Journal of Neural Engineering

PAPER

A low-power communication scheme for wireless, 1000 channel brain–machine interfaces

To cite this article: Joseph T Costello *et al* 2022 *J. Neural Eng.* **19** 036037

View the article online for updates and enhancements.

You may also like

- Vapor Phase Infiltration of Metal Oxides into Nanoporous Polymer Membranes for Organic Solvent Separation
Emily Kathryn McGuinness, Fengyi Zhang, Yao Ma et al.
- Selection of appropriate polyoxymethylene based binder for feedstock material used in powder injection moulding
J Gonzalez-Gutierrez, G B Stringari, Z M Megen et al.
- Code Conversion Method on Process-in-Memory Platform
Jinhui Liu, Chen Zhao, Fangzhou Du et al.

This content was downloaded from IP address 141.214.17.120 on 26/08/2022 at 14:24

Journal of Neural Engineering

PAPER

A low-power communication scheme for wireless, 1000 channel brain–machine interfaces

Joseph T Costello[1], Samuel R Nason-Tomaszewski[2], Hyochan An[1], Jungho Lee[1], Matthew J Mender[2], Hisham Temmar[2], Dylan M Wallace[3], Jongyup Lim[1], Matthew S Willsey[2,4], Parag G Patil[2,4], Taekwang Jang[5], Jamie D Phillips[1,6], Hun-Seok Kim[1], David Blaauw[1] and Cynthia A Chestek[2,3,*]

[1] Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, United States of America
[2] Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, United States of America
[3] Robotics Institute, University of Michigan, Ann Arbor, MI, United States of America
[4] Department of Neurosurgery, University of Michigan Medical School, Ann Arbor, MI, United States of America
[5] Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland
[6] Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: cchestek@umich.edu

Keywords: brain–machine interface, neural interface, low-power, wireless communication

Supplementary material for this article is available online

RECEIVED
23 February 2022

REVISED
25 April 2022

ACCEPTED FOR PUBLICATION
24 May 2022

PUBLISHED
14 June 2022

Abstract

Objective. Brain–machine interfaces (BMIs) have the potential to restore motor function but are currently limited by electrode count and long-term recording stability. These challenges may be solved through the use of free-floating 'motes' which wirelessly transmit recorded neural signals, if power consumption can be kept within safe levels when scaling to thousands of motes. Here, we evaluated a pulse-interval modulation (PIM) communication scheme for infrared (IR)-based motes that aims to reduce the wireless data rate and system power consumption. *Approach*. To test PIM's ability to efficiently communicate neural information, we simulated the communication scheme in a real-time closed-loop BMI with non-human primates. Additionally, we performed circuit simulations of an IR-based 1000-mote system to calculate communication accuracy and total power consumption. *Main results*. We found that PIM at 1 kb/s per channel maintained strong correlations with true firing rate and matched online BMI performance of a traditional wired system. Closed-loop BMI tests suggest that lags as small as 30 ms can have significant performance effects. Finally, unlike other IR communication schemes, PIM is feasible in terms of power, and neural data can accurately be recovered on a receiver using 3 mW for 1000 channels. *Significance.* These results suggest that PIM-based communication could significantly reduce power usage of wireless motes to enable higher channel-counts for high-performance BMIs.

1. Introduction

Brain machine interfaces (BMIs) have shown potential for restoring movement and communication to those who suffer from spinal cord injury. BMIs estimate user intentions by recording from electrodes implanted in cortical regions and processing neural data with a decoding algorithm. These systems have allowed participants to control prosthetic arms [1, 2], write text [3], and functionally stimulate paralyzed limbs [4]. Current BMIs in humans use wired electrode arrays, most commonly the Utah array [5, 6]. While Utah arrays have enabled some success in decoding intentions, their performance is still significantly below able-bodied control; part of this performance drop may be attributed to the low neuronal yield [7, 8] resulting in few tuned channels. While multiple Utah arrays can be implanted to increase channel count, it is infeasible to implant tens of arrays in a single cortical region, which would be required for hundreds to thousands of tuned channels. Thus, recent efforts have looked toward other electrode technologies for chronic recording stability and increasing channel counts to improve performance of intracortical BMIs.

© 2022 IOP Publishing Ltd

To increase the number of recording channels, several groups have developed rigid electrode arrays with thousands of channels: The Argo [9] can record from 65k channels using a microwire bundle with 60–300 μm pitch. Neuropixels 2.0 [10] has up to 10k channels on an implant at a 15 μm pitch. While high channel counts could enable better BMI control, these systems have several limitations, which make them difficult to use for chronic motor BMI. First, with thousands of wires, the interconnect for these systems requires large, chronically open holes in the dura mater which may become failure points. This rigid connectorization may also experience micromotion relative to the brain, which may cause unstable recordings, glial scarring and cell loss near the electrode. From the perspective of BMI performance, these high-density electrodes have a large number of recording sites spaced closely together. Therefore, they record from small brain areas, which may have highly correlated neural activity [11]; dispersing channels across a wider region could record from cells with more independent information, allowing for better decoding across more degrees of freedom [12].

Significant progress has been made in developing thin, flexible electrodes that aim to move with the brain and reduce tissue damage. These include the mesh probe [13], which achieves better biocompatibility through a flexible 3D mesh, the neural matrix [14], a chronic surface array with over 1000 channels, and a polymer-based microelectrode array with 512 channels [15]. Notably, Neuralink has developed a multi-thousand channel recording system using polymer 'thread' electrodes connected to a custom recording ASIC [16]. As the channel count of flexible systems increases, however, the wire interconnects must become increasingly unwieldy, resulting in additional failure points. Despite potentially minimal tissue damage, soft electrodes still require chronic openings in the dura mater for the wire interconnect [17].

Alternatively, wireless dust [18, 19] or 'motes' show promise in solving the challenges of recording stability and high channel counts, without being limited by the interconnect. Each mote receives wireless power, records single-channel neural data, and transmits data to a receiver. Since these devices are free-floating, they can move with the brain for reduced tissue scarring and can be covered by dura, thereby reducing the risk of cerebrospinal fluid leak and infection. They can be implanted in any desired configuration across large regions. However, a primary challenge with wireless motes is minimizing power usage to limit tissue heating to safe levels, while supporting communication with up to thousands of motes. Several wireless power transfer and communication modalities have been previously explored: The original neural dust [18, 19] is a sub-mm mote powered through ultrasound that can be implanted as deep as 5 cm [20]. The neurograin [21] is powered through a radiofrequency (RF) link and could theoretically support up to 770 devices. Finally, the Michigan mote, the focus of this paper (figure 1(a) [22, 23]) and OWIC [24] are powered through infrared (IR) light. Other power modalities and designs have been proposed for stimulation motes, but their power requirements are significantly higher. With each power modality there is a tradeoff between maximum device power and device size, and unique considerations with respect to energy harvesting efficiency and communications methods with device size scaling. Motes using IR power transfer and communications are particularly promising because they can be nearly an order of magnitude smaller than RF and ultrasound-based motes, but have the lowest available power due to the reduced mote area (Singer and Robinson [25]). For example, benchtop and acute testing has demonstrated that IR motes can be smaller than 250 μm maximum dimension [23, 24] but are then limited to 1.5 μW per device [26]. Thus, IR power may be advantageous for recording-based BMIs that require high channel density spread out across a larger area, if power constraints can be met.

