Neuromorphic Wireless Split Computing With Multi-Level Spikes

DENGYU WU [1] **(Member, IEEE), JIECHEN CHEN** [2] **(Member, IEEE),**
BIPIN RAJENDRAN [1] **(Senior Member, IEEE), H. VINCENT POOR** [3] **(Life Fellow, IEEE),**
AND OSVALDO SIMEONE [2] **(Fellow, IEEE)**

[1] King's Laboratory for Intelligent Computing (KLIC), Centre for Intelligent Information Processing Systems (CIIPS), Department of Engineering, King's College London, WC2R 2LS London, U.K.
[2] King's Communications, Learning and Information Processing (KCLIP) Laboratory, CIIPS, Department of Engineering, King's College London, WC2R 2LS London, U.K.
[3] Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA

CORRESPONDING AUTHOR: J. CHEN (jiechen.chen@kcl.ac.uk)

ABSTRACT Inspired by biological processes, neuromorphic computing leverages spiking neural networks (SNNs) to perform inference tasks, offering significant efficiency gains for workloads involving sequential data. Recent advances in hardware and software have shown that embedding a small payload within each spike exchanged between spiking neurons can enhance inference accuracy without increasing energy consumption. To scale neuromorphic computing to larger workloads, split computing—where an SNN is partitioned across two devices—is a promising solution. In such architectures, the device hosting the initial layers must transmit information about the spikes generated by its output neurons to the second device. This establishes a trade-off between the benefits of multi-level spikes, which carry additional payload information, and the communication resources required for transmitting extra bits between devices. This paper presents the first comprehensive study of a neuromorphic wireless split computing architecture that employs multi-level SNNs. We propose digital and analog modulation schemes for an orthogonal frequency division multiplexing (OFDM) radio interface to enable efficient communication. Simulation and experimental results using software-defined radios reveal performance improvements achieved by multi-level SNN models and provide insights into the optimal payload size as a function of the connection quality between the transmitter and receiver.

INDEX TERMS Graded spikes, multi-level spikes, neuromorphic wireless communications, neuromorphic computing, spiking neural networks.

I. INTRODUCTION

A. CONTEXT AND MOTIVATION

CURRENT learning algorithms, computing primitives, and hardware platforms such as GPUs are widely expected to soon fall short in supporting scalable, energy-efficient artificial intelligence (AI) models, especially for edge deployments [1]. This motivates the ongoing exploration of alternative computing paradigms, including in-memory computing [2], neuromorphic computing [3], [4], and quantum computing [5], [6]. Advances in computing technologies are bound to affect a range of fields from the sciences [7] to engineering [8]. This work studies some of the implications of the emergence of neurormorphic computing for telecommunications engineering [9], [10], [11], [12], [13], [14], [15], [16].

As communication networks become increasingly softwarized [19], spiking neural networks (SNNs) present a promising option as co-processors for wireless transmitters and receivers, as explored in [13] and [15]. Neuromorphic computing, therefore, can play an important role in enabling advanced communication functionalities. Conversely, communication networks can support the development of
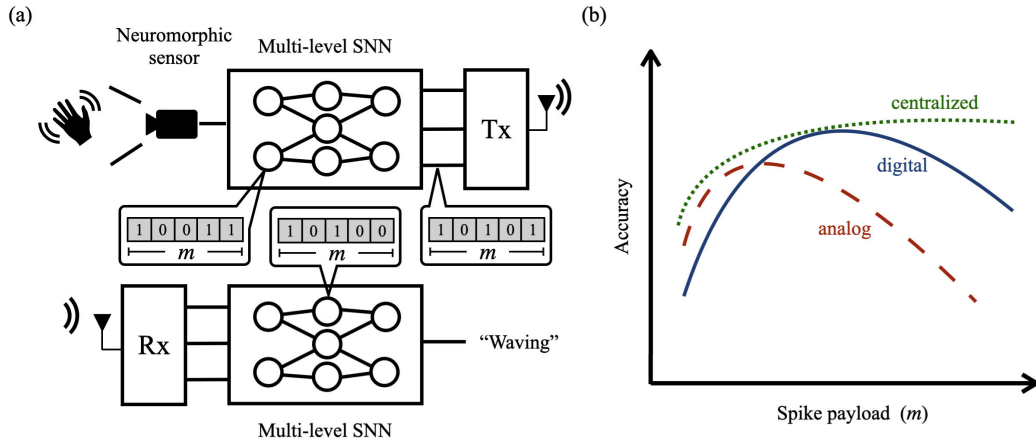
FIGURE 1. **(a) Neuromorphic wireless split computing architecture based on multi-level SNNs: Spikes exchanged between a transmitter and a receiver over a wireless channel include a payload of *m* bits. (b) While the accuracy of a centralized implementation increases monotonically with the spike payload *m* [17], [18], in the presence of communication constraints there is generally an optimized value of *m* that balances the informativeness of each spike with the reduced accuracy of higher-rate transmission.**

distributed computing architectures grounded in neuromorphic principles. In these architectures, communication protocols must be tailored to the unique nature of information exchanged between SNN neurons. Unlike conventional multi-bit clocked messages, spiking neurons encode and transmit information through the timing of individual spikes.

While SNNs can reduce the energy consumption for certain workloads [20], large-scale tasks requiring deeper SNN architectures may still prove too demanding in terms of energy and memory for mobile devices [21]. In these settings, split computing — where the computational workload is distributed across multiple devices — is promising solution. In particular, in [22], the authors demonstrated that splitting a deep SNN architecture across multiple edge devices reduced inference latency by 60.7% and the overall energy consumption per device by 27.7%. However, partitioning an SNN across multiple devices requires the devices to share timing information to maintain the integrity of the neural computations [9], [10], [12], [16], [23], [24], [25].

As shown in Fig. 1(a), in this paper, we focus on a basic distributed computing architecture [26] consisting of an SNN split between two devices, which are connected over a wireless channel. The transmitter-side SNN processes sequential data captured by a neuromorphic sensor, such as an event-driven camera [27], [28], [29], [30]. SNNs can natively process event-driven data via spiking neurons. The receiver-side SNN uses the received radio signal to produce a final inference decision. For example, in the set-up shown in Fig. 1, the transmitter's sensor observes hand gestures, which are estimated at the receiver side.

Conventional SNNs represent information solely in the timing of spikes. However, digital neuromorphic chips, such as Intel's Loihi 2 supports multi-level, or graded, spikes with minimal additional energy cost [17], [18] (see Sec. II-B for

further details). Multi-level spikes encode information both in the timing of the spikes and in their amplitude. There is evidence that biological brains may also leverage spike amplitude variability to encode additional information [31]. As illustrated in Fig. 1(a), multi-level spikes are assigned a payload of *m* bits, while conventional spike carry $m = 0$ additional bits of information. SNNs with multi-level spikes have been shown to improve the accuracy of conventional SNN deployments, particularly when the number of timesteps available for inference is limited [32], [33], [34].

In a split computing architecture, the introduction of multi-level spikes creates the challenge of transmitting a larger amount of information per spike on the wireless interface. As illustrated in Fig. 1(b), while in a centralized implementation, larger values of payload size *m* are generally beneficial in terms of inference accuracy, in a split computing system, an excessively large payload size can cause a performance degradation due to the lower fidelity of higher-rate transmissions on wireless channels. This work addresses this inherent tension by investigating the design of both analog and digital transmission schemes for neuromorphic wireless split computing systems with multi-level spikes.

B. RELATED WORK

1) NEUROMORPHIC WIRELESS SPLIT COMPUTING

Neuromorphic wireless split computing was first studied in [9], in which single-link neuromorphic sensing and computing were integrated with ultra-wideband (UWB) transmission to enable edge-based remote inference. This work was then extended in [10] to a multi-device scenario with frequency-selective channels, demonstrating IR transmission's compatibility in multi-device environments. In [12], wake-up radios were incorporated into the system to further reduce the overall energy consumption of the system.

The work reported in [24] analyzed how spike losses affect the inference accuracy and total neural activity when considering a distributed wireless SNN implementation. Another reference [14] studied a distributed system of edge nodes, each containing a subset of spiking neurons, that communicate with an access point via wireless channels using frequency division multiple access (FDMA) by allocating different frequency bands to different nodes.

A neuromorphic integrated sensing and communications system was studied in [11], in which an SNN was deployed at the receiver to decode the transmitted information and detect the possible presence of a target simultaneously.

There have been also several reported prototypes for neuromorphic split computing. The transmission model in [35] utilized neuromorphic principles, implemented on Intel's Loihi chip combined with software-defined radio (SDR) hardware, to build a full-stack neuromorphic wireless communication system that considers both orthogonal frequency division multiplexing (OFDM) and UWB transmission. Another work [36] experimentally demonstrated a communication approach for large-scale wireless asynchronous microsensor networks, enabling the transmission of binary events from thousands of local nodes with high spectral efficiency and low error rates.

2) MULTI-LEVEL SNNs

A few studies have contributed to advances in multi-level SNNs in centralized implementations. For example, reference [34] proposed a multi-bit transmission mechanism that expands spike representation from a single bit to multiple bits, enriching the information content per spike. In [37], a ternary spiking neuron was introduced to increase information capacity while retaining event-driven, addition-only processing advantages. Additionally, reference [38] designed a spiking neuron that activates integer values during training and maintains spike-driven behavior by extending virtual time-steps during inference for object detection tasks.

C. MAIN CONTRIBUTIONS

This paper investigates for the first time neuromorphic wireless split computing with multi-level SNNs. Unlike conventional neural networks, SNNs are sequential models, processing and transmitting information over time. Furthermore, they use timing information for inter-neuron communication, producing temporally sparse signals. This is fundamentally different from the dense and continuous representations of inter-neuron signals in artificial neural networks (ANNs). For these reasons, SNNs require separate designs and evaluations as compared to conventional neural networks [10], [12].

