An Electrophysiology-Optogenetics Closed-Loop Bi-Directional Neural Interface for Sleep Regulation With 0.2-μJ/class Multiplexer-Based Neural Network

Chao Zhang, Fenyan Zhang, Dawid Sheng, Zhixiong Ma, Yongxiang Guo, Chao Sun, Yuwei Zhang, Wenxin Zhao, Xing Sheng, *Senior Member, IEEE*, Tongfei A. Wang, and Milin Zhang, *Senior Member, IEEE*

***Abstract*—This work proposed a multiplexer-based neural network (MUXnet), a multiplexer-based, multiplier-free neural network (NN) structure applicable to the implementation of all inner product-based NN layers. An on-chip MUXnet-based neural signal processing unit (NSPU) was designed, achieving a state-of-the-art accuracy of 82.4% on a public human sleep staging dataset, with the lowest reported energy consumption of 0.2 μJ per classification. Building on MUXnet, this work introduced the first electrophysiology-optogenetics closed-loop bi-directional neural interface integrating neural signal recording, on-chip deep NN (DNN)-based sleep staging, and optogenetic stimulation. An in vivo sleep regulation experiment was conducted on animal subjects, demonstrating high efficiency in waking subjects from the non-rapid eye movement (NREM) sleep stage.**

***Index Terms*—Closed-loop, multiplier-free neural network (NN), neural interface, optogenetics, sleep staging.**

I. Introduction

CLOSED-LOOP neural modulation plays a crucial role in frontier neuroscience research, particularly in areas such as neural rehabilitation and sleep regulation [1], [2], [3]. A typical closed-loop neural modulation system comprises three key modules: a neural signal recorder for data acquisition, a neural signal processing unit (NSPU) for data processing, and a stimulator for outward interaction.

Traditional electrical stimulation often lacks sufficient spatial resolution, a limitation that can be addressed by using optogenetics as an alternative [4]. Optogenetics [5] has become a powerful tool in neuroscience research due to its ability to mimic natural neural activity. Optical fibers are commonly employed to transmit light signals from external sources, targeting specific brain regions that express opsins, thereby enabling controlled activation or inhibition of neural circuits [6]. However, most optical fiber-based solutions are wired, making them unsuitable for experiments involving freely moving animal subjects. To minimize interference from devices, microscale light-emitting diodes (micro-LEDs) have been adopted for wireless, optogenetics-based neural modulation [7], [8].

Bi-directional neural interfaces supporting wireless optogenetic stimulation have been reported in the literature [9], [10]. However, an NSPU that combines both low power consumption and high accuracy on demanding tasks for intelligent interactions between the recording component and the stimulation module remains absent. A linear least squares (LLSs) classifier employing basic signal features such as amplitude and entropy was reported in [10] for epilepsy seizure detection. Due to its limited number of parameters, the LLS-based NSPU struggles to handle more complex electrophysiological processing tasks, such as sleep staging, which is the core task in sleep regulation, with potential to be crucial for memory enhancement [11] and chronic disease risk management [12].

In electrical stimulation-based neural interfaces, a NeuralTree-based classifier was introduced in [13] for epilepsy modulation, enhancing the traditional decision tree (DT) model by replacing its nodes with small neural networks (NNs), thereby improving representational capacity. Recent works [14], [15] have proposed a matrix-vector multiplication (MVM)-based NSPU and an explainable boosting machine (EBM)-based NSPU for on-chip classification. However, because these solutions rely on a separate architecture for the feature extractor (FE) and classifier, achieving satisfactory classification accuracy in complex tasks remains challenging. End-to-end NNs have been adopted for peripheral nerve signal processing in designs [16] and [17]. Nevertheless, in both cases, the NNs are implemented off-chip, limiting their integration efficiency in fully embedded closed-loop systems. End-to-end NNs offer reliable accuracy for neural

Received 7 January 2025; revised 3 October 2025; accepted 24 November 2025. Date of publication 18 February 2026; date of current version 1 July 2026. This work was supported in part by Beijing Municipal Natural Science Foundation under Grant Z220015. *(Corresponding author: Milin Zhang.)*

Chao Zhang, Dawid Sheng, Yongxiang Guo, Wenxin Zhao, and Xing Sheng are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

Fenyan Zhang, Zhixiong Ma, and Tongfei A. Wang are with the Chinese Institute for Brain Research, Beijing 102206, China.

Chao Sun and Yuwei Zhang are with Beijing Ningju Technology Ltd., Beijing 100190, China.

Milin Zhang is with the Department of Electronic Engineering, Institute for Precision Medicine, Tsinghua University, Beijing 100084, China (e-mail: zhangmilin@tsinghua.edu.cn).

Digital Object Identifier 10.1109/JSSC.2025.3639446

signal processing tasks, but their high power consumption renders them impractical for implantable applications. The inner product operation is fundamental to the NN inference, with multipliers consuming at least twice as much power as adders. To address this issue, multiplier-free NNs have been developed, including binary NNs [18], AdderNet [19], and shift-based deep NNs (DNNs) [20], all of which reduce power consumption by eliminating multipliers during inference. Given that multiplication follows a consistent rule, table lookup methods have also been explored as a viable approach for multiplier-free NNs [21], [22]. Several ASICs have been proposed to implement NN inference using table lookup techniques [23], [24], [25], [26]. However, these existing solutions rely heavily on the lookup tables (LUTs), which require substantial random-access memory (RAM). Since memory access is a significant contributor to power consumption in modern NN processors, these LUT-based approaches face challenges in power efficiency.

This work presents a multiplexer-based NN (MUXnet) to realize multiplier-free on-chip NN inference, significantly reducing the power and memory overhead. In addition, a fully integrated closed-loop system is proposed, which incorporates a neural signal recorder, an end-to-end MUXnet-based NSPU, and an optogenetic stimulation module.

The rest of this work is organized as follows. Section II introduces the design of MUXnet. The details of the proposed closed-loop system are illustrated in Section III. Experimental results are summarized in Section IV, while Section V concludes the entire work.

II. Design of MUXnet

A. Fundamentals of Table Lookup-Based Inner Product

In most NN layers, inner product calculation serves as a fundamental operation, expressed as

$$y = \mathbf{W}\mathbf{x} = \sum_{i=0}^{n-1} W_i x_i \tag{1}$$

where y, $\mathbf{W}$, $\mathbf{x}$, and n are the output, the weight, the input activation, and the vector length for the inner product, respectively. In typical edge devices, the input activation, $\mathbf{x}$, is quantized into a fixed-point vector as

$$\mathbf{x} = \sum_{j=0}^{k-1} \mathbf{x}^j 2^{j-p} \tag{2}$$

where k represents the quantization bit number and p specifies the position of the decimal point. This work utilizes integer quantization, which is equivalent to setting p as 0 in (2). Considering the contribution of every single bit in $\mathbf{x}$, the inner product result, y, can also be decomposed as

$$y = \mathbf{W}\mathbf{x} = \sum_{j=0}^{k-1} \mathbf{W}\mathbf{x}^j 2^j. \tag{3}$$

Equation (3) can also be written as

$$y = \sum_{j=0}^{k-1} y^j 2^j. \tag{4}$$

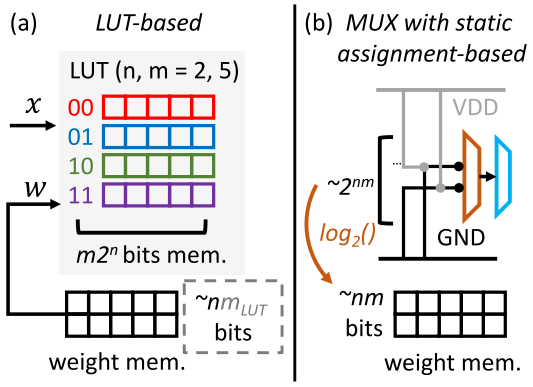


Fig. 1. LUT-based table lookup and MUX-based table lookup. (a) LUT-based. (b) MUX with static assignment-based.

In (4), y^j is defined as

$$y^j = \mathbf{W}\mathbf{x}^j, \quad j \in [0, k-1], \; y^j \in \left[-2^{m-1}, 2^{m-1}\right) \tag{5}$$

where $\mathbf{x}^j$ is the jth bit of the input activation, $\mathbf{x}$. It is important to note that y^j is not a single-bit value, as it represents the inner product result of $\mathbf{W}$ and $\mathbf{x}^j$. The quantization bit number of y^j is assumed to be m. Equation (5) provides the theoretical foundation for the bit-serial input scheme, which is widely utilized in table lookup-based inner product calculations under a weight-reuse manner. This approach significantly reduces the table size by handling a single-bit vector of input activation at each step in the table lookup unit. The vector length for each table-lookup, n, and the quantization bit width for each table-lookup result, m, are fundamental factors in determining the cost of a table lookup-based inner product, and their impact will be analyzed throughout the derivation in this work.