The total power consumed by each mote comes from sampling, filtering, signal processing, wireless communication, and maintaining a clock; the wireless communication scheme affects all of these parts and is important to optimize. An IR communication scheme uses an optical link with the mote as a transmitter and an array of detectors as the receiver, as shown in figure 1(a). Power use for wireless communications presents a demanding requirement for the mote power budget, where the IR approach can be efficiently scaled using light pulses from a microscale light emitting diode on the mote and high-sensitivity photon counting at the receiver. One relatively simple approach for communicating with hundreds to thousands of devices is time-division multiple access (TDMA) where each device has a specific time-slot to transmit data (figure 1(b)). TDMA is used in cell phone communications as well as in the Neurograin system [21]. However, when scaling to 1000s of devices, the on-chip clock rate must be scaled to the megahertz range to support precise transmission timings [21]. While megahertz clocks are feasible for power modalities like RF, they consume too much power for IR (see section 3). Other communication schemes like frequency division multiple access or code division multiple access are challenging to implement with IR light.

Our group has proposed an alternative communication scheme for IR-based communications, pulse-interval modulation (PIM), in which neural data is encoded by the transmission rate of data packets. PIM is asynchronous (figure 1(b)), meaning motes would not have specific transmission times, and could then function with clock speeds on the

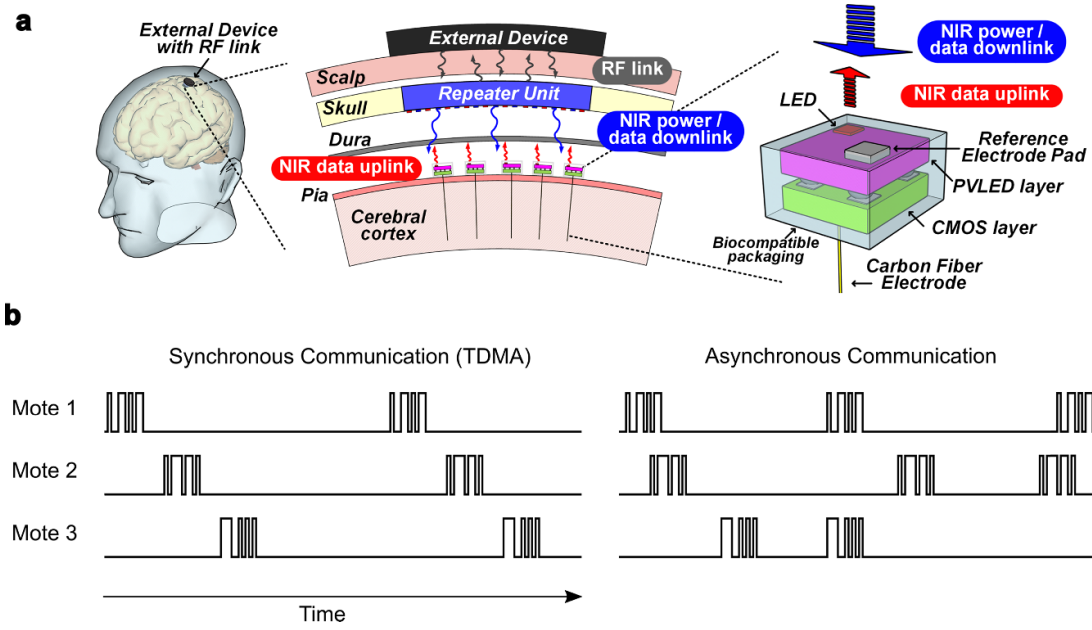

Figure 1. (a) Proposed Michigan mote system. © (2022) IEEE. Reprinted, with permission, from [22]. Motes sit beneath the dura and are powered by near-infrared (NIR) light. They record cortical signals and transmit data through an IR uplink to a receiver ('repeater') unit within the skull. The receiver transmits data from all motes to an external device through an RF link.
(b) Synchronous vs asynchronous communication schemes. Left—In synchronous schemes like TDMA, each mote has a precise timeslot to transmit data. Right—In asynchronous schemes, motes transmit at irregular intervals, and multiple motes may transmit at the same time.

order of 10 kHz regardless of the number of devices [22]. Since PIM encodes neural data in the analog domain, power-hungry analog-to-digital converters are not required. Finally, individual mote power usage and signal fidelity can be adjusted by changing the average packet rate. Specifically, PIM can encode neural features like spiking band power (SBP), or the signal power in 300–1000 Hz frequency band, which our group has previously shown is dominated by single-units, highly correlated with firing rate, and a high-performance BMI input feature [27]. Using SBP allows the sampling rate to be reduced to 2 kSps (when digitally sampling) for significantly reduced power consumption compared to the 10–30 kHz required for traditional BMIs based on recordings of the threshold-crossing signal. Unlike threshold-crossings, SBP does not require setting a channel-specific threshold, or any other commands to be processed by the device, reducing circuitry and communication needed for thresholds. We previously published the integrated-chip design of PIM-motes [22, 23], but we did not investigate performance using significantly lower data-rates in a closed-loop BMI or the feasibility of receiver complexity.

Here, we aimed to determine if IR-based PIM communication, a scheme without precisely clocked transmissions, could support a high channel count BMI in terms of real-time performance and signal recovery within power limits. By simulating PIM communication in a real-time BMI with non-human primates (NHPs), we found that performance matched the state-of-the-art at data rates of only 100 packets/s (pps) (1.3 kbit/s/channel). We found that communication schemes must consider overall lag, since lags as small as 30 ms had a significant performance effect. Additional circuit simulations suggest that, unlike other communication schemes, PIM motes can stay below IR power limits, and despite the complexity of an asynchronous scheme, a receiver could accurately decode data from 1000 motes using 3 mW. Therefore, the receiver power consumption could be of similar magnitude to wired recording systems [28] and is unlikely to be a bottleneck for the mote system.

