same time step (i.e., the quantized voltages on two or more channels are equal) multiple rows and/or columns are activated (collision case in Fig. 2(c)) and the conversion results cannot be recovered (samples are discarded). As discussed in [15], this compression approach is effective for neural signals due to their long-tailed probability distribution. Voltage samples associated with spikes tend to be unique and are typically retained while baseline samples falling within a certain voltage range tend to be discarded (see Fig. 2(d)). This architecture is scalable to a large number of channels ($\gg 1,000$), enabling the next generation of neural interfaces [15].

B. Generalization of wired-OR configurations

The previous subsection reviewed the basic wire-OR readout concept with a single wire in each row and column. However, other configurations are possible, including multiple interleaved wires per row/column as well as diagonals. Diagonal wiring has been explored previously for particle tracking in high-energy physics [16] and we wish to assess its merits for our readout scheme.

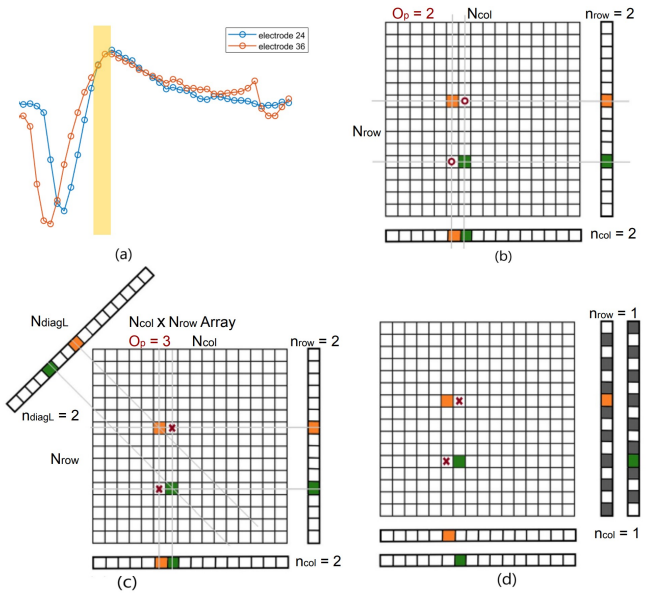

Fig. 3 Collision case for different wiring scheme (a) Action potential of two electrodes. (b) Previously proposed wired-OR. (c) Diagonal wiring. (d) Interleaving wiring.

To investigate further, consider a collision case in which two pixels record the same voltage levels simultaneously (see Fig. 3(a)). Through the wired-OR logic, the pixels are projected onto the row and column index of the data matrix (see Fig. 3(b)) and different wiring configurations can be abstracted as different projections. For the shown case, the address of channels is not uniquely decodable, and results in a collision and losing the corresponding spike samples. One way to decode this case is by adding a diagonal projection through extra wiring (see Fig. 3(c)). In a $N_{\text{col}} \times N_{\text{row}}$ array (32×16 in this work), collisions that trigger the correlated channels are denoted in n_{p}. When $Max(n_{\text{p}}) \leq 2$, the triggered channels are uniquely decodable with one set of added diagonal wiring. A right diagonal projection can be added as another projection (the number of projections/wiring is denoted by O_{p}). For $O_{\text{p}} = 4$, when $Max(n_{\text{p}}) \leq 4$, the triggered channels are uniquely decodable. While there are more uniquely decodable cases when $Max(n_{\text{p}}) > 4$, this would require more complex decoding logic and it is not considered here. This solution effectively disentangles two or more channels recording the same voltage levels. When multiple samples are recorded simultaneously, a prefix code is needed to also record the number of samples. Huffman coding, which is commonly used for lossless data compression is an option for the prefix code encoding. Effectively, the data rate is:

$$R_{\text{p}} = [\text{Huffman Code} + \sum_{O_{\text{p}}} log_2(N_{\text{p}}) \times \alpha_{\text{d,p}}] \times f_{\text{s}} \qquad (1)$$

where N_{p} denotes the dimension of the corresponding projection, and "Huffman Code" being the average bits of Huffman prefix code (bounded by $log_2(O_{\text{p}})$). The resulting data rate depends on the average rate of decodable channels per sample ($\alpha_{\text{d,p}}$) and sampling frequency (f_{s}). For comparison, the data

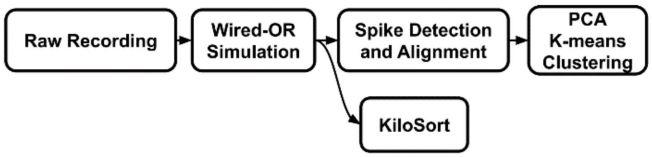

Fig. 4 Neural signal processing pipeline for our simulation study.