B. LUT-Based Table Lookup

In a classic LUT-based inner product, the calculation results are stored in the built-in RAM, as illustrated in Fig. 1(a). The utilization of the LUT allows the external logic to retrieve the results by changing the enabled address line of the built-in RAM, without any computing. A bit-serial input scheme is assumed to be utilized to limit the size of the LUT. Under an inner product task with a vector length of n and output quantization bits of m, each address in the LUT needs to link to m bits of memory. Even with the bit-serial scheme, the number of address lines for the LUT is up to 2^n, since each element of the input vector $\mathbf{x}^j$ features two possible single-bit values (1/0). As a result, an LUT solution requires $m2^n$ bits of built-in RAM.

For on-chip NN inference, the weights of the NN are stored in memory to determine the configuration of the LUT at each step. A weight memory of size nm_{LUT} is required to store the weights used in the above inner product, where m_{LUT} is the bit width of each weight element. Therefore, the LUT-based method requires a total of $m2^n + nm_{\mathrm{LUT}}$ bits of memory, which mainly arises from the fact that the LUT is essentially implemented as RAM.

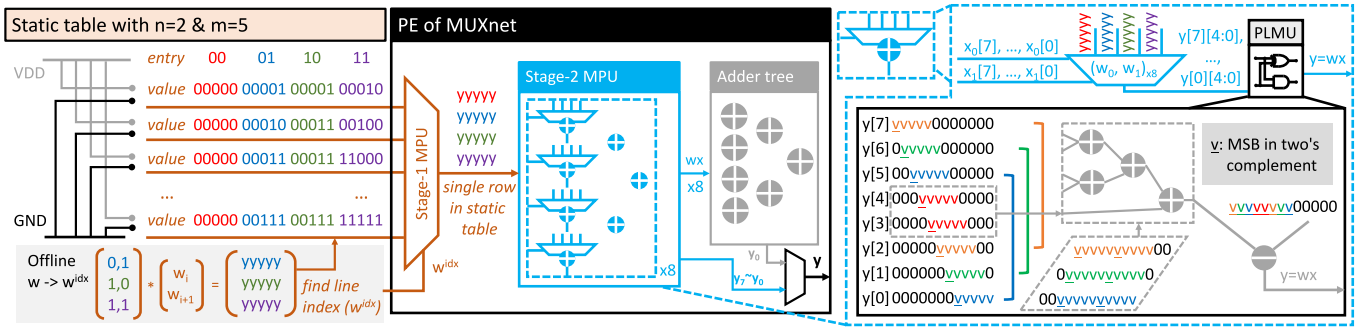


Fig. 2. Details of the ST and PE.

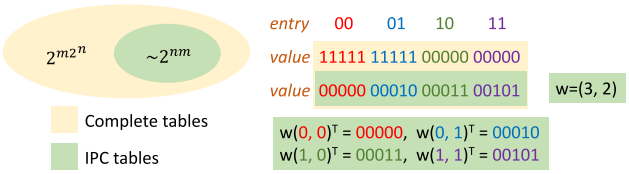


Fig. 3. Complete tables and IPC tables.

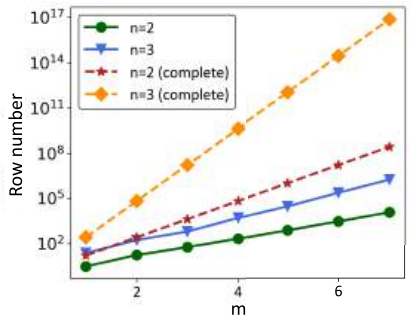


Fig. 4. Row numbers of different tables.

C. Proposed MUX-Based Table Lookup

This work proposes a MUX-based table lookup approach as illustrated in Fig. 1(b) for ultra-low-power scenarios such as neural interfaces. The MUX-based method stores all possible inner product results y^j in (5) on-chip, featuring a post-inner product quantization scheme and a RAM-configuration-free table lookup. The operation of the proposed MUX-based method relies on a static table (ST) as shown in Fig. 2. The entries of the ST represent all possible conditions of the single-bit input activation $\mathbf{x}^j$. Each row in the ST represents the inner product results between all the entries and a network weight vector of length n. When a quantization bit width of m is applied, each value in the row requires m bits. The values in the ST are directly wired to VDD or GND to represent 1 or 0, eliminating the need for memory.

The proposed solution introduces a dilemma between the configurability and the implementability of the ST. Since the ST values are hardwired, all possible inner product results should be implemented on-chip to make sure the on-chip NN can be updated according to newly trained weights. If all possible inner product results can be found on-chip, it is unnecessary to store the actual weights in the weight memory. Instead, the row indices of the ST can be stored, and during computation, the required rows can be selected based on these indices. When the model is updated, it is sufficient to update the row indices in the weight memory according to the new model weights. The derivation of the column count for the ST is similar to the number of address lines in the LUT-based method, which is 2^n, as the input vector consists of n single-bit values (1 or 0) in a bit-serial manner. An m-bit value results in 2^m situations, meaning that the number of situations for one row is 2^{m2^n}, as there are 2^n values in each row. In other words, 2^{m2^n} rows are required to cover all possible situations in a single ST, leading to an extremely high area requirement, making hardware implementation infeasible. However, it should be noticed that the ST is only used for inner product calculations, implying that most of these rows are actually unnecessary. For instance, if the current entry is 00, the corresponding value will not contain any 1 under an inner product rule, as shown in Fig. 3. The set of rows covering all possible situations in inner product calculations is only a small subset of the complete ST. The small subset is denoted as the inner product compatible ST (IPC-ST), which contains approximately 2^{nm} rows. The derivation of the row count is provided in Appendix A. By applying small values of n and m, the size of the IPC-ST becomes quite acceptable compared to the complete ST on hardware, as illustrated in Fig. 4. As a result, although the IPC-ST is non-configurable due to the static assignment with VDD and GND, the proposed MUXnet supports arbitrary weight vectors in an NN by enumerating all possible inner product results on-chip.

According to the above analysis, a weight memory is also required to identify the appropriate line of the IPC-ST at each step of the NN inference in the MUX-based method. The elements in the weight memory are the indices of the IPC-ST rather than the trained weights. After the model is trained, the weights can be mapped to the corresponding IPC-ST indexes, as further explained in Algorithm 1. Since the row count of the IPC-ST is approximately 2^{nm}, the bit length required for each index can be calculated as

$$\log_2(2^{nm}) = nm. \tag{6}$$

In the post-inner product quantization of MUXnet, the quantization target is y^j in (5) instead of network weight $\mathbf{W}$.

Algorithm 1 Mapping Process From Trained Weights to IPC-ST Indexes of the MUXnet

Input:

Trained weights: $\mathbf{W}$

IPC-ST with specified n and m: $\mathbf{T}$

Output:

Values to be stored into the weight memory, $\mathbf{W}^{idx}$ (i.e., the indexes of the IPC-ST)

Find the position of the indexes:

for all weight vector $\mathbf{W}_i$ with length of n **do**
  Calculate the inner product result between $\mathbf{W}_i$ and all situations of the IPC-ST entries $\mathbf{x}$ ($\mathbf{y}_i = \mathbf{W}_i\mathbf{x}$)
  for all row $\mathbf{T}_j$ of the IPC-ST **do**
    Stop when $\mathbf{T}_j = \mathbf{y}_i$
    The value of W_i^{idx} is j
  end for
end for
return $\mathbf{W}^{idx}$

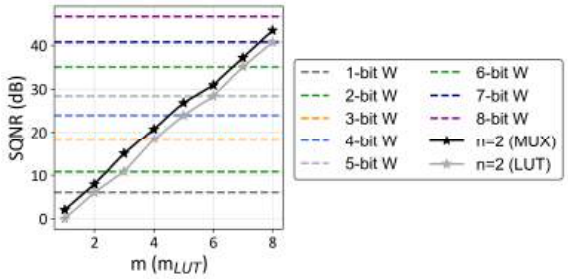


Fig. 5. Relationship between the quantization bit width and the SQNR under n of 2.

In contrast, in a typical LUT-based or multiplier-based inner product, $\mathbf{W}$ is quantized and stored in the weight memory. Since the computation of the network introduces an accumulation of quantization errors, in general, the closer a quantization layer is to the output, the smaller the overall quantization noise becomes. As illustrated in Fig. 5, a numerical experiment was conducted to compare the quantization error level of the MUX-based table lookup method with typical LUT-based and multiplier-based inner product methods. The three methods were evaluated on a linear layer with a shape of (64, 64). First, the output y_{true} of the linear layer was calculated without any quantization. Then the hardware-level outputs y_{mux}, y_{lut}, and $y_{\text{multiplier}}$ were calculated based on quantized input data $\mathbf{x}$ and quantized weight $\mathbf{W}$ (with quantized y^j for the MUX-based method) for the three methods, respectively. In the conducted experiment, the input data $\mathbf{x}$ was quantized into 8 bits, while the vector length for each inner product was fixed at $n = 2$. Both the bit width of the network weights $\mathbf{W}$ and the table lookup results range from 1 to 8, which covers typical weight bit widths in edge devices. To evaluate the quantization error level more robustly, the signal-to-quantization-noise ratio (SQNR) was adopted as the evaluation metric. SQNR is defined as

$$\text{SQNR} = 10\log_{10}\left(\frac{\mathbb{E}\left[|y_{\text{true}}|^2\right]}{\mathbb{E}\left[\left|y_{\text{quantized}} - y_{\text{true}}\right|^2\right]}\right). \tag{7}$$

A higher SQNR indicates smaller quantization distortion relative to the original output, and thus reflects better fidelity of the quantized computation. It can be seen in Fig. 5 that both the MUX-based and LUT-based methods show lower SQNR compared to the multiplier-based method under the same quantization bit width, which raises from that the dynamic ranges of weights have to be reduced by $n\times$ in order to avoid the overflow issue in computation. However, when m_{LUT} equals m, the SQNR of the MUX-based method is always higher than the LUT-based method, showing the benefit of the post-inner product manner. As a result, the proposed MUX-based method at least dramatically reduces the memory usage from $m2^n + nm$ to nm compared to the LUT-based table lookup method.