Previous works [9], [10], [11], [12], which focused on conventional SNNs, adopted a UWB interface due to its low power consumption and compatibility with spike-based transmission. In contrast, in this paper we adopt the standard OFDM interface, which provides a more flexible modulation scheme to accommodate multi-level spikes and is more

widely applicable and available. In particular, OFDM facilitates prototyping using conventional SDR platforms, such as the Universal Software Radio Peripheral (USRP) [39]. We design and evaluate both digital and analog modulation schemes, which are tested via simulation and via an experimental platform.

Overall, the main contributions of this paper are summarized as follows.

- We study for the first time a neuromorphic wireless split computing architecture based on multi-level SNNs. Unlike conventional SNNs with binary spikes, multi-level SNNs are able to process richer information by assigning a multi-bit payload to each spike.
- We detail digital and analog transmission schemes that leverage the sparsity of inter-neuron signals in SNNs, and adapt to the dynamic output of spikes produced over time. The proposed digital modulation scheme is based on the address-event representation (AER) of multi-level spikes [17], [18]. In this implementation, spike addresses and payloads are channel-encoded and modulated on OFDM symbols. If the number of information bits exceeds the available capacity – which is more likely to occur for a larger value of the payload size m – spikes are dropped, causing a potential decrease in accuracy. Upon channel decoding, the transmitted spikes are reconstructed at the receiver and fed to the receiver-side SNN to produce the final inference decision.
- We also detail an analog implementation whereby each output neuron of the transmitter-side SNN is assigned to a fixed subset of OFDM subcarriers and the spikes payloads are transmitted via pulse-amplitude modulation (PAM) on all the assigned subcarriers. This way, the addresses are implicitly transmitted via the location of the PAM symbols across the subcarrier indices. While no spikes are dropped as long as the number of subcarriers is large enough, analog transmission may degrade the quality of the reconstructed spikes due to the reliance of repetition coding.
- We evaluate the performance of the proposed neuromorphic wireless split computing architecture based on multi-level SNNs both via simulations and via a basic prototype using a neuromorphic camera [40] and USRP boards.

D. ORGANIZATION

The remainder of the paper is organized as follows. Section II presents background information about multi-level SNN. Section III describes the neuromorphic wireless split computing system with multi-level spikes under study, while the proposed digital and analog transmission schemes are described in Section IV. Section V explains neuromorphic receiver processing, including channel estimation, equalization, and decoding SNN processing. Experimental setting and results are described in Section VI. Finally, Section VII concludes the paper.
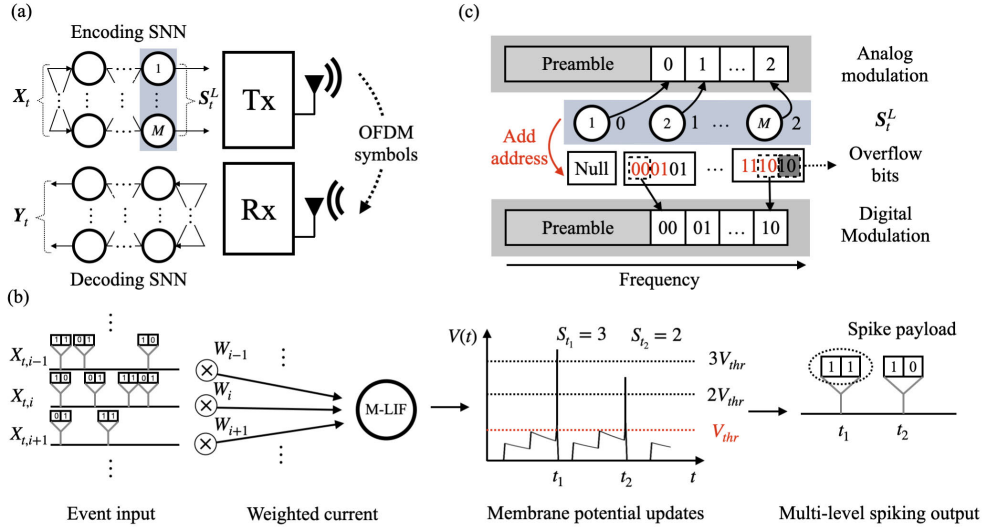
FIGURE 2. Neuromorphic wireless split computing with multi-level spikes: (a) An SNN is split into an encoding SNN and a decoding SNN, which are connected over a wireless channel following a spilt computing architecture. (b) Unlike prior works [9], [10], [11], [12], the SNNs implement spiking neurons that communicate using multi-level spikes [17], [18], adopting a multi-level leaky integrate-and-fire (M-LIF) neuron model. (c) The output of the encoding SNN is transmitted using either analog or digital modulation. In the analog implementation, each output neuron of the encoding SNN is assigned separate OFDM subcarriers. In contrast, in the digital implementation, the AER protocol is used to embed information about the neurons' identities. Overflow bits that do not fit the allocated OFDM symbols are discarded.

II. MULTI-LEVEL SPIKING NEURAL NETWORKS

As illustrated in Fig. 2, this paper studies a neuromorphic wireless split computing system that leverages spiking neuronal models with multi-level, or graded, spike. In this section, we describe first the conventional SNN model based on leaky integrate-and-fire (LIF) neurons (see, e.g., [41], [42]), and then cover the generalized SNN model with multi-level LIF (M-LIF) neurons [17], [18]. The following section will present the proposed split computing architecture based on SNNs with M-LIF neurons.

A. CONVENTIONAL LEAKY INTEGRATE-AND-FIRE NEURON

A conventional LIF neuron accumulates stimuli over time, using an internal state known as membrane potential, and emits a spike once its membrane potential exceeds a certain threshold. LIF neurons can be arranged into arbitrary architectures, and they have been used to implement models such as multi-layer perception transformers [43] and state-space models [44]. In this work, we consider an arbitrary layered architecture, in which each neuron i in layer l generates a spike at time t if the local membrane potential $V_{t,i}^l$ passes the threshold V_{thr}^l. Accordingly, the output of the LIF neuron i in layer l at time t is given by

$$S_{t,i}^l = \begin{cases} 0 \text{ (no spike), if } V_{t,i}^l \leq V_{thr}^l, \\ 1 \text{ (spike),} \quad \text{if } V_{t,i}^l > V_{thr}^l. \end{cases} \quad (1)$$

The membrane potential $V_{t,i}^l$ is updated via the leaky integrator dynamics

$$V_{t,i}^l = \delta V_{t-1,i}^l(1 - S_{t-1,i}^l) + Z_{t,i}^l, \quad (2)$$

where $0 < \delta < 1$ represents the decay factor, and $Z_{t,i}^l$ denotes the input current from the pre-synaptic neurons in the previous, $(l-1)$-th, layer. By (2), if the neuron i at layer l spikes at time $t-1$, i.e., if $S_{t-1,i}^l = 1$, the membrane potential is reset at time t.

Given the vector S_t^{l-1} collecting all the binary outputs $\{S_{t,i}^{l-1}\}_i$ produced in the $(l-1)$-th layer, the input current vector Z_t^l, collecting the currents $\{Z_{t,i}^l\}_i$ feeding into each neuron i in layer l, is given by the linear combination

$$Z_t^l = W^l S_t^{l-1}, \quad (3)$$

where W^l is weight matrix between the $(l-1)$-th and the l-th layer.

Evaluating (3) requires one accumulate operation per spike, whose energy cost we denote as E_{ac}.

B. LEAKY INTEGRATE-AND-FIRE NEURON WITH MULTI-LEVEL SPIKES

Spikes emitted in a conventional LIF-based SNN carry information only via their firing time t defined as in (1). A more general model, implemented in neuromorphic chips, such as Intel's Loihi [17], allows each spike to carry m additional bits of information. According to Fig. 2, a multi-level, or graded, spike is assigned a payload containing m bits of information.

In this way, when $m = 0$, a graded spike reduces to the spike produced by conventional LIF models.

As reported in [17] and [18], digital neuromophic chips such as Intel's Loihi 2 permit spikes to carry integer-valued payloads with marginal additional energy cost. This is due to several architectural optimizations:

- *Efficient spike processing:* Digital chips commonly use the AER protocol for communications between cores [17]. Each AER packet contains address bits, and the few additional bits required to transmit the spike payload – typically less than 8 bits – yield a marginal increase the overall communication cost.
- *Optimized synapse processing:* As discussed in [45], incoming spikes can be mapped to lists of synapse weights that are accumulated for consumption in the next time step. This approach allows for the efficient processing of multi-level spikes without significantly increasing computational complexity.
- *Peripheral modifications:* The overhead to generate multi-bit spikes after multiply-accumulate (MAC) operations primarily involves modifying the sense amplifier and ADC [46]. This overhead scales linearly, contrasting with the quadratic scaling of the number of MAC operations being performed.

Like LIF neurons, multi-level LIF (M-LIF) neurons produce a spike any time the membrane potential crosses a threshold. However, the payload of the spike produced by an M-LIF neuron contains a payload of m bits, which is obtained by quantizing the membrane potential at the time of spiking.

To elaborate, consider a neuron i in layer l of an arbitrary layered architecture, whose membrane potential is denoted by $V_{t,i}^l$. The output $S_{t,i}^l$ associated with neuron i in layer l at time t is given by [33]

$$S_{t,i}^l = \begin{cases} 0 \text{ (no spike)}, & \text{if } V_{t,i}^l \leq V_{\text{thr}}^l, \\ Q^l(V_{t,i}^l), & \text{if } V_{t,i}^l > V_{\text{thr}}^l, \end{cases} \quad (4)$$

where $Q(\cdot)$ is an m-bit quantizer. By (4), a spike contains m bits given by the quantization level $Q^l(V_{t,i}^l)$. We specifically adopt the clipped uniform quantizer

$$Q^l(V_{t,i}^l) = \min(\lfloor \alpha^l V_{t,i}^l 2^m \rfloor, 2^m) \in \{1, \cdots, 2^m\}, \quad (5)$$

where $\lfloor \cdot \rfloor$ is the floor operation, and $\alpha^l \in (0, 1)$ is a per-layer trainable scaling factor. The integer (5), which ranges in the interval $\{1, \ldots, 2^m\}$, is the payload of a spike, which corresponds to m bits.