2. Methods

2.1. Microelectrode array implants

We implanted two male rhesus macaques (Monkeys N and W, as in [29]) with Utah microelectrode arrays (Blackrock Microsystems) in primary motor cortex (M1) using the arcuate sulcus as an anatomic landmark for hand area, as described previously [30, 31]. In each animal, a subset of the 96-channels in M1, with threshold crossings morphologically consistent with action potentials, were used for offline recordings and closed-loop BMI control. Surgical procedures were performed in compliance with NIH guidelines as well as the University of Michigan's Institutional Animal Care & Use Committee and Unit for Laboratory Animal Medicine. Neural features were extracted from the Utah array recordings as described below (section 2.4).

2.2. Behavioral task

We trained Monkeys N and W to acquire virtual targets with virtual fingers by moving their physical fingers (as described in [29, 31]). During all sessions, the monkeys sat in a shielded chamber with their left arms fixed at their sides, flexed at 90° at the elbow, with left-hand fingers placed in a manipulandum. The manipulandum [31] consisted of two flat surfaces, one for each finger group, where each surface was free to rotate about a hinge at the metacarpophalangeal joints and was sized according to the finger group being used to push it (index versus MRS). Finger movements were measured by flex sensors (FS-L-0073-103-ST, Spectra Symbol) attached to both surfaces of the manipulandum and position measurements were recorded by a computer running real-time xPC Target. Flex sensors were calibrated at the start of each experiment session. The computer monitor directly in front of the monkey displayed a virtual monkey hand model (Musculo-Skeletal Modeling Software) controlled by the xPC Target computer. In hand control, the virtual hand mirrored the monkey's hand movements, whereas in brain control the virtual hand was controlled by the decoder.

Each trial began with a spherical target(s) appearing along the path of the virtual finger(s) of interest, where each target occupied 15% of the full arc of motion of the virtual finger(s). For a successful trial, the monkey was required to move its fingers such that the corresponding virtual finger(s) of interest moved into the target(s) and hold its position for 500 ms. On successful trial completion, the monkeys received a juice reward. For closed-loop decoding, the targets were presented in a center-out pattern, in which each finger group alternated between a center 'rest' target and multiple flex or extend targets, three targets in each direction [31]. Monkey N performed a two degree-of-freedom task (controlling index and middle-ring-small (MRS) finger groups [29]) while Monkey W performed a one degree-of-freedom task (all fingers move together).

2.3. Pulse-interval modulation (PIM)

In this work we evaluated simulated-PIM communication in offline and online (real-time) analyses. PIM encodes information by modulating the interval between transmitted data packets. In this application, SBP, or the power in the 300–1000 Hz band, is encoded such that the interval between packets is proportional to 1/SBP; as SBP increases the interval shortens, and as SBP decreases the interval lengthens (figure 2). In a hardware implementation, PIM sends the mote's unique 5–12 bit device identifier within each packet. With PIM, data is encoded by time-intervals rather than transmitted bits, making it extremely bit-efficient and useful for ultra-low-power applications including deep-space satellite communications [32]. By adjusting the average packet-rate one can control the signal reconstruction accuracy for a given time scale; as the packet rate increases, the full-bandwidth signal can be reconstructed with increasing fidelity. Consequently, higher packet-rates allow for better reconstruction of the SBP signal but require more transmission power. The calculations for driving PIM signals can be efficiently implemented on chip through a signal integrator and comparator, sending a data packet every time the signal integrator crosses a threshold. Adjusting this threshold adjusts the packet rate [22] (figure 2). Therefore, on SBP-modulated PIM motes, one PIM-threshold parameter would need to be set in contrast to a spike-threshold-crossing-modulated PIM mote which would require two parameters (the spike-detection-threshold and the PIM-threshold).

Typical BMIs decode average neural activity in 10–100 ms bins, making it unnecessary for motes to transmit every recording sample. Instead, the wireless bit-rate can be significantly reduced by sending only enough data to recover bin-averages rather than individual samples. Thus, PIM-motes use packets of 10–200 pps, substantially lower than the traditional sampling rate of SBP at 2 kSps while maintaining SBP's BMI-related benefits. On the receiving side, for a single channel, the following equation is used to calculate each SBP-PIM bin:

$$\mathrm{SBP}_{\mathrm{PIM}} = \frac{1}{T}\sum_{i=1}^{N+1} w_i \frac{1}{(t_i - t_{i-1})} \tag{1}$$

where T is the time length of the bin, N is the number of packets received within the bin timeframe, t_i is the receiving time of packet i, and w_i is the fraction of the packet's interval contained within the bin (intervals fully within the bin have $w_i = 1$, whereas the first and last interval represent signal in neighboring bins and have $0 < w_i < 1$). The sum is taken over $N + 1$ packets since an additional packet is required to estimate signal within the bin (explained below).

Each PIM packet encodes SBP information from the time preceding the packet arrival; in order to accurately estimate SBP near the end of a bin, the decoder must wait past the end of the bin for a packet to arrive. If a delay is not added, then the end of each bin has an unknown signal value. Thus, when binning PIM packets, a short lag (delay) must be added to account for these packets and improve reconstruction of the true SBP signal (this method was found to better reconstruct SBP bins than without-lag methods). This lag ranges from 15 ms (200 pps) to 50 ms (30 pps) and was determined through offline simulation with real recorded neural data. This extra lag for decoding bins is additive to the BMI's intrinsic lag (for signal processing and other computational time).