D. Table Lookup Flow of MUXnet

The architecture of the process engine (PE) in the proposed MUXnet is illustrated in Fig. 2, comprising a two-stage MUX-based processing unit (MPU) and an adder tree. The computation in the two-stage MPU relies on the IPC-ST. In Stage-1, four MUXs operate in parallel to select rows from the IPC-ST according to the current row indexes stored in the weight memory.

Once a specified row of the IPC-ST is selected by Stage-1, Stage-2 retrieves the inner product results based on the input activations, which serve as the entries of the IPC-ST. Similar to the bit-serial scheme, Stage-2 processes a single bit of the input vectors at a time, so a MUX size of 2^n-to-1 is sufficient for Stage-2. Since an all-zero input always produces an all-zero output according to the inner product rule, the MUX size can be further reduced to $2^n - 1$-to-1. It is important to note that there are no dependencies between the calculations for different bit positions, as described in (2)–(5). Therefore, all bits can be processed by Stage-2 simultaneously. The proposed PE outputs the inner product result of vectors with a length of 64 in one clock cycle, where 64 is empirically chosen in order to reduce tail handling and improve utilization in general scenarios. The required number of MUXs in Stage-2 can be calculated as

$$N_{\text{Stage2-MUX}} = \frac{64}{n} \times k \tag{8}$$

where n and k follow the previous definition in (1) and (2), and are set to 2 and 8 in this work, respectively. As a result, there are 256 3-to-1 MUXs in Stage-2.

Multiple m-bit results are then accumulated through a parallel post-lookup merging unit (PLMU), producing the results for 32 groups of 2×2 inner products. A local adder array within the Stage-2 MPU is then recruited to merge the 32 groups of 2×2 inner product results into eight groups of 8×8 inner product results. Fig. 6 demonstrates an example of implementing two groups of 8×8 inner products using the Stage-2 MPU, where the n multipliers and n–1 adders required in a standard inner product operation are replaced by a single $2^n - 1$-to-1 MUX and one PLMU.

The adder tree in the PE accumulates the eight inner product results generated by the Stage-2 MPU. By controlling whether the adder tree is enabled, the PE supports a convenient switch between the convolutional layers and the linear layers.

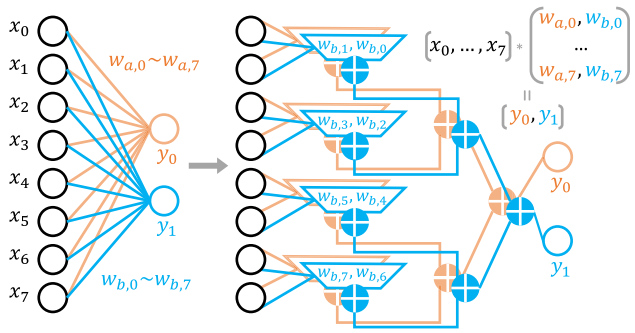


Fig. 6. Example of the inner product in MUXnet.

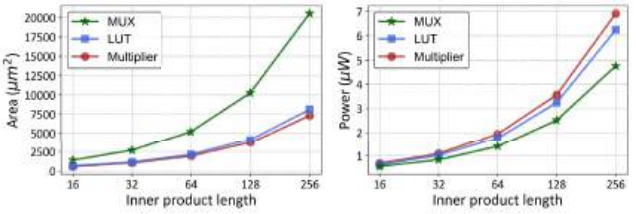


Fig. 7. PE-level comparison between the MUX-based, the LUT-based, and the multiplier-based inner product.

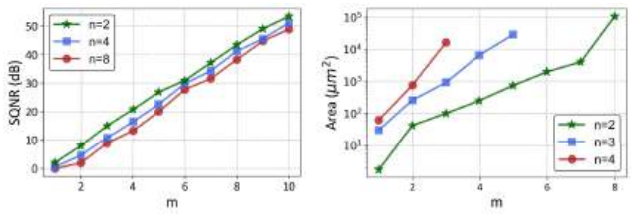


Fig. 8. SQNR and area overhead of the IPC-ST and stage-1 MUX with different n and m.

E. Tradeoff Between Power Consumption and Area With Different Inner Product Methods

The PE-level simulated power and silicon area consumption of the proposed MUX-based inner product operation, the typical LUT-based operation, and the typical multiplier-based operation in a weight reuse scheme are shown in Fig. 7. It can be seen that the MUX-based method shows a larger area since several large MUXs are recruited. Although there is an extra cost of the MUXs, the MUX-based method still shows lower power consumption since it eliminates the requirement of RAM writing and multiplier usage compared to the LUT-based and multiplier-based methods, respectively, ensuring its usability in power-sensitive applications such as neural interfaces. Specifically, in the utilized design point of a 64 × 64 inner product, the MUX-based inner product achieves 0.21× and 0.28× lower power consumption compared to a typical LUT-based and multiplier-based inner product, respectively.

F. Tradeoff Between Quantization Error and Area With Different n and m

Fig. 8 shows the SQNR and area overhead of the IPC-ST and stage-1 MUX with different n and m. It can be summarized that the smaller n and m mean a smaller area of the ST and stage-1 MUX, since the number of rows in the ST is approximately proportional to 2^{nm}. It's natural that a larger m would make the accuracy better, since m is the quantization bit width of y^j in (5). However, the larger n would make the accuracy worse, since given a specific m, when n gets larger, the dynamic range of network weights would be smaller in order to avoid the overflow issue in the quantization of y^j. The benefit of using a larger n is to reduce the size of the adder array after the table lookup.

For each area curve, if m is increased by one at its right end, the number of rows in the IPC-ST would exceed 100 000, making it difficult to obtain simulation results of the area within the limited computational resources. As a result, in order to ensure the flexibility of using different values of m, this work employs n of 2.

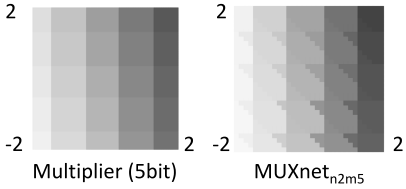


Fig. 9. Non-linear quantization in MUXnet.

G. Implicit Quantization Scheme

According to the post-inner product quantization manner in the MUX-based method, the weight vector $\mathbf{W}$ does not need to be quantized and stored on-chip. Instead, y^j is quantized into m bits and stored on-chip using wire assignment to VDD or GND. Given one weight vector $\mathbf{W}_a$ and another weight vector $\mathbf{W}_b$ with relationship as

$$\mathbf{W}_b = \mathbf{W}_a + \triangle\mathbf{W}. \tag{9}$$

Since y^j is quantized, it's straightforward to get

$$\mathbf{W}_a\mathbf{x}^j = \lim_{\Delta\mathbf{W}\to 0} \mathbf{W}_b\mathbf{x}^j \tag{10}$$

which means the quantization pattern for $\mathbf{W}$ is implicit in the MUX-based method. Fig. 9 illustrates the implicit quantization scheme, taking n of 2, m of 5, and weights between (−2, −2) and (2, 2) as an example. There are two features in Fig. 9, namely position and color. Each position represents one sampled weight vector $\mathbf{W} = (W_0, W_1)$. For each weight vector, all possible y^j are calculated as

$$\mathbf{y}^j_{\text{all}} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} W_0 \\ W_1 \end{pmatrix}. \tag{11}$$

Each color represents one unique $\mathbf{y}^j_{\text{all}}$. In other words, if the positions of two weight vectors show the same color, the two weight vectors can be seen as the same weight vector after quantization. According to the above definition of position and color, it's intuitive to show the quantization manner of the typical inner product method. Since the $\mathbf{W}$ needs to be quantized at the beginning, the division lines of the quantized

Fig. 10. Pre-scaled weight scaling scheme.

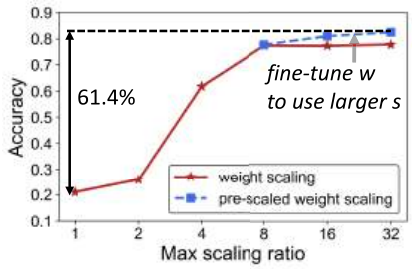


Fig. 11. Results of the proposed pre-scaled weight scaling.