Evaluating the input currents (3) requires the evaluation of a multiply-and-accumulate operation per spike, with the multiplication involving an m-bit number. The energy consumption for this operation can be modeled as

$$E_{mac}(m) = (1 + \gamma(m - 1))E_{ac}, \quad (6)$$

where $\gamma \in [0, 1]$ is a technology-dependent parameter. When $\gamma = 0$, the hardware optimizations mentioned above are maximally efficient, making m-bit synaptic operations as costly as with $m = 1$, i.e., $E_{mac} = E_{ac}$. In contrast, when $\gamma =$
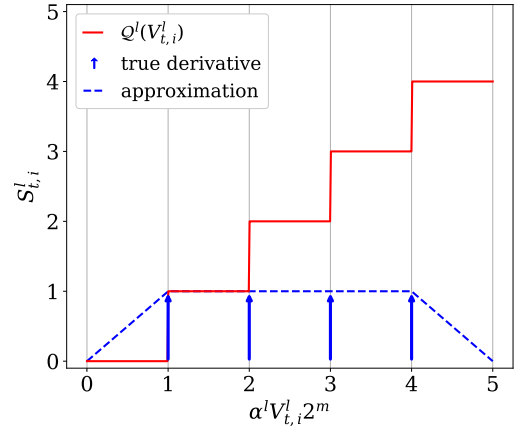

FIGURE 3. Surrogate derivative (8) used for training SNN models with M-LIF neurons ($m = 2$ and $\Gamma = 1$).

1, the energy overhead reflects a basic shift-and-accumulate implementation, which entails an energy cost that is m times that of a single accumulate operation, e.g., $E_{mac}(m) = mE_{ac}$.

C. TRAINING MULTI-LEVEL SNN MODELS

In this work, we assume the availability of a pre-trained multi-level SNN model, which is split between encoder and decoder as discussed in the next section. Training of the multi-level SNN is achieved using backpropagation via a surrogate gradient that provides a smooth approximation for the hard quantization function in (5) [47], [48], [49].

To address the non-differentiability of (4), we first recall the surrogate gradient method used in [47], which applies to a conventional LIF model (1). This method approximates the derivative of the neuron's output $S_{i,t}^l$ in (1) with respect to membrane potential $V_{t,i}^l$ as

$$\frac{\partial S_{i,t}^l}{\partial V_{t,i}^l} \approx \Gamma \max\left(0, 1 - \left|V_{t,i}^l - V_{\text{thr}}^l\right|\right), \quad (7)$$

where $\Gamma > 0$ is a hyper-parameter. The equation (7) replaces the true derivative of the output (1) — a Dirac delta function at threshold V_{thr}^l — with a triangular function centered at the threshold V_{thr}^l with height Γ.

To extend this approach to the M-LIF model (4), we first set for simplicity the threshold as $1/\alpha^l 2^m$. Then, the derivative of the neuron's output is approximated as shown in Fig. 3 as

$$\frac{\partial S_{i,t}^l}{\partial V_{t,i}^l} \approx \begin{cases} \Gamma \max\left(0, 1 - \left|\alpha^l V_{t,i}^l 2^m - 1\right|\right), & \text{if } V_{t,i}^l < \frac{1}{\alpha^l 2^m}, \\ 1, & \text{if } \frac{1}{\alpha^l 2^m} \leq V_{t,i}^l \leq \frac{1}{\alpha^l}, \\ \Gamma \max\left(0, 1 - \left|\alpha^l V_{t,i}^l 2^m - 2^m\right|\right), \\ \qquad \text{if } V_{t,i}^l > \frac{1}{\alpha^l} - \frac{1}{\alpha^l 2^m}. \end{cases} \quad (8)$$

As illustrated in Fig. 3, the discontinuous, impulsive, derivatives, associated with the multi-level spikes (4) $S_{t,i}^l$ are approximated in a manner that extends (7) via a piece-wise function [48], [49].

III. NEUROMORPHIC WIRELESS SPLIT COMPUTING WITH MULTI-LEVEL SPIKES

In this section, we describe the neuromorphic split computing system under study. In order to accommodate a multi-bit spike payload, as well as to facilitate prototyping using SDR technology, we adopt an OFDM radio interface in lieu of the UWB modulation considered in prior works [9], [10], [11], [12].

A. NEUROMORPHIC WIRELESS SPLIT COMPUTING

As illustrated in Fig. 2(a), we consider a neuromorphic wireless split computing system in which an SNN consisting of M-LIF neurons is split between a single-antenna transmitter (Tx) and a single-antenna receiver (Rx). The SNN is trained to solve an inference problem based on data captured by a neuromorphic sensor, such as a dynamic vision sensor (DVS) camera [40] at the Tx. Based on the SNN split, the Rx makes the final inference decision by using the wireless signals received from the Tx. Practical examples of applications of this architecture were presented in [36] and [50], including brain-computer interfaces and biomedical devices.

Neuromorphic sensors generate a spike whenever a relevant event is detected, such as a significant change in pixel brightness. Spikes produced by the sensor are often graded, e.g., indicating the sign of the reported change with a one-bit payload [40].

As illustrated in Fig. 4, we adopt a discrete-time model, where time is divided into sensing slots indexed by integers $t = 1, 2, \ldots$ Each slot corresponds to the time period over which the neuromorphic sensor accumulates information before reporting the presence or absence of events, along with the corresponding payloads.

Accordingly, at the end of each sensing period t, the neuromorphic sensor at the Tx produces a $D \times 1$ vector $X_t = [X_{t,1}, \ldots, X_{t,D}]^T$ representing multi-level spikes produced by each of the D elements of the sensor. For example, a DVS camera produces D signals, each corresponding to one pixel. Each entry $X_{t,i}$ represents the presence ($X_{t,i} > 0$) or absence ($X_{t,i} = 0$) of a spike. When a spike is present, the payload $X_{t,i} \in \{1, \ldots, 2^m\}$ encompasses m bits.

The spiking signal X_t recorded by the neuromorphic sensor is processed by the encoding SNN. Unlike prior works [9], [10], [11], [12], the encoding and decoding SNNs in the proposed system are capable of processing and producing multi-level spikes with m bits. Specifically, as introduced in the previous section, we consider arbitrary layered architectures consisting of M-LIF neurons for both the encoding and decoding SNNs.

Denote by M the number of neurons in the last layer of the encoding SNN, and by $S_t = [S_{t,1}, S_{t,2}, \ldots, S_{t,M}]$ the vector of graded spikes produced by the output layer of the encoding
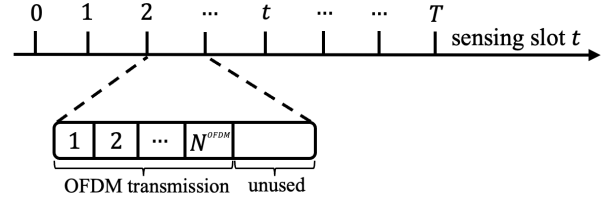

FIGURE 4. Timeline of the proposed neuromorphic wireless split computing system. Time is discretized into sensing slots $t = 1, 2, \ldots, T$, corresponding to the time period over which the neuromorphic sensor accumulates information before reporting the presence or absence of events, along with the corresponding payloads. The spikes produced at time slot $t - 1$ are processed by the Tx, and the outputs of the encoding SNNs are transmitted over the air using N^{OFDM} OFDM symbols to the Rx during the following, t-th, sensing slot. The decoding SNN at the Rx then processes the received signals to produce an inference decision. Each sensing time step t is typically much longer than the duration of an OFDM symbol.

SNN, with $S_{t,i} \in \{0, 1, \ldots, 2^m\}$ for all $i \in \{1, \ldots, M\}$. The multi-level spikes S_t generated by the encoding SNN are modulated by the transmitter onto a baseband signal that is transmitted over a wireless channel using OFDM. Using the received signals, the decoding SNN at the Rx produces the final inference decision.

B. OFDM TRANSMISSION OF MULTI-LEVEL SPIKES

As illustrated in Fig. 4, a fixed number N^{OFDM} of OFDM symbols is available in each sensing slot t to transmit information about the previous slot $t - 1$. The duration of each sensing slot is typically sufficient to accommodate a large number of OFDM symbols, and we assume that the wireless interface is used for other devices and/or services when not occupied by the Tx. For instance, a DVS camera has a typical sensing period of 130 ms [51], while an OFDM symbol for a 5G wireless link with a bandwidth 20 MHz takes 38.09 μs.

Each OFDM symbol consists of $N^D + N^P$ subcarriers, with N^D data subcarriers and N^P pilot subcarriers. We denote the subset of data subcarriers as $\mathcal{N}^D$ and the subset of pilot subcarriers as $\mathcal{N}^P$, with $|\mathcal{N}^D| = N^D$ and $|\mathcal{N}^P| = N^P$.

Assuming that the cyclic prefix (CP) of each OFDM symbol is no shorter than the discrete delay spread of the multi-path channel, the n-th received OFDM symbol at sensing time period t is given by [52]

$$y_t^n = H_t^n x_t^n + w_t^n, \tag{9}$$

where the diagonal channel matrix H_t^n collects on its main diagonal the channel frequency responses across all the subcarriers; x_t^n is the n-th frequency-domain OFDM symbol encompassing both the pilot signals and the payload of the multi-level spikes for slot $t - 1$; and w_t^n is a noise vector, with independent and identically distributed (i.i.d.) complex Gaussian elements having zero means and variances N_0.

IV. DIGITAL AND ANALOG TRANSMISSION OF MULTI-LEVEL SPIKES

In this section, we present digital and analog modulation strategies for encoding multi-level spikes at each sensing time slot. We begin by introducing pilot transmission, which enables channel estimation for OFDM symbol equalization. Next, we discuss two different types of power constraints, and, finally, we detail the digital and analog modulation schemes.