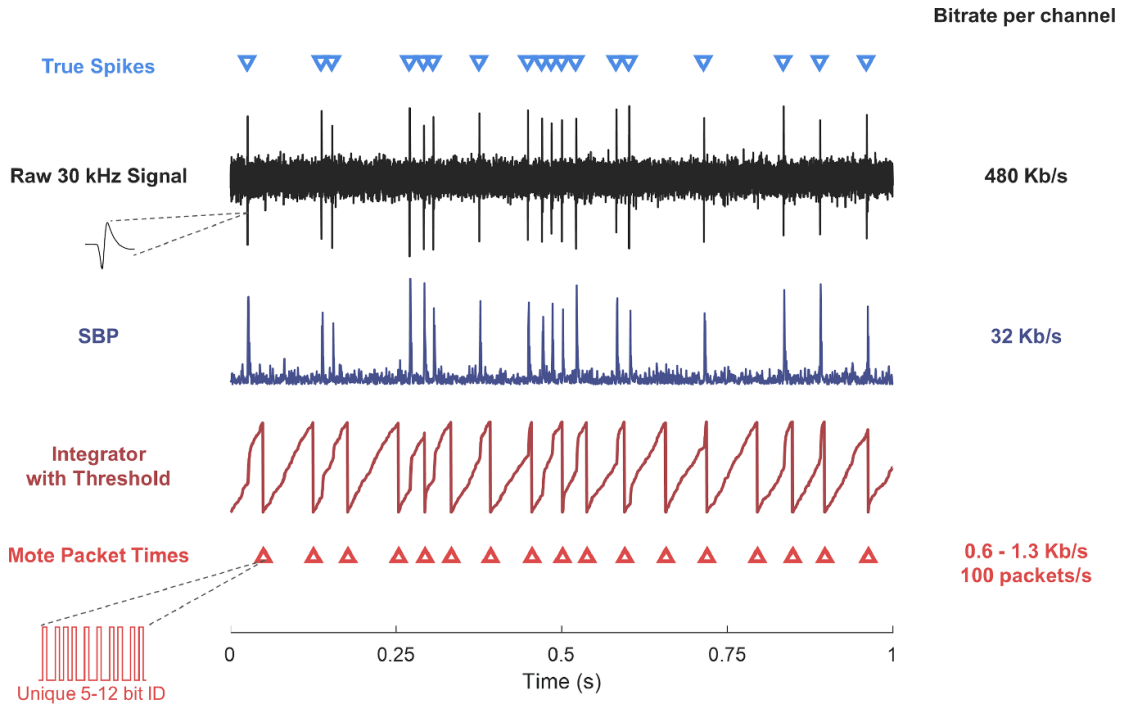

Figure 2. Overview of PIM communication (simulated data). The SBP signal (purple trace) captures much of the same spiking information as the raw 30 kHz signal (black trace) despite a much lower data rate. PIM integrates and thresholds the SBP signal (dark red trace) to determine packet transmission times (red triangles), which contain the mote's unique ID. When SBP is greater, packets are transmitted more frequently.

2.4. Signal processing and feature extraction

We recorded 96-channel raw neural data from the monkeys using a Cerebus v1.0 with firmware version 6.03.01.00 (Blackrock Microsystems) digitizing at 2 kSps. First we applied a second-order Butterworth filter to the raw data with a 300–1,000 Hz pass band and extracted the signal magnitude for SBP. Normal SBP bins were calculated by averaging the 2 kSps SBP signal in non-overlapping 16–100 ms bins. Simulated PIM packet times were calculated by integrating and thresholding the 2 kSps SBP signal (with independent thresholds set for each channel). SBP-PIM was then calculated from the packet-times, and binned similarly while accounting for the extra PIM lag (equation (1)). For real-time closed-loop decoding, we configured the Cerebus to band-pass filter the raw signals to 300–1,000 Hz. This 2 kSps continuous data was streamed to a computer running xPC Target (MathWorks), which calculated SBP and PIM-SBP in 32 ms bins to be used for decoding. Thus, the full signal processing pipeline consisted of recording raw neural data from the implanted Utah array, filtering to calculate SBP, calculating simulated PIM packet times, and averaging SBP and PIM-SBP in bins to be used for decoding.

2.5. Simulated neuron recordings

To examine the ability of PIM to accurately transmit neural information at low data-rates, we simulated single neuron recordings and calculated SBP-PIM signals. Unlike real recordings, simulated recordings enabled comparisons with the true neural firing rate to better determine signal quality at each stage of the system. Using MATLAB (MATLAB 2021a, MathWorks, Natick, MA), we simulated 30 kSps recordings using the method outlined in Nason 2020 (an example trace is shown in figure 2, black trace). Briefly, a 3 ms averaged sorted unit recorded from Monkey W was randomly placed in time at a desired average spiking rate, 6.23 μV thermal white noise was added [33], and the signal amplitude was scaled to match the specified SNR. To calculate the SBP signal (figure 1(c), purple trace), we applied a second-order Butterworth filter to the raw 30 kSps signal with a 300–1,000 Hz passband and extracted the signal magnitude. PIM signals were then simulated by integrating the SBP signal and then thresholding the integrated value, where a packet was sent whenever the integrated sum crossed the threshold (figure 2, red traces). Lastly, to determine how well the SBP and PIM signals captured true firing rate information (figure 3), PIM-based SBP (PIM-SBP) was calculated by inverting the interval between packets, all signals were smoothed using a 50 ms sliding Gaussian window, and the correlation with true firing rate was calculated.

2.6. Offline bin comparisons

In offline analyses, we compared SBP bins with PIM-SBP bins by calculating the Pearson correlation as well

as the variance accounted for (VAF). The VAF was calculated as:

$$\mathrm{VAF} = 1 - \frac{\sum_{1}^{N} (\mathrm{PIM}_i - \mathrm{SBP}_i)^2}{\sum_{1}^{N} \left(\mathrm{SBP}_i - \overline{\mathrm{SBP}}\right)^2} \quad (2)$$

where N is the number of samples (bins), SBP_i is an SBP bin, PIM_i is a PIM bin, and $\overline{\mathrm{SBP}}$ is the mean of SBP bins. Correlations and VAFs were calculated for each channel, and for both metrics a value of 1 indicates that PIM-SBP recovers the same information as SBP. Whereas correlation does not account for scaling errors, VAF takes signal scale into account and is used as the primary metric for comparison.

2.7. Closed-loop decoding

For closed-loop decoding, we applied a standard position/velocity Kalman filter with a position/velocity neural tuning model and optimizations as described previously [30] to predict the movements the monkeys made in the behavioral task. On each experiment day, the monkey first performed 400 hand-control trials to use as training data, where hand positions and neural data were simultaneously recorded. Using the training data, separate SBP and SBP-PIM Kalman filters were trained to predict finger positions from neural features in 32 ms bins. Decoders were evaluated during closed-loop brain-control trials.

A total of 13 testing sessions across 6 d were conducted for Monkey N, and a total of 15 sessions across 9 d were conducted for Monkey W. During each testing session, we alternated between SBP and SBP-PIM decoders in an A-B-A-B fashion with approximately 100 trials during individual decoder runs. This alternating method aimed to equalize any performance changes unrelated to the algorithm that occurred within the session. During analysis, the first ten trials of each run were removed while the monkey adjusted to using the decoder, and the average acquisition-time was taken across all 'A' or all 'B' trials. Since BMI performance varies across days (due to electrode micromotion, animal motivation, etc.), in some analyses, the SBP-PIM decoder performance was normalized to the average SBP decoder performance to enable cross-day analysis (as in figure 6).