W in the figure correspond to the rounding boundaries, i.e., values such as 0.5 or 1.5. According to the experiment in Fig. 5, the 5-bit weights in the typical inner product method are used for comparison with the MUX-based method with m of 5. The average quantization interval is defined as the average area of each color, in which the MUX-based method is 0.69× smaller than the typical inner product method, indicating the potential to reduce the computation error.

H. Weight Scaling

The hardware-level inference exhibits a larger relative quantization error when dealing with small weights. Since MUXnet stores the row indexes of the IPC-ST instead of the trained weights in the weight memory, it can freely perform weight scaling to map a trained weight to an IPC-ST index with a lower relative quantization error, as illustrated in Fig. 10. In addition, before the weight scaling is applied, a pre-scaling step that slightly upscales or downscales the weights to achieve a larger feasible s in the weight scaling is utilized, where s is the scaling ratio in Fig. 10. As demonstrated in Fig. 11, the pre-scaled weight scaling reduces the quantization error without requiring additional memory, resulting in a 61.4% improvement in accuracy for the sleep staging task.

I. Table Decomposition

As derived in Section II-C, the row count of the IPC-ST is estimated to be 2^{nm}, making it feasible to implement the entire IPC-ST on-chip under small values of n and m. However, when low quantization error is required, MUXnet has to recruit a large value for m. The explosive growth of 2^{nm} makes it not only impossible to implement the IPC-ST on-chip, but also difficult to simulate the power and area consumption as m increases. In order to address this issue, a table decomposition scheme is proposed to replace one large IPC-ST with two smaller STs, as outlined in Algorithm 2. In detail, given y^j in (5) quantized to m bits, the table decomposition scheme

Algorithm 2 Procedures of the Table Decomposition

Input:

The inner product length for the IPC-STs: n
The required quantization bit number of IPC-ST: m_i

Output:

A ST, $\mathbf{T}_l$, for the low m_{ol} bits of the inner product result
An IPC-ST, $\mathbf{T}_h$, for the high m_{oh} bits of the inner product result

Get the row values $\mathbf{rv}_l$ and $\mathbf{rv}_h$ for the two output STs:

Determine the value of m_{ol} and m_{oh}, where $m_{ol}+m_{oh} = m_i$
Note all possible weights under a quantization bit of m_i, as **W**
Note all possible single-bit input activations under an inner product length of n, as **x**
Calculate $\mathbf{rv} = \mathbf{W} \times \mathbf{x}$, where **rv** includes all the possible results of the inner product
Calculate $\mathbf{rv}_l = \mathbf{rv}\ mod\ 2^{m_{ol}}$
Calculate $\mathbf{rv}_h = \mathbf{rv}\ //\ 2^{m_i - m_{oh}}$

Generate the two STs $\mathbf{T}_l$ and $\mathbf{T}_h$:

Init a blank ST $\mathbf{T}_l$ for the low part of the inner product
for all row $\mathbf{rv}_l^i$ of the $\mathbf{rv}_l$ **do**
Insert $\mathbf{rv}_l^i$ into $\mathbf{T}_l$ when $\mathbf{rv}_l^i$ is not in T_l
end for
Use the same produces to get $\mathbf{T}_h$ for the high part of the inner product
return $\mathbf{T}_l$, $\mathbf{T}_h$, m_{ol}, m_{oh}

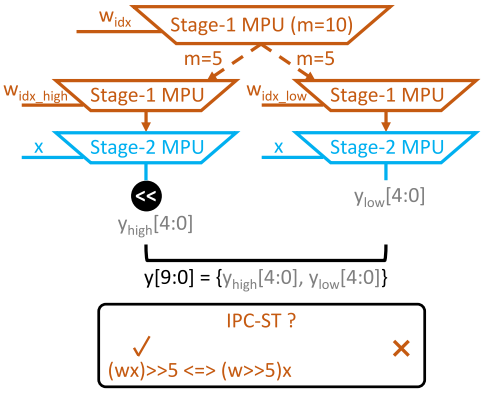


Fig. 12. Illustration of the table decomposition.

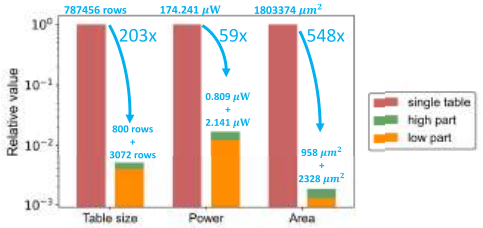


Fig. 13. Benefits of the proposed table decomposition.

calculates y_h^j and y_l^j, with bit widths of m_{oh} and m_{ol} in Algorithm 2, respectively. Then y^j can be reconstructed by bit-level concatenation of y_h^j and y_l^j. Fig. 12 illustrates an example of decomposing an IPC-ST with m of 10 into two

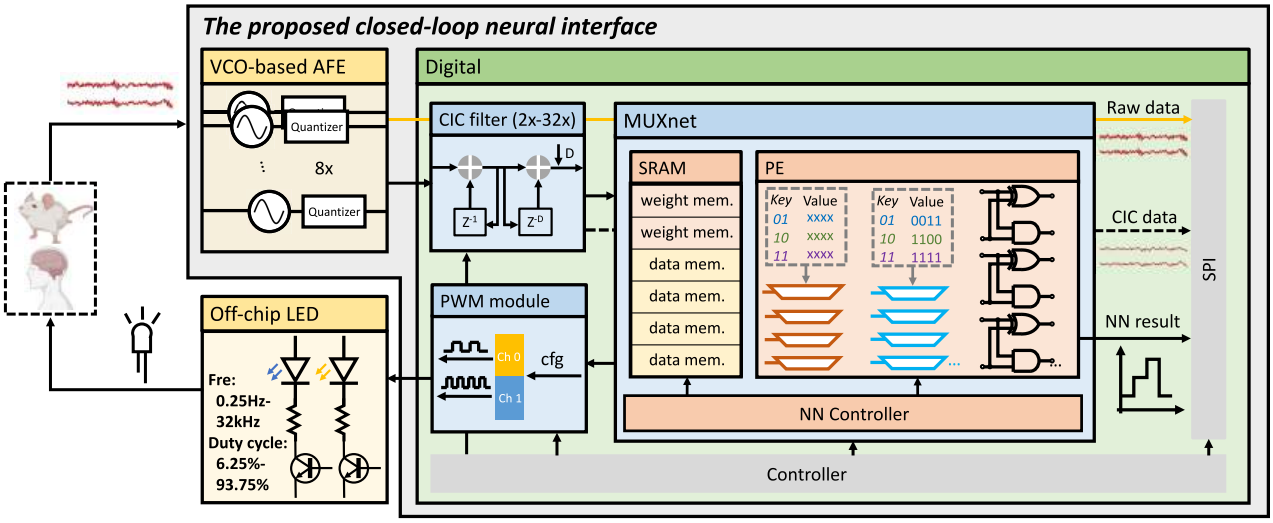


Fig. 14. Block diagram of the proposed neural interface.

STs, where m takes the value 5, with all three STs using n of 2. It should be noticed that the ST for the lower 5 bits becomes non-IPC, resulting in a slight increase in the row count. Fig. 13 shows the estimated benefits of the table decomposition approach, demonstrating a dramatic reduction of 59× in power consumption and 548× in area occupation.

III. Closed-Loop Bi-Directional System

A. System Overview

Fig. 14 shows the block diagram of the proposed closed-loop system. An eight-channel voltage-controlled oscillator (VCO)-based analog front-end (AFE) with a sampling rate of up to 24 kHz is integrated for neural signal acquisition, supporting a wide range of recording scenarios from non-invasive EEG to invasive action potentials of single neurons. A cascaded integrator-comb (CIC) decimation filtering module is implemented for 2×–32× data compression, relieving the computational workload of the subsequent signal processing module. The proposed MUXnet is implemented for multiplier-free, low-power NN inference, functioning as the on-chip NSPU. It features a standard NN processor architecture, consisting of separate memory, PE, and controller units. The memory is divided into six SRAM sub-blocks, each of which can be dynamically gated when the active memory address is distant from the local address.

In order to perform the electrophysiology-optogenetics-based closed-loop control, a dual-channel pulsewidth modulation (PWM) module is designed to generate stimulation waveforms with adjustable frequencies ranging from 0.25 Hz to 32 kHz and duty cycles ranging from 6.75% to 93.25%. The system includes an implantable, wireless optogenetic probe equipped with a blue thin-film micro-LED, which targets specified tissue based on the classification results from the NSPU or the direct register configuration.

The proposed system shares data with an external device through an SPI secondary. The output data from the SPI can be selected from the raw data recorded by the AFE, the data filtered and compressed by the CIC filter, or the classification results of the NSPU, supporting applications with varying data resolution requirements and power budgets.