rate for W interleaved wires (see Fig. 3(d), where $W = 2$) is [16]:

$$R_{\mathrm{W}} = [log_2(N_{\mathrm{row}}/W) + log_2(N_{\mathrm{col}})]\alpha_{\mathrm{d,W}} f_{\mathrm{s}} \quad (2)$$

For B-bit ADC resolution, the corresponding compression ratio is:

$$CR = \frac{N_{\mathrm{col}} \times N_{\mathrm{row}} \times B \times f_{\mathrm{s}}}{R} \quad (3)$$

Knowing the estimated average rate of decodable channels per sample in a neural dataset, one can calculate the output data rate of wired-OR with (1) and its corresponding compression ratio with (3), guiding a configuration choice with higher compression ratio. For example comparing the cases of $O_{\mathrm{p}} = 4$ and $W = 4$, $O_{\mathrm{p}} \times \alpha_{\mathrm{d,p}} \approx \alpha_{\mathrm{d,W}}$. Then, $\frac{CR_{\mathrm{O=4}}}{CR_{\mathrm{W=4}}} \approx \frac{[log_2(N_{\mathrm{row}}/4)+log_2(N_{\mathrm{col}})]\times\alpha_{\mathrm{d,W}}}{\mathrm{HuffmanCode}+log_2(N_{\mathrm{p}})\times O_{\mathrm{p}}\times\alpha_{\mathrm{d,p}}} \approx 1.3$, given the array used in the primate retina dataset. This matches our simulation results in Section III, showing the compression ratio of diagonal wiring achieves 1.3x higher compression ratio than the 4-interleaving wires configuration with similar performance. The trade-offs between compression ratio and signal fidelity is further evaluated through simulations in the next section.

III. Simulation Results

To evaluate the performance of the wired-OR readout architecture, we use 512-channel data recorded in *ex vivo* experiments with primate retina. The readout scheme discussed in Section II is emulated by re-processing the recorded raw data in software. We then take the wired-OR compressed data and apply neural signal processing steps such as spike detection, waveform estimation, and spike classification. To show the efficacy of wired-OR compression for a retinal prosthesis application, we also applied automated spike sorting and cell-type classification using KiloSort [17]. Our analysis pipeline is shown in Fig.4.

A. Spike detection and alignment

The most commonly used spike detection method for implantable neural signal processors is thresholding. Conversely, the wired-OR scheme discards baseline samples that cannot

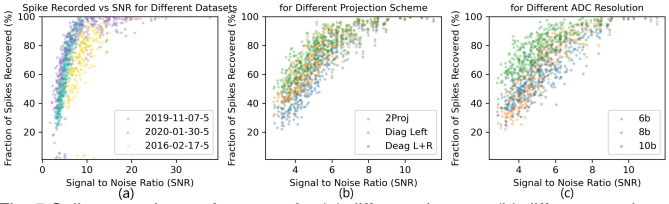

Fig. 5 Spike capturing performance for (a) different datasets (b) different numbers of diagonal wires (c) different ADC resolutions.

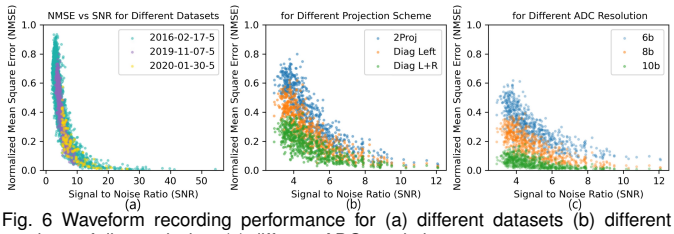

Fig. 6 Waveform recording performance for (a) different datasets (b) different numbers of diagonal wires (c) different ADC resolutions.

be uniquely decoded by construction. The recorded wired-OR samples can be analyzed for spike alignment and further single-channel analysis to separate the spikes. We adopt the simplest spike detection and alignment based on the wired-OR readout strategy. For any given channel, when uniquely decodable samples are recorded, we search for the minimum-value sample in the window of the next 30 samples (this number is empirically chosen) to align the spikes. We then analyze the percentage of spikes captured by comparing it to raw, full-bandwidth dataset with the knowledge of spike times detected using KiloSort. The results are shown in Fig. 5, which shows the percentage of spikes captured for each identified neuron in KiloSort. The event SNR is approximated by [1]:

$$SNR = \frac{V_{spike\ peak\ amplitude}}{V_{\sigma,\,channel}}$$

Here, the spike peak amplitude is found from the electrode with the largest negative peak, and the noise is the median absolute deviation when no action potential is seen on the channel. The percentage of spikes captured shows a consistent correlation to the event SNR across recorded datasets collected in the span of 4 years (see Fig. 5(a)), which makes the performance metrics analyzed here translatable given the SNR. The performance of different wiring schemes is shown in Fig. 5(b), where "Diag Left" refers to wired-OR plus one extra set of diagonal wires as shown in Fig. 3(c), and "Diag L+R" extends this to two directions of diagonal wiring. Since adding diagonal wiring improves the capability of wired-OR architecture to decode cases when multiple channels are triggered, more spikes can be captured.