A. PILOT TRANSMISSION

To enable the receiver to perform effective channel estimation and equalization for both analog and digital modulation, the Tx sends pilot symbols $\{x_{t,i}^p\}_{i \in \mathcal{N}^P}$, known to the Rx, at a given power level P^{max} on designated pilot subcarriers indexed by the integers $i \in \mathcal{N}^P$. As discussed in the next section, the pilot subcarriers serve as references to obtain a channel estimate, which is leveraged for the equalization of data symbols.

B. POWER CONSTRAINTS

At each sensing slot t, the graded spikes S_{t-1} are modulated into symbols $\{x_{t,i}^n\}_{i \in \mathcal{N}^D}$ transmitted on the subset $\mathcal{N}^D$ of data subcarriers in each OFDM symbol indexed as $n = 1, 2, \ldots, N^{OFDM}$. We consider two types of power constraints on the data subcarriers.

1) AVERAGE PER-SYMBOL POWER CONSTRAINT

The average per-symbol power constraint limits the overall power used for transmission across all data subcarriers in a given OFDM symbol. Denoting the collection of N^D symbols transmitted on all data subcarriers of the n-th OFDM symbol at sensing slot t by $\boldsymbol{x}_t^n$, this constraint is given by

$$\frac{1}{N^D} \|\boldsymbol{x}_t^n\|^2 \leq P^{max}, \tag{10}$$

where P^{max} represents the maximum allowable average transmission power per data subcarrier. This constraint ensures that the total transmitted power remains within acceptable limits, helping to control energy consumption and interference.

2) PEAK POWER CONSTRAINT

While the average per-symbol power constraint (10) allows for adaptive power allocation among data subcarriers, in practice, the Tx has a maximum power limit that cannot be exceeded for each subcarrier. To address this type of constraint, we also study a peak power requirement, which restricts the maximum power transmitted on each data subcarrier to a fixed value P^{max}, i.e.,

$$\|x_{t,i}^n\|^2 \leq P^{max} \tag{11}$$

for each subcarrier i. This peak power constraint helps maintain the peak-to-average power ratio (PAPR) within reasonable bounds, thereby mitigating potential issues like amplifier non-linearities, which can adversely affect signal quality and system performance [52].

3) DYNAMIC POWER CONSTRAINTS

Under either constraint (10) or (11), the total maximum power allocated at each slot t remains the same, i.e.,

$$\frac{1}{N^D N^{OFDM}} \sum_{n=1}^{N^{OFDM}} \|\boldsymbol{x}_t^n\|^2 \leq P^{max}. \tag{12}$$

Accounting for the sequential nature of processing in SNNs, we also explore dynamic power allocation strategy across different sensing slots. Specifically, denoting as P_t the power allocated in slot t, we impose the total power budget constraint

$$\frac{1}{T} \sum_{t=1}^{T} P_t \leq P^{max}, \tag{13}$$

across T time slots. This enables the dynamic allocation of transmit power levels P_t over the time index t. Within each slot t, the power constraints (10) or (11) is enforced with the power P_t in lieu of P^{max}.

C. DIGITAL MODULATION OF MULTI-LEVEL SPIKES

For digital modulation, we adopt a standard AER protocol to encode the multi-level spikes [53]. Accordingly, each graded spike $S_{t,i} > 0$ at time t is associated with an AER packet containing the address of the i-th spiking output neuron, as well as with the payload of m bits. The address of the i-th output neuron is encoded into $\lceil \log_2(M) \rceil$ bits, as there are M possible spiking neurons at the output layer. Hence, The resulting AER packet contains $\lceil \log_2(M) \rceil + m$ bits.

Following the timeline in Fig. 4, the AER packets corresponding to all spikes generated at each sensing slot $t - 1$ from the output layer of the encoding SNN are encoded and modulated using N^{OFDM} OFDM symbols at sensing slot t. The N^{OFDM} symbols must thus ideally encode a number of AER packets equal to $\sum_{i=1}^{M} \mathbb{1}(S_{t-1,i} > 0)$. This yields a total of

$$B_t^{ToT} = (\lceil \log_2(M) \rceil + m) \sum_{i=1}^{M} \mathbb{1}(S_{t-1,i} > 0) \tag{14}$$

bits to be transmitted. However, due to the varying level of sparsity of the output neurons at different times $t - 1$, the allocated OFDM symbols may be insufficient to deliver all these bits. In particular, using a binary channel code with rate $0 < r \leq 1$ and a modulation scheme with 2^B constellation points, the number of bits that can be transmitted by the Tx to the Rx is $B^{OFDM} = N^{OFDM} N^D Br$.

If the spiking signals of the output neurons are sparse enough, so that the number of bits B_t^{ToT} does not exceed the capacity B^{OFDM} of the available OFDM symbols, all the AER packets are encoded for transmission. Otherwise, the largest subset $\mathcal{S}_t$ of AER packets is selected so as to guarantee the condition $(\lceil \log_2(M) \rceil + m)|\mathcal{S}_t| \leq B^{OFDM}$. The subset $\mathcal{S}_t$ is selected uniformly at random among the $\sum_{i=1}^{M} \mathbb{1}(S_{t-1,i} > 0)$ AER packets. Note that, in case the output signals are very sparse, i.e., if $B_t^{ToT} < B^{OFDM}$, some subcarriers remain unused.

D. ANALOG MODULATION OF MULTI-LEVEL SPIKES

As discussed in the previous subsection, digital modulation represents each multi-level spike using an AER format, requiring additional bits to specify addresses, as well as redundant bits for channel coding. In contrast, as detailed next, analog modulation directly maps the multi-level spikes onto a set of PAM symbols. These PAM symbols are then loaded onto the OFDM symbols by mapping subcarriers to output neurons of the encoding SNN, without the need for addressing or coding overhead.

With analog modulation, each of the M output neurons of the encoding SNN is assigned to a subset of data subcarriers in the OFDM symbols corresponding to the current time slot t. We denote the mapping between neuron i and a subset of subcarriers $\mathcal{N}^D(i) \subseteq \mathcal{N}^D$. The subsets $\mathcal{N}^D(i)$ are disjoint, so that each subcarrier is uniquely assigned to one neuron. This requires the assumption $N^D N^{OFDM} \geq M$. Moreover, each set $\mathcal{N}^D(i)$ contains $\lceil N^D N^{OFDM}/M \rceil$ subcarriers.

Each multi-level spike $S_{t,i} > 0$ from output neuron i at sensing slot t is mapped into a 2^m-PAM symbol, which is transmitted on all subcarriers in subset $\mathcal{N}^D(i)$. The transmission of the same symbol on multiple subcarriers amounts to a form of repetition coding. In contrast, if there is no spike, i.e., if $S_{t,i} = 0$, the corresponding subcarriers in subset $\mathcal{N}^D(i)$ are idle. Therefore, sparser spiking signals entail a larger number of unused subcarriers.

V. NEUROMORPHIC RECEIVER PROCESSING

In this section, we describe the processing applied by the receiver for both digital and analog modulation schemes. The receiver first estimates the channel using standard signal processing methods, allowing for equalization of the received symbols. Once equalization is complete, the receiver demodulates the data symbols to estimated multi-level spikes, which are then passed to the decoding SNN to make the final inference decision.

A. CHANNEL ESTIMATION

Denote by $\boldsymbol{h}_t^n = [H_{t,1}^n, \ldots, H_{t,N+N^p}^n]^T$ the diagonal elements of the frequency-domain channel matrix $\boldsymbol{H}_t^n$ in (9). Note that $H_{t,i}^n$ corresponds to the channel gain for the i-th subcarrier in the n-th OFDM symbol at time t. Following the standard approach [52], the receiver first estimates the channel vector $\hat{\boldsymbol{h}}_t^n$ on the pilot subcarriers. Based on these estimates, the channel gains for the data subcarriers are determined via interpolation [54].

Specifically, the receiver extracts from the received signal $\boldsymbol{y}_t^n$ in (9) the received pilot symbols $\boldsymbol{y}_{t,n}^p = \{y_{t,i}^n\}_{i \in \mathcal{N}^P}$. The channels on the pilot subcarriers are estimated using the least squares (LS) method. The resulting estimated channel $\hat{h}_{t,i}^n$ for pilot subcarrier $i \in \mathcal{N}^P$ of the n-th OFDM symbol is given by

$$\hat{h}_{t,i}^n = \frac{y_{t,i}^n}{x_{t,i}^p}, \tag{15}$$

where $x_{t,i}^p$ is the known pilot symbol transmitted at time t on pilot subcarrier $i \in \mathcal{N}^P$.

With the estimated channels $\{\hat{h}_{t,i}^n\}_{i \in \mathcal{N}^P}$ on the pilot subcarriers for each n-th OFDM symbol, the channel response on the data subcarriers is estimated via linear interpolation. To elaborate, write the set of pilot subcarriers $\mathcal{N}^P = \{i_1, i_2, \ldots, i_P\} \subset \{1, 2, \ldots, N + N^p\}$ and assume pilots are interleaved with data subcarriers. For a data subcarrier $i \in \mathcal{N}^D$, we find the largest pilot subcarrier $i_p < i$ and the smallest pilot subcarrier $i_{p+1} > i$, where $i_p < i < i_{p+1}$. The estimated channel response for the data subcarrier $i \in \mathcal{N}^D$ at time t is then given by

$$\hat{h}_{t,i}^n = \hat{h}_{t,i_p}^n + \frac{\hat{h}_{t,i_{p+1}}^n - \hat{h}_{t,i_p}^n}{i_{p+1} - i_p}(i - i_p). \tag{16}$$

B. EQUALIZATION

After obtaining the channel estimate, zero forcing (ZF) equalization is applied to mitigate the effect of the channel on the received symbols. Specifically, the equalized data symbol for the i-th data subcarrier in the n-th OFDM symbol at sensing slot t is given by the ratio

$$\hat{x}_{t,i}^n = \frac{y_{t,i}^n}{\hat{h}_{t,i}^n}. \tag{17}$$

For both digital and analog modulation, each equalized symbol $\hat{x}_{t,i}^n$ is demodulated into one of the 2^B constellation points using maximum likelihood detection. This detection involves selecting the constellation point that is closest to the equalized symbol in terms of Euclidean distance, ensuring that the received symbol is mapped to its most likely transmitted value.