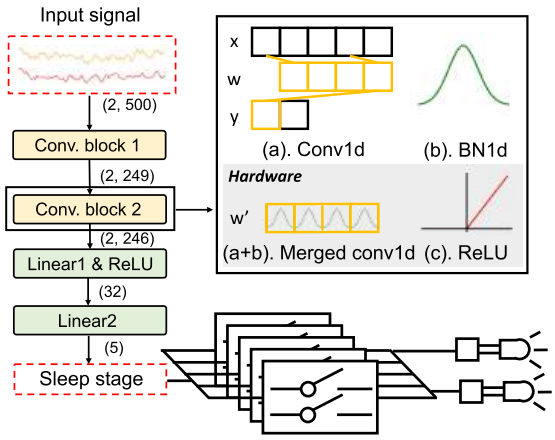


Fig. 15. Computing flow.

B. MUXnet-Based NN Architecture and Closed-Loop Protocol

Fig. 15 illustrates the architecture of the designed four-layer convolutional NN (CNN) for sleep staging. A dual-mode MUXnet with m of 5 or 10 and n of 2 is implemented to ensure both robust sleep staging and low hardware overhead in energy and area. The convolutional layers utilize m of 10 to achieve a lower quantization error, while the linear layers utilize m of 5 to reduce the memory usage. The IPC-ST with m of 10 is decomposed into two smaller STs with m of 5 according to the proposed table decomposition approach in Section II-I.

The 1-D batch normalization (BN) layers are integrated into the 1-D convolutional layers to eliminate the need for separate memory storage of the BN parameters, as well as to reduce the latency and energy costs associated with the BN inference. The proposed system allows each optogenetic

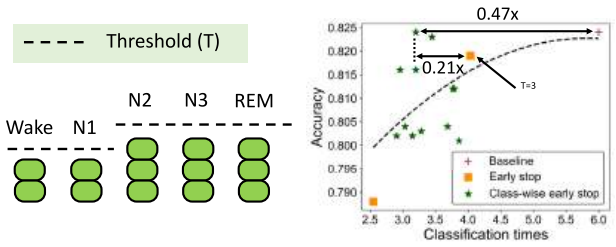


Fig. 16. Class-wise early stop voting scheme.

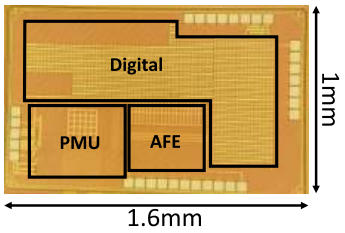


Fig. 17. Photography of the fabricated neural interface.

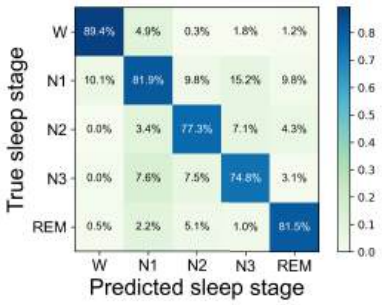


Fig. 18. Confusion matrix on DREAMS dataset.

channel to be triggered independently based on any prediction (0–9) produced by the NSPU, making the system adaptable to arbitrary closed-loop scenarios.

C. Early Stop Voting Scheme

In the human sleep staging task, the predictions from six 5-s segments are combined to determine the sleep stage of one 30-s segment.A class-wise early stop voting scheme controlled by the on-chip controller is proposed to reduce the classification times by up to 46.7%, as shown in Fig. 16. In a six-segment voting scheme, once the final decision can be determined from the first few 5-s segments, the remaining segments do not need to be processed. Each predicted class is assigned a unique threshold stored in registers to reach a majority vote. This voting scheme not only aligns task lengths by constraining the on-chip NN to short segment classification, but also reduces the average energy consumption by removing the unnecessary on-chip data buffering requirements and the redundant network inferences.

IV. Experimental Results

A. Chip Measurement

The proposed design was fabricated using standard 40-nm CMOS technology, occupying a silicon area of 1.6 mm^2,

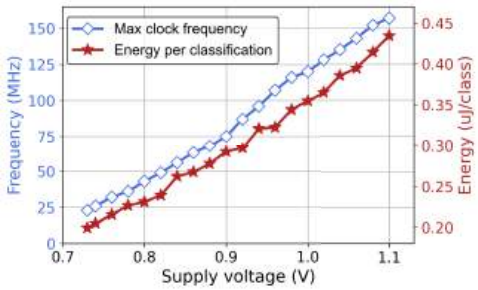


Fig. 19. Energy consumption and the maximum frequency under different supply voltages.

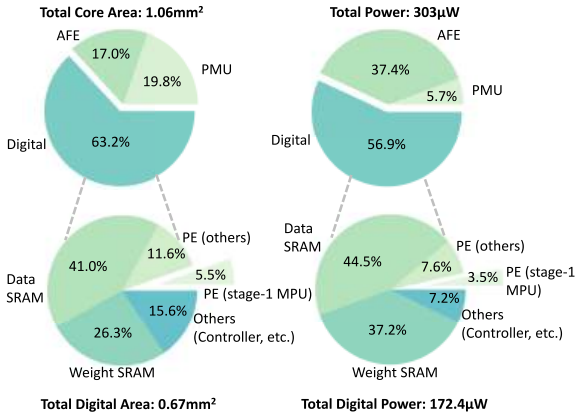


Fig. 20. Area and power breakdown of the designed neural interface.

as shown in Fig. 17. The designed system was verified on a public 20-subject database (DREAMS) [27] and a private mice dataset. An average sleep staging accuracy of 82.4% and 84.3% was achieved, respectively. The confusion matrix of the classification on the DREAMS dataset was shown in Fig. 18. An average power dissipation of 172.4 μW, with an energy consumption of 0.2 μJ per classification, was measured for the digital part of the chip under a supply voltage of 0.73 V and a system clock of 23 MHz, as shown in Fig. 19. The area and power breakdown is shown as Fig. 20.

The AFE exhibits a noise level of 0.96 μV_{rms} in the 0.5 Hz–1-kHz range, an ENOB of 13.8 bits, a linearity of 85 dB after non-linearity calibration (NLC) using a fifth-order polynomial, a PSRR of 74 dB, and a CMRR of 106 dB. The area and power consumption for each channel are 0.015 mm^2 and 10.45 μW, respectively. The PMU integrates two LDOs for the digital circuit and AFE, consuming a static current of 300 nA and 1.75 μA, respectively.

B. In Vivo Test

The proposed neural interface was integrated with an RF chip for wireless in vivo verification, as shown in Fig. 21. A flexible printed circuit (FPC) provided the interconnects for neural recording and stimulation.

One EEG channel and one EMG channel from mice were recorded in the experiments. EEG and EMG electrodes were implanted in mice using a homemade head-mount, which was stabilized by three stainless steel epidural screw electrodes. The first two electrodes (frontal and ground) were located

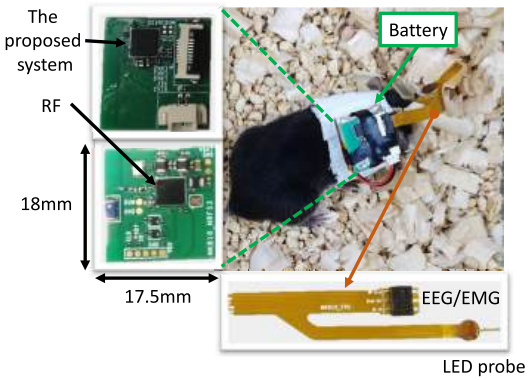


Fig. 21. Setup for in vivo test.

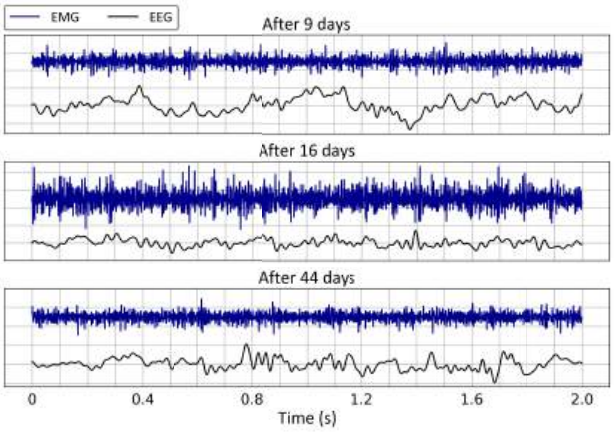


Fig. 22. EMG and EEG during a long-term verification.

TABLE I
INDICATORS FOR THE LONG-TERM RECORDING TEST

Day	9	16	44
	EEG		
Line-noise ratio	0.0002	0.0494	0.0009
Flat-line fraction	0	0	0
Amplitude-outlier fraction	0.0005	0.0003	0
	EMG		
Envelope SNR (dB)	20.9	27.3	21.0
Median frequency (Hz)	110.8	51.3	143.1
Clip fraction	0.001	0.001	0.001

0.5 mm posterior to the bregma and 1.5 mm on either side of the central suture. The third electrode was located 4.5 mm posterior to the bregma and 1.5 mm on the left side of the central suture. EMG activity was monitored using stainless-steel Teflon-coated wires inserted bilaterally into the nuchal muscle. The head mount was secured to the skull with dental acrylic. During implantation, the tip of the LED probe was carefully lowered above the PVT (coordinates, bregma: AP = −0.3 mm; ML = +0 mm; and DV = −3.7 mm) and secured to the skull surface using a thin layer of ultraviolet ray-curing dental cement. All mice were allowed to recover for at least seven days before the recording experiments began.