The percentage of spikes captured for an SNR range of 3-12 given different ramp signal resolutions is shown in Fig. 5(c). Increasing ramp signal resolution is shown to improve performance because the finer the voltage levels, the less chance of multiple channels falling into the same quantization voltage levels. For SNR in the range of 7-10, at least 80% of the spikes are captured for all configurations of wired-OR. For a typical neural recording system design such as

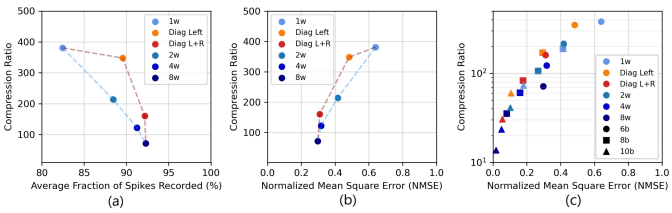

Fig. 7 Pareto frontier of wired-OR performance for (a) average fraction of spikes captured vs. compression ratio (b) average NMSE vs. compression ratio with different wiring schemes (c) average NMSE vs. compression ratio for all configurations

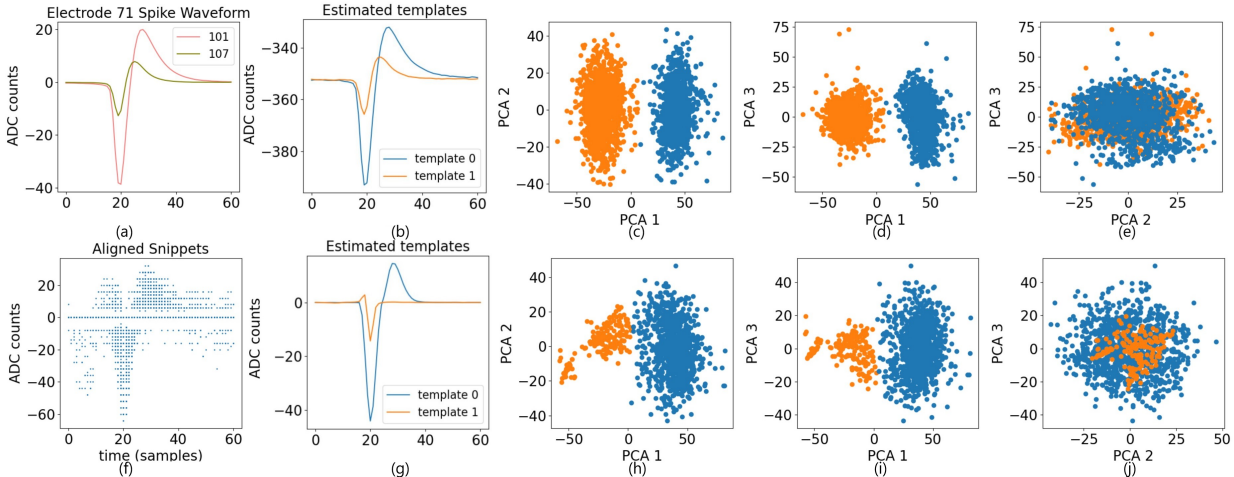

Fig. 8 Single channel analysis. (a) Spike waveform extracted through KiloSort. (b) K-means estimated templates from raw recorded spikes. (c-e) PCA analysis of clustered raw spikes with number of components is set to 3. (f) Detected and aligned spikes from compressed data. (g) K-means estimated templates from detected spikes of compressed data. (h-j) PCA analysis of clustered compressed spikes.

the Neuropixels probe, where the SNR is around 8 [1], over 90% of spikes are predicted to be captured by adding diagonal wiring.