As previously mentioned, PIM introduces an extra lag not present in the normal SBP decoder, which had a small negative impact on performance. In order to quantify this effect, in some sessions, we added the extra lag to the SBP decoder for an 'equalized lag' comparison. While equalizing lag may have slightly lowered SBP performance, this enabled a better comparison of the features. These extra lags are in addition to the inherent lag of 0.5 * bin-size and BMI system lags present in both decoders.

2.8. Feasibility of synchronous communication schemes

To determine if TDMA or other synchronous communication methods would be feasible for a IR mote in terms of power, we simulated the power required for the relatively fast clocks used in these schemes. For TDMA, in a best case scenario, each mote would send an 8 bit sample once every 32 ms. Assuming a 100% overhead for synchronization pulses and guard bits (reasonable due to relatively high clock variability of >1% [21, 22] and similar to the scheme in [21]), the data-rate for 1000 motes is 500 kbps. To account for signal phase synchronization, it is reasonable to require the clock to be 8× faster than the data rate [34], for a best-case (minimum) mote-clock speed of 4 MHz. Then, to estimate power of the on-chip clock, we simulated a ring oscillator in Cadence (Cadence Virtuoso 6.1.7.500.2100 and Cadence Spectre 15.1.0) using TSMC 180 nm CMOS technology. While leaving the mote digital processor connected to the clock generator for realistic output capacitance, we varied clock frequency by adjusting the current flowing through the oscillator, and measured the clock power.

2.9. Relationship between decode performance and receiver error rate

To evaluate the effect of receiver error rate on final decode performance, in an offline Kalman filter decode using 96-channel data from Monkey N, we uniformly randomly inserted false-positive packet detections and uniformly randomly removed true packets for false-negative detections. We independently varied the rate of false-positives and false-negatives, and observed the change in finger-velocity correlation and MSE.

2.10. Communication simulation and receiver design

We evaluated the feasibility of PIM communication in a 1000 channel system by simulating the IR optical transmission, detection, and filtering performed by the receiver. In the physical system, motes would be implanted on the cortical surface (beneath the dura) and the receiver unit would be within the skull several millimeters above. Here, we placed 1008 simulated motes in a 28 × 36 grid with 600 μm pitch, a simulated receiver was placed 2 mm vertically above the motes, and simulated dura mater in-between. All simulations were performed in MATLAB (MATLAB 2021a, MathWorks).

Each simulated mote transmitted at an average rate of 100 pps, where each packet contained 13 pulses (encoding the 12 bit ID through two different intervals corresponding to 0/1 bits). Packet times were generated by calculating PIM times using real recorded 30 kHz neural data; to generate 1000 channels of data, the 96-channel data was copied with time-shifts

to randomize across time. A small jitter of 16 μs $\left(\frac{1}{2} \times \frac{1}{30\,\text{kHz}}\right)$ was added to packet times to better simulate the variation in mote clock synchronization.

The simulated receiver used an array of single photon avalanche diodes (SPADs) to detect the mote light pulses. The total number of SPADs ranged from approximately 1000–9000 (see optimization in section 2.12 below). An example layout of motes and SPADs is shown in figures 7(a) and (b). Upon detecting one or more photons, the SPAD goes into a high state for the specified dead-time before being reset and able to detect photons again. Thus, SPADs are sampled at a rate equal to 1/dead-time. We simulated dead times of 100–5000 ns, corresponding to sampling rates of 200 kHz to 10 MHz.

To simulate SPAD detections of light pulses, photons were probabilistically received at each SPAD. Figure 7(a) shows the probability of a SPAD detecting a mote, based off their relative locations. This probability was calculated using the following equations:

$$N_{\text{rp}} = \eta_{\text{Dura}} \times N_{\text{tp}} \frac{A}{\pi D^2} \cos\phi \cos\psi \, 1_{\psi \leqslant \psi_{\text{FOV}}} \quad (3)$$

$$P(\text{detection}) = 1 - (1 - \text{PDE}_{\text{SPAD}})^{N_{\text{rp}}} \quad (4)$$

where N_{rp} is the number of received photons, N_{tp} is the number of transmitted photons (from the mote source), A is the SPAD detector area, D is the SPAD-to-source distance, ϕ is the radiant angle between source and SPAD, ψ is angle between the SPAD normal and source, ψ_{FOV} is the detector field-of-view (FOV), η_{Dura} is the estimated transmittance through human dura (0.3, held constant across difference path lengths as a first order approximation) [26], and PDE_{SPAD} is the photon detection efficiency of the SPAD (approximately 0.1 for 1000 nm light as in [26]). This optical link model (equation (3)) is based off [35], with further explanation on its application to motes in [26]. In equation (4), the $P(\text{detection})$ is the probability of detecting at least one photon in the pulse, which is equal to $1 - P(\text{no detections})$, where each detection is considered independent. In addition to this model, SPADs were assumed to have a dark count rate of 1000 counts/s. We assumed the mote plane to be parallel to the SPAD plane, such that $\psi = \phi$. The radius of the circular SPAD detector was set at 30–40 μm such that motes directly underneath a SPAD had >99% probability of detection. Mote light pulses consisted of 1.9×10^6 photons (assuming an injected charge of 18 pC [22] with an LED quantum efficiency of 1.7% for 1000 nm light [26]).

As seen in figure 7(a), there is relatively wide dispersion of light such that a single mote's IR pulse may be received by multiple neighboring SPADs with varied probabilities, and each SPAD may receive pulses from multiple motes. As additional motes are added, or the average packet transmission rate is increased, the probability of a SPAD receiving multiple packets simultaneously (a packet 'collision') increases. Thus, the challenge in recovering transmitted mote packets is identifying a given mote's packet time despite collisions.

To accurately recover packet-times at the simulated receiver, we employed a two-step filter process, as depicted in figure 7(b). First, temporal matched filters were run on each digital SPAD output. Each filter was matched to the ID of a mote; filters performed the convolution of the temporal ID pattern with the input signal. When all signal bits corresponding to known ID times are high, a match is declared. While this detects true packet times with high probability, false-positive matches can occur when light interference is present (for example, the case where the input signal is all high-bits results in a filter match). If a SPAD is local to multiple motes, then separate temporal filters matched to different IDs are run on the same SPAD output through temporal multiplexing.

In the second filter stage, false-positive matches were minimized by combining the filter outputs from multiple local SPADs in a 'spatial' filter. The binary temporal filter outputs were weighted by their probability of reception and summed (with 1 bit weights, this is simply the sum of local filter outputs). When this sum was greater than a preset threshold, a match was declared (figure 7(b) right and figure 7(c)). Finally, accuracy was slightly improved by removing any matches within 2 ms of another match; these must be false-positive matches since packets are 2–4 ms long (determined by the ID length) and a packet cannot be transmitted before the previous packet has finished transmitting.