A long-term recording test was conducted to verify the stability of the proposed setup in mice. Recordings were performed on the same mouse at 9, 16, and 44 days post-implantation. Each session comprised at least 20 min of continuous data sampled at 1 kHz. The quality control scheme was performed using a custom script in MNE-Python. For each EEG channel, the line-noise ratio, flat-line fraction, and amplitude-outlier fraction were computed as indicators. For each EMG channel, the envelope signal-to-noise ratio (SNR), median frequency, and clip fraction were computed as indicators. Further information about these indicators and the thresholds for marking one channel as poor quality based on these factors is provided in Appendix B. The signal segments for each of the three days of recording and the corresponding indicators of signal quality control are shown in Fig. 22 and Table I, respectively. Based on the recorded data, even at 44 days post-implantation, neither the EEG nor the EMG channels were marked as poor quality, demonstrating the long-term reliability of the system.

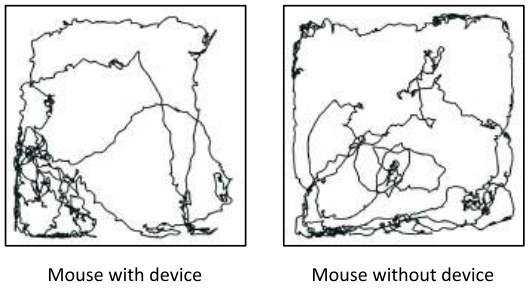


Fig. 23. Traces in the open-field test.

The setup for the in vivo test weighs 3.8 g, including a 2.5-g battery. An open-field test was conducted to assess whether wearing the device affects the spontaneous locomotion of mice. Two mice were recruited for the open-field test. One served as the experimental group and wore the device for data acquisition; the other served as the control and did not wear the device. The two mice were tested separately. The experiment was conducted in a white box with dimensions of 39 × 39 × 47 cm. There were no obstacles other than the side walls of the box. At the start of each trial, the mouse was placed at the center of the floor and then left undisturbed. Each session lasted 9 min and was video-recorded by a camera positioned above the box. After the session, the locomotor trajectory was extracted and calibrated using the tool in [30]. Fig. 23 shows that both the mouse with the implanted setup and the mouse in the control group exhibited unrestricted movement within the box, confirming that the setup does not affect their mobility.

A sleep regulation test was performed to verify the proposed neural interface in closed-loop applications. For optogenetic manipulation, AAV2/9-EF1a-DIO-hChR2-mCherry was injected into the aPVT of the mouse two weeks before implantation to enable the mouse to be awakened by light stimulation during the experiments. A 488 nm blue LED-based optogenetic modulation, with a frequency of 10 Hz and a duty cycle of 40%, was triggered when a non-rapid eye movement (NREM) phase was detected.Significant changes in EEG and EMG signals were observed following optogenetic stimulation, as shown in Fig. 24, demonstrating the effectiveness of the proposed work for closed-loop applications. The stimulation was performed on the same mouse eight times in total, with an

TABLE II
COMPARISON WITH STATE-OF-THE-ART ASICS

	ISSCC'22 [13][a]	ISSCC'23 [17][a]	ISSCC'24 [14][a]	ISSCC'18 [9][b]	ISSCC'19 [10][b]	JSSC'17 [28][c]	TCAS-I'19 [29][c]	This work
General								
Technology	65nm	65nm	180nm	130nm	180nm	180nm	180nm	**40nm**
Area (mm^2)	8	5	4.86	1.2	13.67	10.3	11.74	**1.6**
NSPU								
Algorithm	FE + NeuralTree	Off-chip CNN	FE + EBM	Filter	FE + LLS	FE + DT	FE + NN-based DT	**End-to-end CNN**
Application	Epilepsy	Peripheral nerve	Memory enhance	-	Seizure detection	Sleep staging	Sleep staging	**Sleep staging**
Dataset	CHB-MIT iEEG.org	-	-	-	-	DREAMS	SHHS, NTUH	**DREAMS**
Accuracy	>95.6% >94.0%	86.7%	>98.5%	-	-	<80%[d]	77.1%-81.1%	**82.4%**
Supply (V)	1.2	-	1.8/3.3	-	1.8	1.25	1.2	**0.73**
Power (μW)	-	-	319.2	-	114@375Hz	426@1.5kHz	4.96@10kHz	**172.4@23MHz**
Energy (μJ/class)[e]	0.227	-	0.19	-	-	10.5	149	**0.2**
On-chip SRAM	3.17kB	None	None	None	None	None	None	**14kB**
Recorder & Stimulator								
AFE channel	256	64	8(x3)	10	2	-	-	**8**
Area (mm^2/ch)	0.014	0.02	-	0.060	-	-	-	**0.015**
Power (μW/ch)	1.51	5.2	-	11.2	-	-	-	**10.45**
Stim. channel	16	64	8(x3)	4	1	-	-	**2**
Closed-loop	NeuralTree-based	Off-chip NN-based	EBM-based	N	LLS-based	-	-	**NN-based**
Stim. type	Electrical	Electrical	Electrical	Optogenetic	Optogenetic	-	-	**Optogenetic**
Device weight	-	-	-	-	6g	-	-	**3.8g**
In-vivo validation	Y	Y	Y	Y	Y	-	-	**Y**

[a] Neural interface designs in other tasks.
[b] Optogenetics based neural interface designs.
[c] ASICs for sleep staging.
[d] The reported 98.5% is the consistency between hardware and software.
[e] Thirty second per classification.

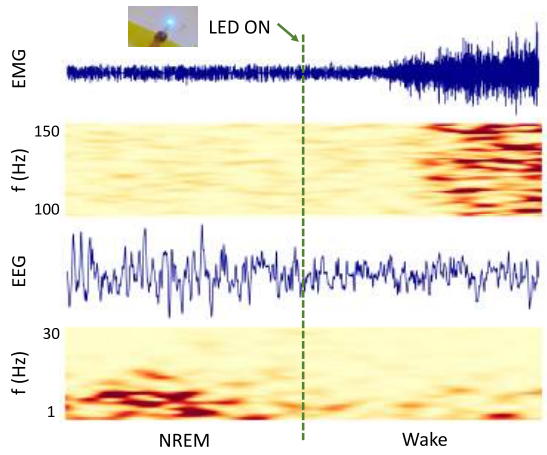


Fig. 24. Results of the sleep regulation test.

average delay of 14.88 ± 5.56 s from the onset of stimulation to wakefulness.

C. Comparison With State-of-the-Art

Table II compares the proposed work with state-of-the-art designs in the literature. As an optogenetic closed-loop neural interface validated on the sleep regulation task, this work is compared against existing sleep staging chips as well as prior optogenetic closed-loop neural interfaces. Benefiting from the end-to-end NN architecture and multiplier-free inference based on the MUX-based method, the proposed design achieves an average accuracy of 82.4% on the open-source sleep staging task, with an average energy consumption of 0.2 μJ per classification—outperforming existing sleep staging chips [28], [29]. Compared with prior optogenetic closed-loop neural interfaces [9], [10], this work implements an end-to-end NN-based NSPU, enabling the execution of more complex tasks and thereby enhancing the generality of the neural interface. In addition, Table II also summarizes BMI designs across different tasks [13], [14], [17]. Although this work addresses more complex tasks, its energy consumption per classification remains comparable to that of these prior studies.

V. CONCLUSION

This work proposed MUXnet, a MUXnet for multiplier-free NN inference, serving as an NSPU in a closed-loop neural interface system. Featuring post-inner product quantization and RAM-configuration-free table lookup, the proposed method achieves a lower quantization error level and lower power consumption than a typical LUT-based table lookup method. The proposed design achieves a state-of-the-art average accuracy of 82.4% and the lowest energy consumption of 0.2 μJ per classification in the ASIC-level sleep staging task. Moreover, the proposed closed-loop neural interface integrates low-noise signal acquisition, end-to-end NN-based on-chip signal processing, and optogenetic stimulation, making it applicable to a variety of general neural interface scenarios. Extensive in vivo testing, including a closed-loop sleep regulation experiment, demonstrates the effectiveness and robustness of the proposed work, establishing it as a powerful tool for cutting-edge neuroscience research.

APPENDIX A
ROW NUMBER OF THE IPC-ST

As mentioned in Section II-A, n and m are two fundamental parameters in the MUX-based table lookup method, where n

denotes the vector length for the inner product and m is the quantization bit width of the single-bit inner product result in a bit-serial manner. The column count of an IPC-ST corresponds to all possible conditions of the single-bit input activation, which totals 2^n. In other words, an IPC-ST contains 2^n entries. Each row of the IPC-ST indicates a specific weight vector, and the values within a row are the inner product results between the implicit weight vector and all entries, with each value quantized to m bits.