1) DIGITAL IMPLEMENTATION

For the digital implementation, each equalized symbol $\hat{x}_{t,i}^n$ is demodulated into one of the 2^B constellation points, resulting in a B-bit sequence. The bits from all data subcarriers over N^{OFDM} OFDM symbols form a total of $N^{OFDM} N^D B$ bits. These bits are then processed through channel decoding to reconstruct the estimated AER packets.

Each estimated AER packet contains a payload of m bits and an address i, which are mapped to the estimated multi-level spike $\hat{S}_{t,i}$. If no multi-level spike is detected for a neuron i, the neuron's input is set to $\hat{S}_{t,i} = 0$, indicating no activity. This results in a vector $\hat{\boldsymbol{S}}_t = [\hat{S}_{t,1}, \hat{S}_{t,2}, \ldots, \hat{S}_{t,M}]^T$ of estimated spikes, which serves as the input to the decoding SNN.

2) ANALOG IMPLEMENTATION

In the analog implementation, the $\lceil N^D N^{OFDM}/M \rceil$ equalized subcarriers corresponding to the same symbol are averaged to enhance signal quality. After averaging, the symbols are demodulated by using maximum likelihood detection to determine the most likely value from a 2^B-level PAM constellation. The demodulated PAM symbols are then mapped

to the estimated multi-level spikes, resulting in the vector $\hat{S}_t = [\hat{S}_{t,1}, \hat{S}_{t,2}, \ldots, \hat{S}_{t,M}]^T$ of estimated spike activities of the M input neurons for the decoding SNN.

C. DECODING SNN PROCESSING

The estimated spikes $\hat{S}_t = [\hat{S}_{t,1}, \hat{S}_{t,2}, \ldots, \hat{S}_{t,M}]^T$, derived from either analog or digital modulations, are processed by the decoding SNN. For a classification task, the decoding SNN has C output neurons, with each output neuron representing a specific class.

Focusing on classification, this work uses a membrane potential-based decision rule for classification [55]. Following this approach, the classification decision is determined by identifying the output neuron that has the highest integrated membrane potential over all sensing slots $t = 1, \ldots, T$. The decision rule can be expressed mathematically as

$$\hat{c} = arg\,max_{i \in \{1,\ldots,C\}} \sum_{t=1}^{T} V_{t,i}, \tag{18}$$

where $V_{t,i}$ represents the membrane potential of the i-th output neuron at sensing slot t.

VI. SIMULATION RESULTS AND EXPERIMENTS

In this section, we report results from simulations and real-world experiments with the main aim of investigating the potential advantages of multi-level spikes in neuromorphic wireless split computing.[1]

A. SETTING

1) INFERENCE TASK

The system is configured to classify event-based inputs, captured by a neuromorphic vision sensor. Specifically, we consider the standard DVS128 Getsure dataset [51], which consists of data logged by a DVS camera for a duration of 6 seconds when presented one out of 11 possible gestures [56]. An SNN with M-LIF neurons having five convolutional layers followed by four fully connected layers is pre-trained by using the approach discussed in Sec. II-C. We split the network, configured as 2C4-64C3-128C3-128C3-128C3-128C3-AP2-FC512-FC256-FC128-FC11, at the first fully connected (FC) layer. In this notation, C denotes a convolutional layer, and AP is the average-pooling layer, with the numbers before and after each layer description specifying the number of input channels and kernel size, respectively. The first six layers form the encoding SNN, while the remaining layers constitute the decoding SNN. Accordingly, the encoding SNN's output layer contains $M = 512$ neurons. We train different SNN models for different pairs (T, m) consisting of number $T \in \{2, 4, 6, 8, 10\}$ of sensing slots and payload size $m \in \{0, 2, 4, 6, 8\}$.

To define the sensing slots, each original recording from the DVS camera is divided into four segments of 1.3 seconds each. Each segment is further divided into 10 frames, each of

---

[1]Code is available at https://github.com/kclip/neurocomm-msnn

duration 130 ms, with each frame representing a single sensing slot. The events within each sensing slot are accumulated and directly fed to the first layer of the SNN. For each sensing slot, each neuron in the hidden layer can emit at most one multi-level spike.

2) SIMULATION SETTING

In the simulation results, the number of OFDM symbols per sensing slot is set to $N^{OFDM} = 5$, and each OFDM symbol consists of $N^D = 512$ data subcarriers and $N^P = 75$ pilot subcarriers. Pilot symbols are interleaved with data subcarriers, so that a pilot is placed every 8 data symbols. We consider a five-path frequency-selective channel, where each path amplitude follows a Rayleigh distribution with the same average power so that the average channel norm equals 1. The signal-to-noise ratio (SNR) is defined as the ratio of the peak or average per-subcarrier symbol power P^{max} over the noise power, i.e., SNR $= P^{max}/N_0$. If not stated otherwise, the average SNR is set to 25 dB.

We also consider dynamic power allocation strategies with a exponentially decreasing power

$$P_t = a \cdot b^{T-t}, \tag{19}$$

where b determines the exponential decay rate of allocated power, and a is a scaling factor selected to ensure the power constraint (13). The rationale for considering this type of power allocation is that errors made in earlier sensing slots may have a cascading effect, causing further degradation compared to errors affecting later slots.

For digital transmission, we use quadrature phase shift keying (QPSK) modulation, providing $B = 2$ bits per subcarrier, along with low-density parity-check (LDPC) coding with rate $r = 1/2$ using the implementation in Nvidia's Sionna [57]. After channel estimation, equalization and demodulation, the bit sequence is decoded using the belief propagation algorithm provided in [57]. As detailed in Sec. V-B.1, the recovered bit stream is mapped to the AER packets, and the corresponding multi-level spikes are fed to the decoding SNN.

For analog transmission, as presented in Sec. IV-D, each m-bit multi-level spike is quantized into one of the 2^m PAM constellation points, with the absence of a spike for a neuron corresponding to idle subcarriers. Specifically, each output neuron of the encoding SNN is mapped to one subcarrier in each OFDM symbol. Thus, each PAM symbol is transmitted N^{OFDM} times. The received OFDM symbols are equalized and averaged before being demodulated into PAM symbols. Finally, the demodulated PAM symbols are mapped back to multi-level spikes.

3) EXPERIMENTAL SETTING

To validate the proposed system in a real-world scenario, we implement the proposed end-to-end neuromorphic wireless remote inference system using USRP SDRs [58]. As illustrated in Fig. 5, one USRP board is configured as the
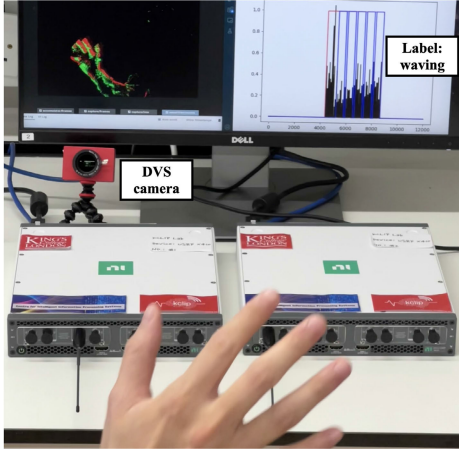
FIGURE 5. The experimental setup includes a DVS sensor, a transmitter and a receiver. The screen visualizes the event-based input of the DVS sensor (left), along with the corresponding received OFDM signal at the receiver and the gesture type detected by the decoding SNN (right).

FIGURE 6. Accuracy versus number of sensing slots T for a centralized implementation of a conventional SNN classifier with $m = 0$-bit payloads, as well as for multi-level SNN classifiers with $m = 2, 4, 6, 8$-bit payloads.

FIGURE 7. Energy consumption for a conventional SNN ($m = 0$) and for multi-level SNNs ($m = 2, 4, 6, 8$) compared to an ANN with the same architecture. The parameter γ reflects the hardware efficiency in processing multi-level spikes, with $\gamma = 0$ corresponding to a maximally efficient system. The test accuracy is reported in parenthesis.

Tx connected to the DVS camera, while the other USRP board serves as the Rx.

We set up the radios to exchange one frame per sensing slot of 130 ms. To permit synchronization of the receiver, we assume the transmission of digital pilot signals by following the standard implementation detailed in [59]. Each frame consists of a known preamble for synchronization, followed by N^{OFDM} OFDM symbols with $N^D = 512$ data subcarriers and $N^P = 75$ pilot subcarriers as for the simulations described above. The carrier frequency is set to 3.58 GHz, with the gain configured to 50 dB, and both the transmit and receive rates set to 10^6 symbols per second. With these choices, an OFDM symbol lasts for 0.714 ms, which is significantly shorter than the duration of a single sensing slot.

The distance between the transmitter and receiver was set to approximately 1 meter in an indoor environment with a line-of-sight (LoS) path, ensuring minimal external interference. The system bandwidth is determined by the sampling rate of 1 MHz, with the active subcarriers spanning an effective bandwidth of approximately 0.7 MHz. The signal-to-noise ratio (SNR) at the receiver was observed to be approximately 30 dB. Note that since the USRP is an uncalibrated device, the configured gain value does not correspond to an exact transmit power level [58].