B. Spike waveform estimation

To analyze the performance of wired-OR in recording the waveform of action potentials, we studied the normalized mean square error in the spike waveforms for each cell-electrode pair compared to that from the uncompressed dataset. As before, the wired-OR performance shows a strong correlation to the event SNR, as shown in Fig. 6(a). Additional wiring and higher ADC resolution reduce the NMSE (see Fig. 6(b-c)). And as one would expect, the improvement in performance comes at the cost of compression ratio. As shown in Fig. 7(a-b), for both diagonal and interleaved wiring, increasing the number of wires lowers the average NMSE across the entire dataset, as well as captures more spikes, but also decreases the achievable compression ratio. Previously proposed wired-OR and interleaving wiring scheme results are demonstrated in blue, and diagonal wring results are shown in warm colors. As expected, diagonal wiring results surpass the previous Pareto frontier. Comparing the configuration of diagonal wiring in both directions (4Proj) to 4 interleaved wires, fewer wires and higher compression are possible while further lowering the average NMSE. The trade-off between performance and compression ratio for all studied configurations is summarized in Fig. 7(c).

C. Spike classification

After the spike events are extracted and aligned, a common procedure in neural signal processing is to reduce the high dimensionality of the recorded neural data. We applied K-means clustering and principal component analysis (PCA) to demonstrate the effect of separability among recorded neurons after compression. An example of electrode 71 is demonstrated with the knowledge of spike waveform extracted from uncompressed dataset shown in Fig. 8(a). We compare compressed data shown in Fig. 8(f-j) to raw data processed in the same way shown in Fig.8(b-e). With simple spike detection and clustering algorithm, relatively large amplitude spikes that are recorded at the soma of neurons can be clearly separated from waveforms recorded from the axon or artifact signals. Over 99% of spikes of the cell with ID# 101 (demonstrated as an example) are captured and correctly matched to its cluster. This shows that although the compression is lossy, wired-OR captures spikes without thresholding and still retains sufficient information to sort different recorded units. Different from previously proposed architectures [2], [14], no computation for the threshold is needed.

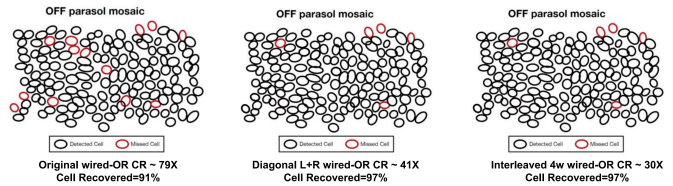

Fig. 9 Receptive field mosaic for OFF parasol cells from several wired-OR configurations.

D. Spike sorting and cell-type classification.

To assess the performance of diagonal wiring in cell-receptive-field mapping application in retinal prostheses, we also passed the compressed data through state-of-the-art spike sorting and cell-type classification using KiloSort. The recovered receptive field mosaic of the collection of cells of one type (OFF parasol) is illustrated in Fig. 9. For the same percentage of cells recovered, diagonal wiring achieves higher compression than interleaved wiring.

IV. Conclusion

We conducted a simulation study of the data compressive wired-OR readout architecture using a range of neural signal processing methods. We considered various different wire configurations and found that diagonal wiring is more effective than interleaved wiring. For a typical event SNR of 7-10, the wired-OR readout captures at least 80% of the spikes at ∼150x compression, while maintaining sufficient waveform fidelity for spike sorting.

REFERENCES

[1] J. J. Jun, N. A. Steinmetz, J. H. Siegle, D. J. Denman, M. Bauza, B. Barbarits, A. K. Lee, C. A. Anastassiou, A. Andrei, Ç. Aydın, M. Barbic, T. J. Blanche, V. Bonin, J. Couto, B. Dutta, S. L. Gratiy, D. A. Gutnisky, M. Häusser, B. Karsh, P. Ledochowitsch, C. M. Lopez, C. Mitelut, S. Musa, M. Okun, M. Pachitariu, J. Putzeys, P. D. Rich, C. Rossant, W.-L. Sun, K. Svoboda, M. Carandini, K. D. Harris, C. Koch, J. O'Keefe, and T. D. Harris, "Fully integrated silicon probes for high-density recording of neural activity," *Nature*, vol. 551, no. 7679, pp. 232–236, Nov. 2017.

[2] K. Sahasrabuddhe, A. A. Khan, A. P. Singh, T. M. Stern, Y. Ng, A. Tadić, P. Orel, C. LaReau, D. Pouzzner, K. Nishimura, K. M. Boergens, S. Shivakumar, M. S. Hopper, B. Kerr, M.-E. S. Hanna, R. J. Edgington, I. McNamara, D. Fell, P. Gao, A. Babaie-Fishani, S. Veijalainen, A. V. Klekachev, A. M. Stuckey, B. Luyssaert, T. D. Y. Kozai, C. Xie, V. Gilja, B. Dierickx, Y. Kong, M. Straka, H. S. Sohal, and M. R. Angle, "The argo: a high channel count recording system for neural recording in vivo," *J. Neural Eng.*, vol. 18, no. 1, p. 015002, Feb. 2021.