According to (5), the range of each value, y^j, in the IPC-ST is determined by

$$y^j \in \left(-2^{m-1}, 2^{m-1}\right). \tag{12}$$

The y^j is derived from the product of $\mathbf{W}$ and $\mathbf{x}^j$ in (5), where each element of $\mathbf{x}^j$ is either 0 or 1. In order to prevent overflow, the range of each weight element, W_i, is shown as

$$W_i \in \left(-\frac{2^{m-1}}{n}, \frac{2^{m-1}}{n}\right), \quad i \in [0, n-1]. \tag{13}$$

The row count of the IPC-ST is essentially the unique weights in the quantized space, which depend solely on n and m. To clarify the derivation, examples under varying values of n are provided. It should be noted that when n equals 1, the operation becomes a direct replacement of multiplication with addition in a bit-serial approach, which lies beyond the scope of MUXnet.

A. $n = 2$

When n takes the value 2, the range of each weight element, W_i, is shown as

$$W_i \in \left(-2^{m-2}, 2^{m-2}\right). \tag{14}$$

Each W_i falls into one of two types of intervals. A type-I interval has a length of 1, with its center at an integer [e.g., $(-0.5, 0.5)$ and $(0.5, 1.5)$]. A type-II interval holds a length of 0.5, with one of its boundary points being an integer [e.g., $(-2, -1.5)$ and $(1.5, 2)$]. A single-bit input activation vector under $n = 2$ is noted as $(x_0, x_1)^T$. The inner product result between a weight vector and a single-bit input activation is denoted as $y_{x_0x_1}$, representing the corresponding value in the IPC-ST. For example, when the single-bit input activation is $(0, 1)^T$, the corresponding table value is y_{01}. According to the inner product rule, the y_{00} always holds a value of zero. The y_{01} and y_{10} are actually the quantized forms of w_1 and w_0, respectively. Once the intervals containing the weight elements are known, the y_{01} and y_{10} are also determined. As a result, given the intervals in which the weight elements reside, the corresponding row number in IPC-ST is determined by the number of situations for y_{11}, which can be identified through grid searching.

Table III lists the corresponding row numbers based on different intervals where the weight elements reside.The number of intervals with a length of 1 is calculated by

$$(L_1) = \left\lceil 2^{m-2} \right\rceil \times 2 - 1 \tag{15}$$

while the following equation shows the number of intervals with a length of 0.5:

$$(L_{0.5}) = 2 \times \left\lceil 0.5 - \left(2^{m-2} \bmod 1\right) \right\rceil. \tag{16}$$

TABLE III
CORRESPONDING ROW NUMBERS FOR DIFFERENT INTERVALS THAT THE WEIGHT ELEMENTS LOCATE ($n = 2$)

Interval length		Side[a]	Corresponding row number
W_0	W_1		
1	1	-	3
1	0.5	-	2
0.5	1	-	2
0.5	0.5	Same	2
0.5	0.5	Different	1

[a] The positions of the two intervals compared with 0.

TABLE IV
CORRESPONDING ROW NUMBERS FOR DIFFERENT INTERVALS THAT THE WEIGHT ELEMENTS LOCATE ($n = 3$)

The number of weight elements in a specified interval length		Side[a]	Corresponding row number
1, 5/6, 2/3	1/6		
3	0	-	23
2	1	-	12
1	2	-	5
0	3	Same	1
0	3	Different	2

[a] The positions of the three intervals compared with 0.

Consequently, the total row number of the IPC-ST with $n = 2$ is expressed as

$$T_{n=2} = 3C^1_{(L_1)}{}^2 + 2C^1_2 C^1_{(L_1)} C^1_{(L_{0.5})} + (2+1)\, C^1_{(L_{0.5})}. \tag{17}$$

B. $n = 3$

Similar to the case where $n = 2$, the range of the weight element is given by

$$W_i \in \left(-\frac{2^{m-1}}{3}, \frac{2^{m-1}}{3}\right). \tag{18}$$

Table IV lists the corresponding row numbers for different intervals in which the weight elements are located. To improve clarity, Table IV presents the count of weight elements in a specified interval length instead of their exact positions.The count of intervals with a length of 1 or 2/3 is calculated by

$$(L_{1,2/3}) = \left(\max\left(1, \text{round}\left(\frac{2^{m-1}}{3}\right)\right) - 1\right) \times 2 + 1. \tag{19}$$

The following equations provide the number of ranges with lengths of 1/6 and 5/6, respectively,

$$(L_{1/6}) = \left(\left(\frac{2^{m-1}}{3} - \frac{1}{6}\right) \bmod \frac{1}{2} == 0\right) \,?\, 2 : 0 \tag{20}$$

$$(L_{5/6}) = \left(\left(\frac{2^{m-1}}{3} + \frac{1}{6}\right) \bmod \frac{1}{2} == 0\right) \,?\, 2 : 0 \;(m > 1). \tag{21}$$

Therefore, the total row count of the IPC-STs with $n = 3$ is expressed as

$$\begin{aligned} T_{n=3} = {} & 23\left(C^1_{(L_{1,2/3})} + C^1_{(L_{5/6})}\right)^3 \\ & + 12C^1_3 C^1_{(L_{1/6})}\left(C^1_{(L_{1,2/3})} + C^1_{(L_{5/6})}\right)^2 \\ & + 5C^1_3 C^2_{(L_{1/6})} + 1C_{(L_{1/6})} + 2C^1_3 C_{(L_{1/6})}. \end{aligned} \tag{22}$$

It can be observed that both (17) and (22) are dominated by their first terms, which grow in proportion to 2^{nm}. For larger values of n, the derivation process follows a similar approach as in cases where n holds values 2 and 3.

APPENDIX B
DETAILS OF THE LONG-TERM RECORDING TEST

For each EEG channel, the following indicators were computed.

1) *Line-Noise Ratio:* Power within a $\pm$1-Hz window around the mains frequency divided by broadband EEG power (1–45 Hz, excluding $\pm$2 Hz around mains).
2) *Flat-Line Fraction:* Fraction of 1-s windows whose peak-to-peak amplitude was smaller than 1 μV.
3) *Amplitude-Outlier Fraction:* Fraction of samples whose absolute amplitude exceeded 200 μV.

A channel was flagged as poor quality if any of the following conditions held: line-noise ratio $>$ 0.05, flat-line fraction $>$ 0.10, or amplitude-outlier fraction $>$ 0.02.

For each EMG channel, the following indicators were computed.

1) *Envelope SNR:* The rms of the top 10% versus bottom 10% of the envelope after full-wave rectification and low-pass filtering at 10 Hz.
2) *Median Frequency:* Frequency at which the cumulative EMG PSD (10–450 Hz) reaches 50% of total power.
3) *Clip Fraction:* Fraction of samples at or above the 99.9th-percentile absolute amplitude (heuristic indicator of near-saturation).

A channel was flagged as poor quality if any of the following conditions held: envelope SNR $<$ 10 dB, clip fraction $>$ 0.02, median frequency $<$ 40 or $>$ 250 Hz on the 10–450-Hz PSD.

REFERENCES

[1] M. Capogrosso et al., "A brain–spine interface alleviating gait deficits after spinal cord injury in primates," *Nature*, vol. 539, no. 7628, pp. 284–288, 2016.

[2] F. Weber, S. Chung, K. T. Beier, M. Xu, L. Luo, and Y. Dan, "Control of REM sleep by ventral medulla GABAergic neurons," *Nature*, vol. 526, no. 7573, pp. 435–438, Oct. 2015.

[3] Y. Dong, J. Li, M. Zhou, Y. Du, and D. Liu, "Cortical regulation of two-stage rapid eye movement sleep," *Nature Neurosci.*, vol. 25, no. 12, pp. 1675–1682, Dec. 2022.

[4] O. Yizhar, L. E. Fenno, T. J. Davidson, M. Mogri, and K. Deisseroth, "Optogenetics in neural systems," *Neuron*, vol. 71, no. 1, pp. 9–34, Jul. 2011.

[5] E. S. Boyden, F. Zhang, E. Bamberg, G. Nagel, and K. Deisseroth, "Millisecond-timescale, genetically targeted optical control of neural activity," *Nature Neurosci.*, vol. 8, no. 9, pp. 1263–1268, Sep. 2005.

[6] S. Park et al., "One-step optogenetics with multifunctional flexible polymer fibers," *Nature Neurosci.*, vol. 20, no. 4, pp. 612–619, Apr. 2017.

[7] L. Li et al., "Transfer-printed, tandem microscale light-emitting diodes for full-color displays," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 18, May 2021, Art. no. 2023436118.

[8] X. Cai et al., "A dual-channel optogenetic stimulator selectively modulates distinct defensive behaviors," *iScience*, vol. 25, no. 1, Jan. 2022, Art. no. 103681.

[9] G. Gagnon-Turcotte, C. Ethier, Y. De Koninck, and B. Gosselin, "A 13μm CMOS SoC for simultaneous multichannel optogenetics and electrophysiological brain recording," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 466–468.

[10] S.-Y. Lee et al., "22.7 A programmable wireless EEG monitoring SoC with open/closed-loop optogenetic and electrical stimulation for epilepsy control," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 372–374.