As mentioned, we select the pre-trained SNN models fixed as a function of the parameters T and m throughout this section. However, as we will detail, we found that this approach does not work well with analog modulation when implemented using SDRs. Therefore, for analog transmission, we also considered fine-tuning the encoding and decoding SNNs to the given deployment via end-to-end training with measured channels. To this end, we first measure a set of $10,000$ channel samples via the receiving USRP board. These samples are then used to simulate channel transmission during fine-tuning. In order to differentiate the
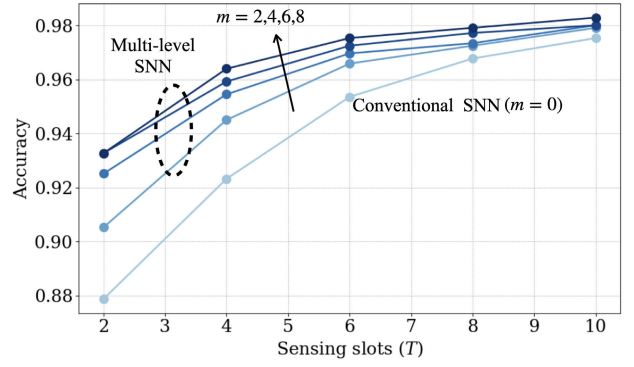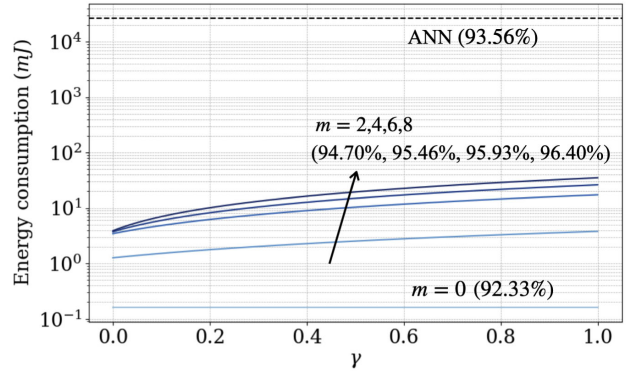
quantization error loss through the PAM modulation mapping producing the transmitted symbols, we approximate the underlying quantizer via a temperature-scaled softmax function as in [60].

B. RESULTS

1) NOISELESS CHANNEL

To start, Fig. 6 illustrates the accuracy performance of the pre-trained SNN model as a function of the number T of sensing slots in a fully centralized implementation. This performance serves as a benchmark for the wireless split computing system to be investigated next. Increasing T enhances the informativeness of the input by extending the sensing period. Accordingly, the accuracy of the SNN classifier increases with the input duration T.

The figure compares results obtained with SNN models processing multi-level spikes with a different size m of the spike payload in bits. Compared to the conventional SNN model with $m = 0$, multi-level spikes consistently achieve
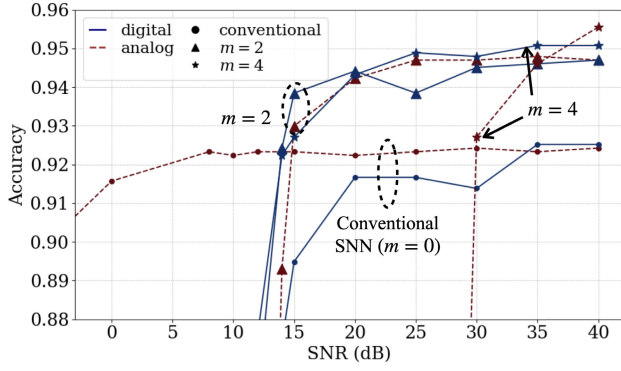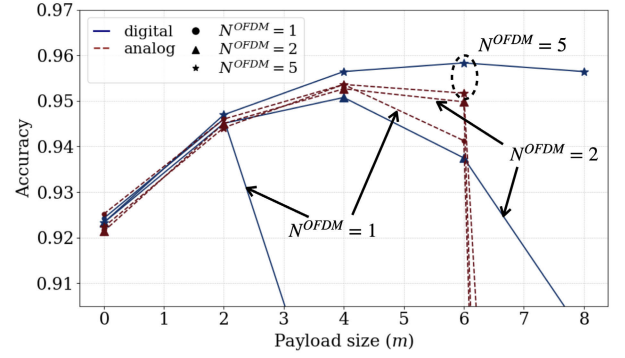
(a) Per-block power constraint
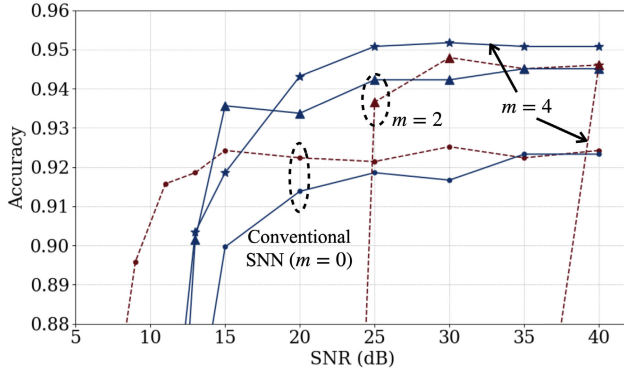
(b) Peak power constraint

FIGURE 8. Accuracy versus SNR for the neuromorphic wireless split computing architecture for analog and digital transmission schemes and: (a) per-block power constraint, and (b) peak power constraint (simulation, $T = 4$).

(a) Per-block power constraint

(b) Peak power constraint

FIGURE 9. Accuracy versus payload size m for the neuromorphic wireless split computing architecture for analog and digital transmission schemes, using different number of OFDM symbols and: (a) per-block power constraint, and (b) peak power constraint (simulation, $T = 4$).
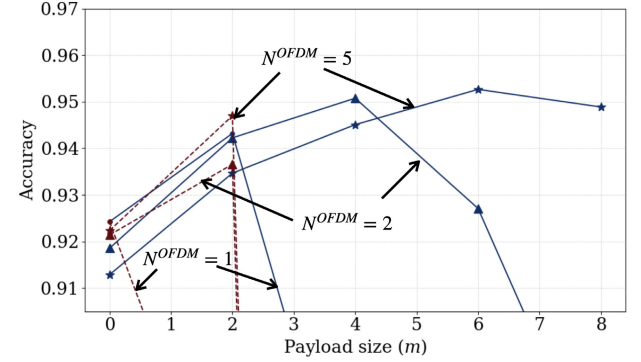
higher accuracy, particularly at earlier sensing slots. For example, at $T = 4$, the conventional SNN model reaches an accuracy of 92.33%, while a multi-level SNN with $m = 2$ attains 94.70% and $m = 8$ yields an accuracy of 96.40%. In contrast, for $T = 10$, the conventional SNN achieves an accuracy of 97.54%, which is similar to the result obtained in [61] using parametric LIF neurons, while a multi-level SNN achieves 98.30%.

The outlined performance gains attained with multi-level spikes may not be retained in a wireless split architecture due to an inherent trade-off between the number of transmitted bits and the accuracy of the decoded bits. To illustrate the potential benefits of multi-level spikes in this context, based on Fig. 6, in the following we set $T = 4$, adopting the corresponding pre-trained SNNs models for different values of m.

To assess inference energy consumption, following [62], we count the number of accumulate operations carried out by the SNN, and adopt the model (6) described in Section II-B with $E_{ac} = 0.1$ pJ [63]. For comparison, in a manner similar to [10], we also consider the performance of an ANN with the same architecture of the SNN with a ReLU activation function and with a softmax output layer. The ANN is trained

on the same data. The ANN takes the same input of the SNN at each sensing slot t, and makes a final decision by selecting the index with the highest average output of the softmax output layer over T slots. The energy for multiply-and-accumulate operations for the ANNs is set to $E_{mac} = 3.2$ pJ [63].

Fig. 7 illustrates that increasing the number of bits m from 0 to 8 leads to a modest increase in energy consumption, while yielding a substantial accuracy increase from 92.33% to 96.40%. In all cases, the energy consumption remains lower than that of the ANN model. Furthermore, the lower accuracy of the ANN model is due to its lack of a memory mechanism for temporal processing.

2) SIMULATION RESULTS

We first analyze the impact of the average SNR in Fig. 8 for both analog and digital modulation schemes with $N^{\text{OFDM}} = 2$ OFDM symbols under a per-block power constraint and a peak power constraint. A conventional SNN with analog modulation provides the best performance at lower SNR levels, especially under an average power constraint. In fact, a peak power constraint limits the power that can be used per spike, while an average power constraint makes it possible
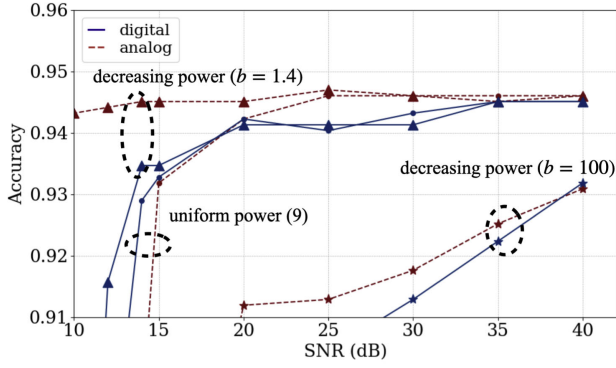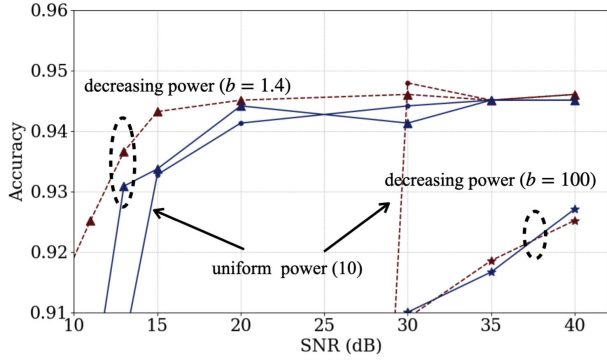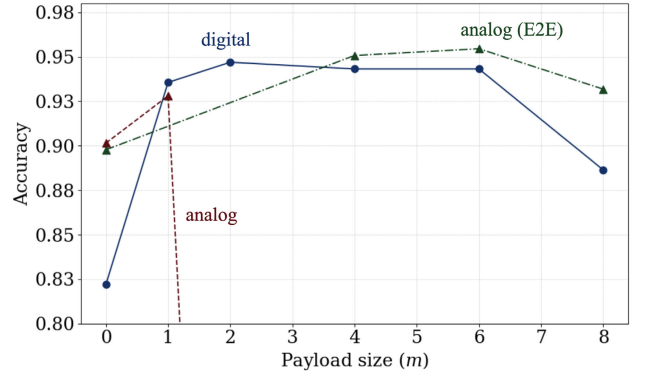
FIGURE 10. Accuracy versus SNR for the neuromorphic wireless split computing architecture for analog and digital transmission schemes using: (a) per-block power constraint, and (b) peak power constraint (simulation, $m = 2$ and $T = 4$). For both types of power constraints, we consider time-uniform power allocation, as well as decreasing power allocation schemes as in (19) with different exponential decay rates b.