In summary, the full simulation was performed as follows: First, simulated motes were placed in a grid and previously recorded neural data was used to generate packet times for each mote at 100 pps. For each pulse time within each packet, a number of photons were calculated. Then, at each SPAD, photons were probabilistically received using the relative mote locations and the light dispersion model, resulting in a binary vector of detections across time, from all motes. The temporal matched filter was then applied to this vector, and finally, the filter outputs from multiple SPADs were combined in the spatial filter to result in a binary vector of packet detections across time, for each mote.

2.11. Receiver power calculation

We estimated the power on the receiver unit required for detecting and filtering transmitted packets in a 1000 mote system, using gate-level power simulations. The estimated total power was taken as the sum of power from running SPADs, reading from and writing to SRAM, running temporal matched filters, and running spatial filters. Each SPAD used an estimated 1.2 nW per detection [36–43], where detection rates ranged from approximately 2k to 115k detections/s. As shown in figure 7(b), each SPAD

is sampled at 0.4–10 MHz and stored in an SRAM buffer, where SRAM rates are reduced by 16× using 16 bit words for each read or write. SRAM power was calculated using TSMC 28 nm HPC+ technology by summing the power required for read and write operations at a 0.5 V system voltage.

In order to more accurately estimate the power of the non-standard temporal and spatial filter circuits we performed gate-level simulations. The digital filter circuits were designed in Verilog HDL, synthesized using Synopsys DesignCompiler (Version Q-2019.12-SP4) in TSMC 0.028 μm Logic High Performance Compact Mobile Computing Plus (0.9 V/1.8 V)(CLN28HT) technology, and simulated using Cadence Xcelium simulator Version 18.03–008, with power estimated using Synopsys PrimeTime PX Version Q-2019.12. Figure 7(b) summarizes the receiver filter pipeline. Specifically, the input vectors of the gate-level simulations were the SPAD detections of the previously described simulation. Each filter consisted of thirteen 16 bit buffers, where the data in each buffer was read in from SRAM locations corresponding to the preset mote ID times. A 13 bit AND determined if all locations contained high-detections, thus performing the temporal matched filter. 1–5 separate filters were run on the output of each SPAD with different mote IDs, where SRAM read operations were temporally multiplexed. These binary temporal filter outputs were multiplied by 1 bit spatial weights, and summed (initial simulations showed no significant performance difference between 1 bit and higher-bit weights).

2.12. Receiver power optimization

To optimize receiver power, we first ran simulations of the following parameter combinations: SPAD pitch from 200 to 600 μm, FOV from 45° to 180°, and sampling rate from 1 to 10 MHz. For each simulation, we varied the number of temporal filters per mote (the number of local SPADs 'listening' for an individual mote) from 0 to 32, and found the minimum number to maintain less than 1% packet error rate (sum of false-positives and false-negatives). For each number of temporal filters, the spatial filter threshold was optimized to minimize the error rate. Then, given the specific simulation parameters and number of temporal filters per mote, receiver power was calculated. The optimal receiver parameters were chosen as those that achieved the minimum power.

We performed additional simulations with a 30° FOV; with this limited FOV, only 1 mote is visible per SPAD so that interference was eliminated and spatial filter is unnecessary. In these simulations we used a SPAD pitch of 600 μm, a FOV of 30°, and varied the SPAD diameter from 60 to 100 μm, the sampling rate from 0.4 to 1 MHz, and the number of ID bits from 5 to 12. The power optimization was performed the same as above.

3. Results

3.1. Pulse-interval modulation encodes neural information at low data-rates

We began by investigating how much the data rate could be reduced by using PIM, while maintaining features strongly correlated with the true neural firing rate. First, we simulated single neuron recordings and compared low data rate signals to the true firing rate. In this analysis, simulated recordings were used to enable comparisons with the true (known) firing rate. Figure 3 shows an example simulated neuron with a firing rate of 20 spikes per second and SNR of 5 (dark blue trace). SBP (light purple trace) was calculated by bandpass filtering, downsampling and taking the signal magnitude of the simulated broadband signal, and the SBP-PIM signals (red/orange traces) were calculated from the SBP signal. Previous work demonstrated that SBP maintains high BMI performance while reducing the wireless data rate from 480 kbps to 32 kbps [27]. In this example, SBP clearly follows the same trend as the true firing rate, exhibiting a correlation of 0.95. The 200 pps SBP-PIM signal has the same 0.95 correlation with true firing rate, showing that it recovers the firing rate information in the SBP signal with an additional order of magnitude reduction in data rate (2.6 vs 32 kbps).

Next, we explored how well SBP-PIM could transmit real neural data at low-data rates. Using neural data previously recorded from NHPs performing a virtual finger task, we varied the packet rate and measured the VAF and correlation and between binned SBP-PIM and binned SBP in 32 ms bins. A VAF or correlation of 1 would imply that SBP-PIM perfectly recovered the original SBP signal, and could thus achieve the same BMI performance. As the PIM packet rate drops, the signal resolution at finer timescales drops, reducing the correlation with SBP. As in figure 4, for packet rates greater than 100 pps the curve is relatively flat, showing little benefit to increasing the data rate above this level. The 100 pps signal has only a small reduction in ability to recover SBP (VAF of 0.92, correlation of 0.96) with a major reduction in data rate (1.3 vs 32 kbps), indicating its potential use as a power-efficient BMI feature. However, at lower packet rates the correlation drops more rapidly (VAF of 0.8 and correlation of 0.9 at 45 pps).