[11] H.-V.-V. Ngo, T. Martinetz, J. Born, and M. Mölle, "Auditory closed-loop stimulation of the sleep slow oscillation enhances memory," *Neuron*, vol. 78, no. 3, pp. 545–553, May 2013.

[12] L. Xie et al., "Sleep drives metabolite clearance from the adult brain," *Science*, vol. 342, no. 6156, pp. 373–377, Oct. 2013.

[13] U. Shin et al., "A 256-channel 0.227μJ/class versatile brain activity classification and closed-loop neuromodulation SoC with 0.004 mm^2-1.51 μW/channel fast-settling highly multiplexed mixed-signal front-end," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 65, Feb. 2022, pp. 338–340.

[14] Y. Hou et al., "33.4 A multi-loop neuromodulation chipset network with frequency-interleaving front-end and explainable AI for memory studies in freely behaving monkeys," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2024, pp. 548–550.

[15] Y. Hou, Y. Zhu, X. Ji, A. G. Richardson, and X. Liu, "A wireless sensor-brain interface system for tracking and guiding animal behaviors through closed-loop neuromodulation in water Mazes," *IEEE J. Solid-State Circuits*, vol. 59, no. 4, pp. 1093–1109, Apr. 2024.

[16] M. ElAnsary et al., "Bidirectional peripheral nerve interface with 64 second-order opamp-less $\delta\sigma$ ADCs and fully integrated wireless power/data transmission," *IEEE J. Solid-State Circuits*, vol. 56, no. 11, pp. 3247–3262, Nov. 2021.

[17] J. Xu et al., "Fascicle-selective bidirectional peripheral nerve interface IC with 173dB FOM noise-shaping SAR ADCs and 1.38pJ/b frequency-multiplying current-ripple radio transmitter," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2023, pp. 31–33.

[18] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 4107–4115.

[19] H. Chen et al., "AdderNet: Do we really need multiplications in deep learning?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1468–1477.

[20] M. Elhoushi, Z. Chen, F. Shafiq, Y. H. Tian, and J. Y. Li, "DeepShift: Towards multiplication-less neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2359–2368.

[21] T. Luo, Z. Wang, D. Gerlinghoff, R. S. M. Goh, and W.-F. Wong, "Blunet: Arithmetic-free inference with bit-serialised table lookup operation for efficient deep neural networks," OpenReview, 2021. [Online]. Available: https://openreview.net/forum?id=_zL5mZ95FV6

[22] S. Dey, P. Dasgupta, and P. P. Chakrabarti, "DietCNN: Multiplication-free inference for quantized CNNs," 2023, *arXiv:2305.05274*.

[23] D. Shin, J. Lee, J. Lee, and H.-J. Yoo, "14.2 DNPU: An 8.1 TOPS/W reconfigurable CNN-RNN processor for general-purpose deep neural networks," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 240–241.

[24] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU: A 50.6 TOPS/W unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 218–220.

[25] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU: An energy-efficient deep neural network accelerator with fully variable weight bit precision," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 173–185, Jan. 2019.

[26] R. Guo et al., "15.4 a 5.99-to-691.1TOPS/W tensor-train in-memory-computing processor using bit-level-sparsity-based optimization and variable-precision quantization," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 242–244.

[27] S. Devuyst, "The dreams databases and assessment algorithm," Jan. 2005, doi: 10.5281/zenodo.2650142.

[28] S. A. Imtiaz, Z. Jiang, and E. Rodriguez-Villegas, "An ultralow power system on chip for automatic sleep staging," *IEEE J. Solid-State Circuits*, vol. 52, no. 3, pp. 822–833, Mar. 2017.

[29] S.-Y. Chang et al., "An ultra-low-power dual-mode automatic sleep staging processor using neural-network-based decision tree," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 9, pp. 3504–3516, Sep. 2019.

[30] C. Laney. (2021). *Animal Tracking: Object Tracking With Opencv in Open Field Behavioral Test*. [Online]. Available: https://github.com/colinlaney/animal-tracking

Chao Zhang received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2020 and 2025, respectively.

His research focuses on ultra-low power electroencephalogram (EEG)-based brain–machine interface (BMI) design.

Fenyan Zhang received the B.S. from Sichuan University, Chengdu, China, in 2018, and the M.S. degree in neurology from Chongqing Medical University, Chongqing, China, in 2023. She is currently pursuing the Ph.D. with the Chinese Institute for Brain Research, Beijing, China.

Her current research focuses on unraveling the neural mechanisms that regulate the circadian rhythms of sleep and wakefulness.

Dawid Sheng received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2023.

In 2021, he joined Xing Sheng Research Group, Beijing, and is involved in the development of advanced biphotonic devices.

Zhixiong Ma received the Ph.D. degree in biochemistry and molecular biology through a joint program between the National Institute of Biological Sciences (NIBS), Beijing, China, and Beijing Normal University, Beijing.

He subsequently pursued post-doctoral research focused on pharmacological investigations of autism and epilepsy as part of a collaborative initiative between Peking University Health Science Center, Beijing, and Purdue University, West Lafayette, IN, USA. Thereafter, he joined the Chinese Institute for Brain Research (CIBR), Beijing, where he investigated the mechanisms underlying sleep and biological rhythms. He later transitioned to the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, as an Assistant Researcher, employing electroencephalography (EEG) recordings, neurotransmitter fluorescent probes, and optogenetic techniques to elucidate the mechanisms of sleep and epilepsy.

Yongxiang Guo received the B.S. degree in electronic information science and technology from Beijing Normal University, Beijing, China, in 2021, and the M.S. degree in electronic engineering from Tsinghua University, Beijing, in 2024.

His research interests include various low-power circuit and system designs for biomedical applications.

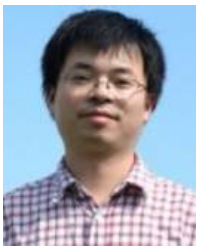
Chao Sun received the B.S. degree in electronic engineering and the M.S. degree in integrated circuit engineering from Tsinghua University, Beijing, China, in 2005 and 2008, respectively.

He is currently an SoC Engineer with Beijing Ningju Technology Ltd., Beijing, and focuses on analog and mixed-signal circuit designs oriented for various applications.

Yuwei Zhang received the M.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2018.

Since 2020, he has been with Beijing Ningju Technology Ltd., Beijing, as a PMU SoC Engineer, focusing on low-power integrated circuit design for biomedical applications.

Wenxin Zhao received the B.S. degree in physics from Beihang University, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree in electronic engineering with Tsinghua University, Beijing.



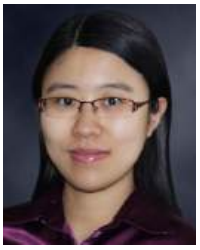
Xing Sheng (Senior Member, IEEE) received the bachelor's degree from Tsinghua University, Beijing, China, in 2007, and the Ph.D. degree from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2012.

He was a Post-Doctoral Researcher with the University of Illinois at Urbana–Champaign, Champaign, IL, USA, from 2012 to 2015. He is currently a Professor with the Department of Electronic Engineering and a PI with the IDG/McGovern Institute for Brain Research, Tsinghua University. His current research interests include the exploration of implantable micro- and nano-scale optical and electronic devices for neural signal sensing and modulation.

Tongfei A. Wang received the B.S. degree in biological sciences from Peking University, Beijing, China, in 2005, and the Ph.D. degree from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2012, where he was advised by Prof. Martha Gillette.

He completed post-doctoral training under Prof. Lily Jan at the University of California, San Francisco, San Francisco, CA, USA, from 2012 to 2020. He has been an Assistant Investigator with the Chinese Institute for Brain Research (CIBR), Beijing, since 2020. His research aims to decipher the neural circuitry regulating energy metabolism. Current projects in his laboratory focus on the functional architecture of hypothalamic networks in mice, specifically exploring how thermoregulation interacts with energy metabolism.

Dr. Wang is a member of the Chronobiology Branch of the Chinese Society for Cell Biology. He serves on the Cellular and Molecular Physiology Committee for the Chinese Association for Physiological Sciences.



Milin Zhang (Senior Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2004 and 2006, respectively, and the Ph.D. degree from the Electronic and Computer Engineering Department, The Hong Kong University of Science and Technology (HKUST), Hong Kong.

After finishing her doctoral studies, she worked as a Post-Doctoral Researcher at the University of Pennsylvania (UPenn), Philadelphia, PA, USA. She joined Tsinghua University in 2016. She is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. Her research interests include designing biomedical sensing-oriented SoC designs and various non-traditional imaging sensors.

Dr. Zhang serves and has served as a TPC Member for ISSCC, CICC, A-SSCC, and CASS. She has received the Best Paper Award of the BioCAS Track from the 2014 International Symposium on Circuits and Systems (ISCAS), the Best Paper Award (first place) from the 2015 Biomedical Circuits and Systems Conference (BioCAS), and the Best Student Paper Award (second place) from ISCAS 2017. She is the Chapter Chair of the SSCS Beijing Chapter.