FIGURE 11. Comparison between analog and digital transmission schemes for the neuromorphic wireless split computing architecture. (a) Accuracy versus payload size m with $N^{OFDM} = 2$, and (b) accuracy versus number of OFDM symbol N^{OFDM} per sensing slot with payload size $m = 6$ (USRP implementation, $T = 4$).
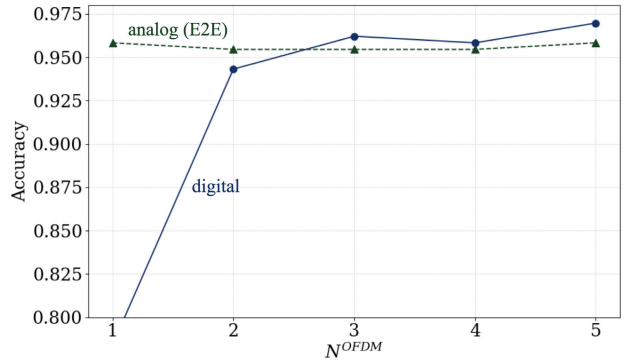
to allocate power by leveraging the sparsity of the spiking signals. At higher SNRs, it becomes essential to rely on muti-level spikes. In general, increasing the SNR calls for the selection of a larger value of m. This is aligned with the performance of the centralized implementation shown in Fig. 6.

The optimal value of the payload size m generally depends on the SNR, on the modulation schemes, and on the available spectrum, which is controlled by the number of OFDM symbols N^{OFDM}. To elaborate on this, Fig. 9 presents the accuracy as a function of the bit width m when the SNR is fixed at 25 dB, while varying also the number of OFDM symbols. The results indicate that for both modulation schemes, there is an optimal value of m that strikes the best balance between increased inference accuracy and decreased transmission reliability caused by a larger value of m.

In this regard, digital modulation is more sensitive to a decrease in spectral resources, showing a significant accuracy drop when $m > 2$ and $N^{OFDM} = 1$, as limited resources force some of the spikes to be discarded. However, when the number of OFDM symbols is sufficiently large, such as

$N^{OFDM} = 5$, allowing most spikes to be transmitted, digital modulation can better capitalize on an increasing value of m. Under either power constraint, analog modulation also benefits from an increase in the number of OFDM symbols N^{OFDM}, becoming more robust through repetition coding.

Fig. 10 presents the accuracy performance as a function of SNR for analog and digital transmission schemes under different power allocation strategies with $m = 2$ and $T = 4$. The results show that the decaying power allocation strategy in (19) can outperform fixed power allocation with a suitably chosen decay rate b, here $b = 1.4$. This validates that errors occurring in earlier sensing slots can have a cascading effect, leading to larger degradation compared to those in later slots. However, an excessively large value of b, here $b = 100$, can underperform uniform power allocation.

3) EXPERIMENTAL RESULTS

We now turn to the results obtained from the real-world experiments with USRP radios. Fig. 11(a) presents the accuracy versus the number of bits m for $N^{OFDM} = 2$. As discussed, for the analog implementation, we show the performance with pre-trained models, as well as with end-to-end (E2E)

fine-tuning with measured channels. As seen in the simulation in Fig. 9, there exists an optimal value of m for all schemes. Furthermore, the figure highlights the importance of incorporating channel data in the optimization of a deployment that relies on analog transmission.

Fig. 11(b) shows the accuracy versus the number of OFDM symbols, N^{OFDM}, for $m = 6$ bits, focusing on E2E fine-tuning for analog transmission. It is observed that digital modulation requires a sufficiently large number of OFDM symbols in order not to be limited by the accuracy degradation caused by spikes being dropped at the transmitter. In contrast, analog transmission can obtain the best performance even with only one OFDM symbol.

VII. CONCLUSION

In this paper, we have studied a neuromorphic wireless split computing architecture that leverages multi-level SNNs. Multi-level SNN models are known to achieve higher accuracy than conventional SNNs, especially in the presence of strict constraints on the sensing period. However, in a split computing system, these gains may be offset by the challenges of exchanging multi-level spikes between the SNN models deployed across two separate devices. To address this problem, we have developed digital and analog modulation schemes optimized for an OFDM radio interface. Simulations and experiments with software-defined radios have accordingly revealed optimal configurations in terms of the size of the spike payload for both analog and digital transmission schemes. Analog transmission was seen to perform better at lower SNR levels and for smaller payload sizes, while digital transmission was seen to be more effective at higher SNR levels and for larger payloads. Finally, experimental results have demonstrated the need for channel-specific fine-tuning of the SNN models for analog transmission. Future work may consider extensions to multi-terminal settings [10], the use of UWB for short-range low-power communications, and the problem of dynamic power optimization across sensing slots, e.g., via reinforcement learning.

REFERENCES

[1] K. Bourzac, "Fixing AI's energy crisis," *Nature*, vol. 628, no. 7980, pp. S20–S22, Oct. 2024.

[2] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnol.*, vol. 15, no. 7, pp. 529–544, Jul. 2020.

[3] S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, and R. Douglas, *Event-Based Neuromorphic Systems*. Hoboken, NJ, USA: Wiley, 2014.

[4] O. Simeone et al., "Learning algorithms and signal processing for brain-inspired computing [from the guest editors]," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 12–15, Nov. 2019.

[5] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, pp. 195–202, Sep. 2017.

[6] O. Simeone, "An introduction to quantum machine learning for engineers," *Found. Trends Signal Process.*, vol. 16, nos. 1–2, pp. 1–223, 2022.

[7] D. Baker, D. Hassabis, and J. Jumper, *WATCH: David Baker, Demis Hassabis and John Jumper Win 2024 Nobel Prize in Chemistry for Breakthrough Work on Proteins*. Arlington County, VA, USA: PBS NewsHour, 2024. [Online]. Available: https://www.pbs.org/newshour/science/watch-live-the-winner-of-the-2024-nobel-prize-in-chemistry-is

[8] C. Bartolozzi, G. Indiveri, and E. Donati, "Embodied neuromorphic intelligence," *Nature Commun.*, vol. 13, no. 1, p. 1024, Feb. 2022.

[9] N. Skatchkovsky, H. Jang, and O. Simeone, "End-to-end learning of neuromorphic wireless systems for low-power edge artificial intelligence," in *Proc. 54th Asilomar Conf. Signals, Syst., Comput.* Monterey, CA, USA: IEEE Computer Society, Nov. 2020, pp. 166–173.

[10] J. Chen, N. Skatchkovsky, and O. Simeone, "Neuromorphic wireless cognition: Event-driven semantic communications for remote inference," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 2, pp. 252–265, Apr. 2023.

[11] J. Chen, N. Skatchkovsky, and O. Simeone, "Neuromorphic integrated sensing and communications," *IEEE Wireless Commun. Lett.*, vol. 12, no. 3, pp. 476–480, Mar. 2023.

[12] J. Chen, S. Park, P. Popovski, H. V. Poor, and O. Simeone, "Neuromorphic split computing with wake-up radios: Architecture and design via digital twinning," *IEEE Trans. Signal Process.*, vol. 72, pp. 4635–4650, 2024.

[13] Z. Song, O. Simeone, and B. Rajendran, "Neuromorphic in-context learning for energy-efficient MIMO symbol detection," in *Proc. IEEE 25th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2024, pp. 1–5.

[14] Y. Liu, Z. Qin, and G. Y. Li, "Energy-efficient distributed spiking neural network for wireless edge intelligence," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 10683–10697, Mar. 2024.

[15] A. Gupta, O. Dizdar, Y. Chen, and S. Wang, "SpikingRx: From neural to spiking receiver," 2024, *arXiv:2409.05610*.

[16] M. Martini, J. Adhuran, and N. Khan, "Lossless compression of neuromorphic vision sensor data based on point cloud representation," *IEEE Access*, vol. 10, pp. 121352–121364, 2022.

[17] S. B. Shrestha, J. Timcheck, P. Frady, L. Campos-Macías, and M. Davies, "Efficient video and audio processing with Loihi 2," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jan. 2023, pp. 13481–13485.

[18] B. H. Theilman et al., "Spiking physics-informed neural networks on Loihi 2," in *Proc. Neuro Inspired Comput. Elements Conf. (NICE)*, Apr. 2024, pp. 1–6.

[19] O-RAN Alliance. (2020). *O-RAN: Towards an Open and Smart RAN*. [Online]. Available: https://www.o-ran.org/resources

[20] D. Wu, X. Yi, and X. Huang, "A little energy goes a long way: Build an energy-efficient, accurate spiking neural network from convolutional neural network," *Frontiers Neurosci.*, vol. 16, May 2022, Art. no. 759900.

[21] M. Davies et al., "Advancing neuromorphic computing with Loihi: A survey of results and outlook," *Proc. IEEE*, vol. 109, no. 5, pp. 911–934, May 2021.

[22] D. Yu, X. Du, L. Jiang, W. Tong, and S. Deng, "EC-SNN: Splitting deep spiking neural networks for edge devices," in *Proc. 33rd Int. Joint Conf. Artif. Intell.*, Aug. 2024, pp. 5389–5397.

[23] N. Skatchkovsky, H. Jang, and O. Simeone, "Federated neuromorphic learning of spiking neural networks for low-power edge intelligence," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8524–8528.

[24] T. Borsos, M. Condoluci, M. Daoutis, P. Hága, and A. Veres, "Resilience analysis of distributed wireless spiking neural networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Austin, TX, USA, Apr. 2022, pp. 2375–2380.

[25] Y. Venkatesha, Y. Kim, L. Tassiulas, and P. Panda, "Federated learning with spiking neural networks," *IEEE Trans. Signal Process.*, vol. 69, pp. 6183–6194, 2021.