While the previous analyses used 32 ms bins, BMI systems can operate at a range of bin sizes, typically 20–100 ms. Therefore, we also examined SBP vs SBP-PIM correlations for different bin sizes. As seen in figure 4, longer time bins predictably require less-frequent PIM updates (a lower data rate) to achieve the same correlation with SBP; to achieve a 0.95 correlation, 32 ms bins require 70 pps while 100 ms bins require only 20 pps. In general, for relatively accurate recovery of the original signal (e.g. 0.95 correlation), the packet rate should be roughly twice

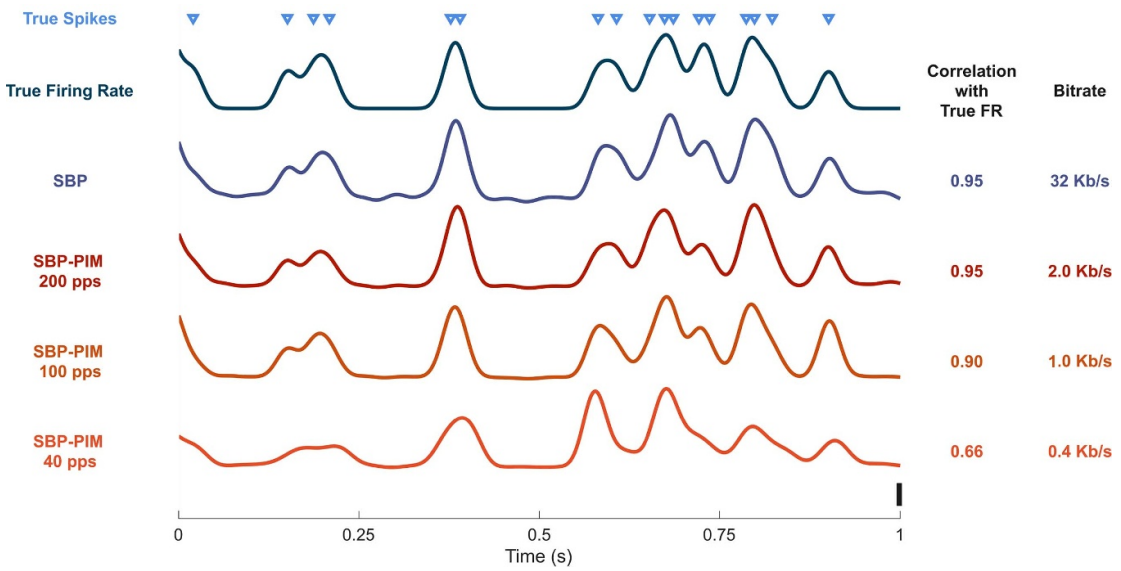

Figure 3. SBP and SBP-PIM correlation with true firing rate. Spikes were simulated and features were smoothed with a 50 ms sliding gaussian window. As the packet rate is dropped, data rate is dropped significantly while correlation with true FR only drops slightly. All signals were normalized to unit amplitude with the black scale bar indicating a normalized amplitude of 1.

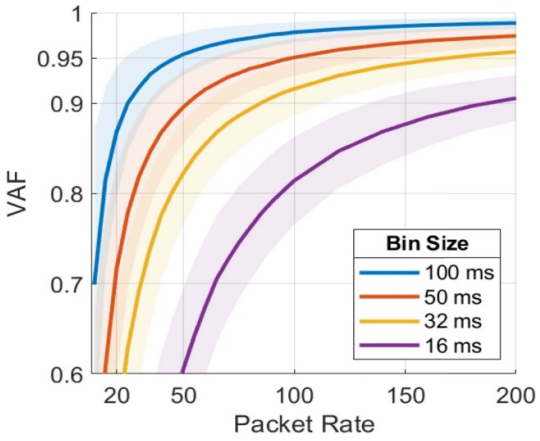

Figure 4. VAF between SBP bins and SBP-PIM bins across packet rates and bin sizes. Data is from real neural recordings from Monkeys N and W performing the finger task.

the bin-rate. This presents a tradeoff between data rate and performance since shorter time bins (on the order of 20 ms) enable significantly higher closed-loop BMI performance [44, 45], but require a higher data rate (and greater power consumption to filter PIM signals).

3.2. Pulse-interval modulation maintains online decoding performance at 100 pps

While offline simulations indicate that low PIM data-rates maintain strong correlations to binned features, it is unclear how much signal fidelity is required to maintain performance of a real-time BMI, which incorporates closed-loop user feedback. To evaluate SBP-PIM's ability to control a real-time closed-loop BMI, Monkeys N and W performed a finger target acquisition task using SBP and SBP-PIM Kalman filter decoders. The SBP-PIM decoder simulated receiving and binning packets in real-time. We started with a PIM packet rate of 100 pps due to its high offline correlation with binned-SBP. During these trials, Monkey N performed a two-DOF task where he had to move each virtual finger to the target position and hold for 500 ms for a successful trial. Figure 5 shows example trials for each decoder; shaded boxes indicate the target positions and solid lines indicate the decoded position for each finger group. At 100 pps, both decoders had a 99% success rate, where SBP had an average acquisition time of 1.36 s and SBP-PIM had an average time of 1.26 s, showing no significant difference *($P > 0.05$; two-tailed two-sample t-test)*. The 100 pps signal represents a bit rate of 1.3 kbps. Thus, despite the small loss of information in the offline comparison, the online decoder maintained performance.

To determine how low the PIM data rate could be dropped without closed-loop performance loss, we varied the packet rate from 30 to 200 pps in separate testing sessions. For the 2-DOF task performed by Monkey N, 100 pps SBP-PIM showed no significant difference in performance compared to SBP on two of three different sessions, as seen by the near zero percent increase in acquisition times in figure 6(a). However, for lower packet rates of 30–80 pps, performance was more than 17% slower with significant differences *($P < 0.05$ for acquisition times; two-tailed two-sample t-test)*. Monkey W, who had weaker neural signals, only performed a one-DOF BMI task, and there were no obvious performance trends within this range (tests with 60–150 pps showed with no significant differences on select days, *$P > 0.05$ for acquisition times; two-tailed two-sample t-test)*.