[3] E. Musk and Neuralink, "An integrated Brain-Machine interface platform with thousands of channels," *J. Med. Internet Res.*, vol. 21, no. 10, p. e16194, Oct. 2019.

[4] D.-Y. Yoon, S. Pinto, S. Chung, P. Merolla, T.-W. Koh, and D. Seo, "A 1024-channel simultaneous recording neural soc with stimulation and real-time spike detection," in *2021 Symposium on VLSI Circuits*, 2021, pp. 1–2.

[5] X. Yuan, A. Hierlemann, and U. Frey, "Extracellular recording of entire neural networks using a Dual-Mode microelectrode array with 19 584 electrodes and high SNR," *IEEE J. Solid-State Circuits*, vol. 56, no. 8, pp. 2466–2475, Aug. 2021.

[6] E. M. Trautmann, S. D. Stavisky, S. Lahiri, K. C. Ames, M. T. Kaufman, D. J. O'Shea, S. Vyas, X. Sun, S. I. Ryu, S. Ganguli, and K. V. Shenoy, "Accurate estimation of neural population dynamics without spike sorting," *Neuron*, vol. 103, no. 2, pp. 292–308.e4, Jul. 2019.

[7] P. Gao and S. Ganguli, "On simplicity and complexity in the brave new world of large-scale neuroscience," *Curr. Opin. Neurobiol.*, vol. 32, pp. 148–155, Jun. 2015.

[8] N. Even-Chen, D. G. Muratore, S. D. Stavisky, L. R. Hochberg, J. M. Henderson, B. Murmann, and K. V. Shenoy, "Power-saving design opportunities for wireless intracortical brain–computer interfaces," *Nature Biomedical Engineering*, vol. 4, no. 10, pp. 984–996, Oct 2020. [Online]. Available: https://doi.org/10.1038/s41551-020-0595-9

[9] E. K. Lee, H. Balasubramanian, A. Tsolias, S. U. Anakwe, M. Medalla, K. V. Shenoy, and C. Chandrasekaran, "Non-linear dimensionality reduction on extracellular waveforms reveals cell type diversity in premotor cortex," *eLife*, vol. 10, pp. 1–52, 2021.

[10] A. M. Litke, N. Bezayiff, E. J. Chichilnisky, W. Cunningham, W. Dabrowski, A. A. Grillo, M. Grivich, P. Grybos, P. Hottowy, S. Kachiguine, R. S. Kalmar, K. Mathieson, D. Petrusca, M. Rahman, and A. Sher, "What does the eye tell the brain?: Development of a system for the large-scale recording of retinal output activity," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 4, pp. 1434–1440, Aug. 2004.

[11] E. S. Frechette, A. Sher, M. I. Grivich, D. Petrusca, A. M. Litke, and E. J. Chichilnisky, "Fidelity of the ensemble code for visual motion in primate retina," *J. Neurophysiol.*, vol. 94, no. 1, pp. 119–135, Jul. 2005.

[12] D. G. Muratore and E. J. Chichilnisky, "Artificial retina: A future Cellular-Resolution Brain-Machine interface," in *NANO-CHIPS 2030: On-Chip AI for an Efficient Data-Driven World*, B. Murmann and B. Hoefflinger, Eds. Cham: Springer International Publishing, 2020, pp. 443–465.

[13] B. Gosselin, "Recent advances in neural recording microsystems," *Sensors (Basel)*, vol. 11, no. 5, pp. 4572–4597, Apr. 2011.

[14] J. Wang, Y. Hua, and Z. Zhu, "A 10-bit reconfigurable ADC with SAR/SS mode for neural recording," *Analog Integrated Circuits and Signal Processing*, vol. 101, no. 2, pp. 297–305, Nov. 2019.

[15] D. G. Muratore, P. Tandon, M. Wootters, E. J. Chichilnisky, S. Mitra, and B. Murmann, "A Data-Compressive Wired-OR readout for massively parallel neural recording," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1128–1140, Dec. 2019.

[16] P. Giubilato and W. Snoeys, "OrthoPix: A novel compressing architecture for pixel detectors," in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*. IEEE, Oct. 2012, pp. 1735–1742.

[17] M. Pachitariu, N. A. Steinmetz, S. N. Kadir, M. Carandini, and K. D. Harris, "Fast and accurate spike sorting of high-channel count probes with kilosort," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.