[26] Y. Matsubara and M. Levorato, "Split computing for complex object detectors: Challenges and preliminary results," in *Proc. 4th Int. Workshop Embedded Mobile Deep Learn.*, Sep. 2020, pp. 7–12.

[27] Prophesee. *Event-Based Vision Applications*. Accessed: Oct. 15, 2024. [Online]. Available: https://www.prophesee.ai/

[28] IniVation. *Extreme Machine Vision*. Accessed: Oct. 15, 2024. [Online]. Available: https://inivation.com/

[29] SynSens. *Speck*. Accessed: Oct. 15, 2024. [Online]. Available: https://www.synsense.ai/products/speck-2/

[30] Innatera. *Ultra Low Power Intelligence for the Sensor Edge*. Accessed: Oct. 15, 2024. [Online]. Available: https://innatera.com/

[31] T. Tchumatchenko and A. K. Magnusson, "Membrane potential fluctuations govern spike amplitude variability in cortical pyramidal neurons," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 5, pp. 564–569, 2014.

[32] J. Qu, Z. Gao, T. Zhang, Y. Lu, H. Tang, and H. Qiao, "Spiking neural network for ultralow-latency and high-accurate object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 4934–4946, Mar. 2025.

[33] Z. Hao, X. Shi, Y. Liu, Z. Yu, and T. Huang, "LM-HT SNN: Enhancing the performance of SNN to ANN counterpart through learnable multi-hierarchical threshold model," 2024, *arXiv:2402.00411*.

[34] Y. Xiao, X. Tian, Y. Ding, P. He, M. Jing, and L. Zuo, "Multi-bit mechanism: A novel information transmission paradigm for spiking neural networks," 2024, *arXiv:2407.05739*.

[35] A. Rácz, A. Veres, P. Hága, T. Borsos, and Z. Kenesi, "A full-stack neuromorphic prototype architecture for low-power wireless sensors," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2022, pp. 353–358.

[36] J. Lee et al., "An asynchronous wireless network for capturing event-driven data from large populations of autonomous sensors," *Nature Electron.*, vol. 7, no. 4, pp. 313–324, Mar. 2024.

[37] Y. Guo et al., "Ternary spike: Learning ternary spikes for spiking neural networks," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 11, pp. 12244–12252.

[38] X. Luo, M. Yao, Y. Chou, B. Xu, and G. Li, "Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection," 2024, *arXiv:2407.20708*.

[39] A. Marwanto, M. Adib Sarijari, N. Fisal, S. Kamilah Syed Yusof, and R. A. Rashid, "Experimental study of OFDM implementation utilizing GNU radio and USRP–SDR," in *Proc. IEEE 9th Malaysia Int. Conf. Commun. (MICC)*, Dec. 2009, pp. 132–135.

[40] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic DVS events," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1312–1321.

[41] J. Chen, S. Park, and O. Simeone, "Knowing when to stop: Delay-adaptive spiking neural network classifiers with reliability guarantees," *IEEE J. Sel. Topics Signal Process.*, vol. 19, no. 1, pp. 88–102, Jan. 2025.

[42] D. Wu, Y. Qi, K. Cai, G. Jin, X. Yi, and X. Huang, "Direct training needs regularisation: Anytime optimal inference spiking neural network," 2024, *arXiv:2405.00699*.

[43] Z. Song, P. Katti, O. Simeone, and B. Rajendran, "Xpikeformer: Hybrid analog-digital hardware acceleration for spiking transformers," 2024, *arXiv:2408.08794*.

[44] S. Shen et al., "SpikingSSMs: Learning long sequences with sparse and parallel spiking state space models," 2024, *arXiv:2408.14909*.

[45] G. Orchard et al., "Efficient neuromorphic signal processing with Loihi 2," in *Proc. IEEE Workshop Signal Process. Syst. (SiPS)*, Oct. 2021, pp. 254–259.

[46] A. Mehonić et al., "Roadmap to neuromorphic computing with emerging technologies," *APL Mater.*, vol. 12, no. 10, Oct. 2024, Art. no. 109201.

[47] G. Bellec, D. Salaj, A. Subramoney, R. Legenstein, and W. Maass, "Long short-term memory and learning-to-learn in networks of spiking neurons," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2018, pp. 1–11.

[48] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.

[49] T. Bu, W. Fang, J. Ding, P. Dai, Z. Yu, and T. Huang, "Optimal ANN-SNN conversion for high-accuracy and Ultra-low-latency spiking neural networks," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2023, pp. 1–20.

[50] Y. He et al., "An implantable neuromorphic sensing system featuring near-sensor computation and send-on-delta transmission for wireless neural sensing of peripheral nerves," *IEEE J. Solid-State Circuits*, vol. 57, no. 10, pp. 3058–3070, Oct. 2022.

[51] A. Amir et al., "A low power, fully event-based gesture recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7243–7252.

[52] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[53] J. A. Perez-Carrasco, B. Acha, C. Serrano, L. Camunas-Mesa, T. Serrano-Gotarredona, and B. Linares-Barranco, "Fast vision through frameless event-based sensing and convolutional processing: Application to texture recognition," *IEEE Trans. Neural Netw.*, vol. 21, no. 4, pp. 609–620, Apr. 2010.

[54] O. Simeone, Y. Bar-Ness, and U. Spagnolini, "Pilot-based channel estimation for OFDM systems by tracking the delay-subspace," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 315–325, Jan. 2004.

[55] S. Deng, Y. Li, S. Zhang, and S. Gu, "Temporal efficient training of spiking neural network via gradient re-weighting," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–17.

[56] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Jan. 2008.

[57] J. Hoydis et al., "Sionna: An open-source library for next-generation physical layer research," 2022, *arXiv:2203.11854*.

[58] *Ettus USRP X410*. Accessed: Nov. 12, 2024. [Online]. Available: https://www.ni.com/docs/en-U.S./bundle/ettus-usrp-x410-specs/page/specs.html

[59] M. Matthe. *The Schmidl & Cox Synchronization Technique for OFDM*. Accessed: Nov. 12, 2024. [Online]. Available: https://dspillustrations.com/pages/posts/misc/schmidlcox-synchronization-for-ofdm.html

[60] J. Huang, S. Park, and O. Simeone, "Calibration-aware Bayesian learning," in *Proc. IEEE 33rd Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2023, pp. 1–6.

[61] W. Fang, Z. Yu, Y. Chen, T. Masquelier, T. Huang, and Y. Tian, "Incorporating learnable membrane time constant to enhance learning of spiking neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2661–2671.

[62] H. Yang et al., "Lead federated neuromorphic learning for wireless edge artificial intelligence," *Nature Commun.*, vol. 13, no. 1, p. 4269, Jul. 2022.

[63] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.

DENGYU WU (Member, IEEE) received the Ph.D. degree from the University of Liverpool, U.K. He is currently a Research Associate with the King's Laboratory for Intelligent Computing (KLIC), Centre for Intelligent Information Processing Systems (CIIPS), Department of Engineering, King's College London, U.K. His research interests include neuromorphic computing and algorithm-hardware co-design for AI systems, particularly in the development of efficient AI.
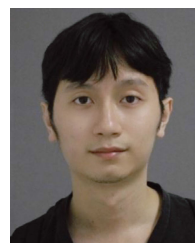
JIECHEN CHEN (Member, IEEE) received the Ph.D. degree from King's College London, U.K. He is currently a Research Associate with the King's Communications, Learning and Information Processing (KCLIP), Centre for Intelligent Information Processing Systems (CIIPS), Department of Engineering, King's College London. His research interests include neuromorphic computing and its applications in energy-efficient wireless communications, particularly in edge-based Internet of Things (IoT) systems.

BIPIN RAJENDRAN (Senior Member, IEEE) received the B.Tech. degree from IIT Kharagpur in 2000 and the M.S. and Ph.D. degrees in electrical engineering from Stanford University in 2003 and 2006, respectively. From 2006 to 2012, he was a Master Inventor and a Research Staff Member with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA, and has since held faculty positions in India and USA. He is currently a Professor of intelligent computing systems and an EPSRC Fellow at King's College London (KCL). He also co-directs the Centre for Intelligent Information Processing Systems, Department of Engineering, and the King's Institute for Human and Synthetic Minds. He has co-authored more than 100 peer-reviewed articles, one monograph, an edited book, and holds 59 U.S. patents. His research interests include brain-inspired computing, neuromorphic systems, and hardware accelerators.
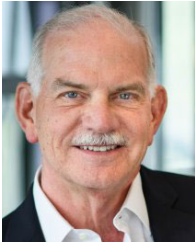
H. VINCENT POOR (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Caltech. His research interests include information theory, machine learning and network science, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022). He is a member of the National Academy of Engineering and the National Academy of Sciences, and a foreign member of the Royal Society and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.

OSVALDO SIMEONE (Fellow, IEEE) was a Faculty Member with the Electrical and Computer Engineering (ECE) Department, New Jersey Institute of Technology (NJIT), from 2006 to 2017. He is currently a Professor of information engineering and co-directs the Centre for Intelligent Information Processing Systems, Department of Engineering, King's College London. He is also a Visiting Professor with the Connectivity Section, Department of Electronic Systems, Aalborg University. He is the author of the textbook *Machine Learning for Engineers* (Cambridge University Press), four monographs, two edited books, and more than 200 research journals and magazine papers. He is a fellow of the IET and EPSRC. Among other recognitions, he was a co-recipient of the 2022 IEEE Communications Society Outstanding Paper Award, the 2021 IEEE Vehicular Technology Society Jack Neubauer Memorial Award, the 2019 IEEE Communication Society Best Tutorial Paper Award, the 2018 IEEE Signal Processing Best Paper Award, and the 2015 IEEE Communication Society Best Tutorial Paper Award. He was also awarded an Open Fellowship by the EPSRC in 2022 and a Consolidator grant by European Research Council (ERC) in 2016. He was a Distinguished Lecturer of the IEEE Communications Society in 2021 and 2022 and the IEEE Information Theory Society in 2017 and 2018.