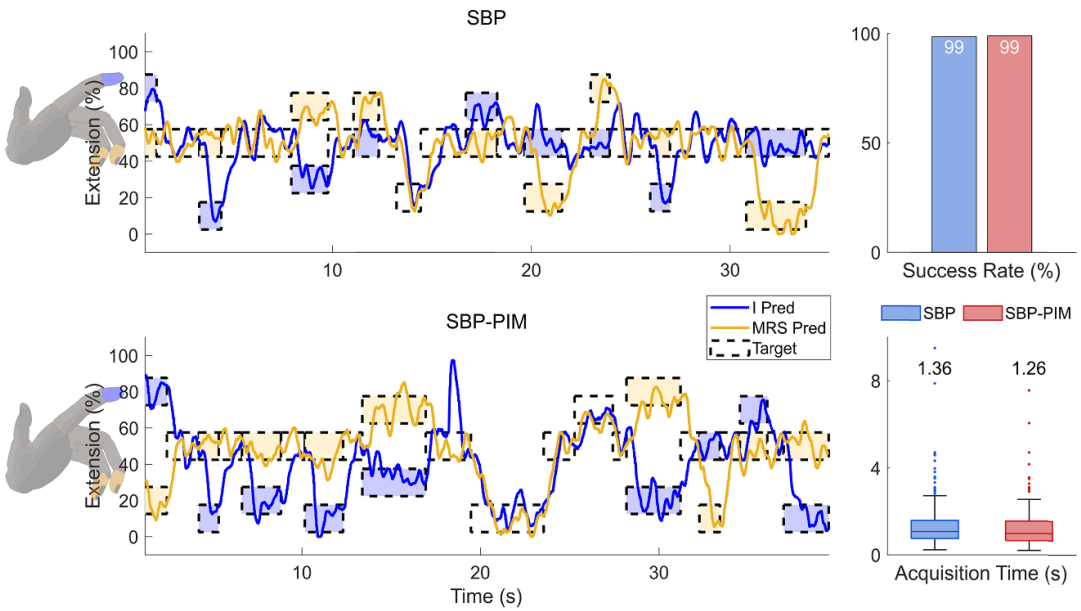

Figure 5. Online two-DOF decoding with Monkey N using SBP and SBP-PIM decoders. Blue and yellow lines indicate the predicted finger position and boxes indicate the target position. No significant differences for success rate or acquisition time were found ($P > 0.05$; two-tailed two-sample t-test).

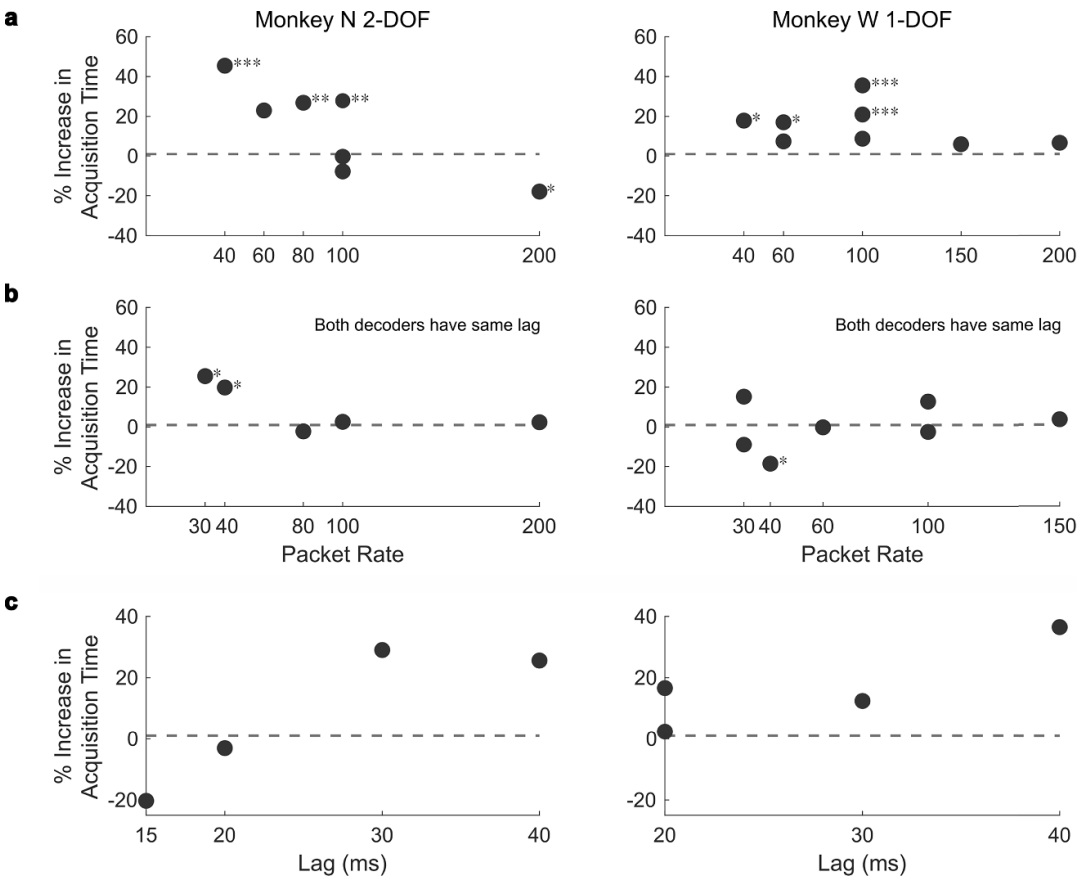

Figure 6. Online SBP-PIM decoder performance. (a) Online PIM-SBP acquisition times as a fraction of SBP acquisition times. In these trials, the PIM-SBP decoder had an additional communication lag not present in the SBP decoder. The dashed 0% line indicates SBP-PIM had the same performance as SBP. Asterisks denote a significant difference from SBP at * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$; two-tailed two-sample t-test. (b) Online PIM-SBP acquisition times as a fraction of SBP acquisition times when both decoders had the same lag. The red 0% line indicates SBP-PIM had the same performance as SBP. Asterisks denote a significant difference from SBP at * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$; two-tailed two-sample t-test. (c) Increase in acquisition times (AT) from additional lag. The change in AT was calculated by taking the difference in AT between decoders with and without the specified lag.

In these initial comparisons, using SBP-PIM features resulted in an extra delay of 15–50 ms (see section 2) before any neural activity would affect the decode due to the asynchronous clocking, whereas the SBP decoder had 0 extra delay. To separate the effects of lag from feature quality, we added an identical lag to the SBP decoder. In these tests, Monkey N achieved near-equivalent performance for 80 pps or higher *(average acquisition times within 3%; no significant difference: $P > 0.05$; two-tailed two-sample t-test)*, and Monkey W, had similar performance across the full range of 30–150 pps (figure 6(b)). Thus, 80 pps PIM might enable equivalent performance if there was not the additional issue of feature lag. This is somewhat surprising, because these results suggest that lags as small as 30 ms have negative performance effects. For example, Monkey N using 80 pps SBP-PIM with a 30 ms lag was 27% slower than SBP, but was 2% faster when lags were equalized (figures 6(a) and (b)). Figure 3(d) summarizes the relationship between increase in acquisition time and lag, with a greater than 12% increase in acquisition times when a 30 ms lag is present.

It is important to note that offline decodes of similar datasets erroneously suggest that lower packet rates could be used without performance loss. An offline velocity Kalman filter using 50 pps SBP-PIM had the same or better VAF compared to an equivalent SBP decoder (supplemental figure 1 (available online at stacks.iop.org/JNE/19/036037/mmedia)). Thus, offline analyses fail to incorporate the effects of lag, time-averaging due to lower packet rates, and closed-loop control.

3.3. Feasibility of communication schemes at 1000 devices

3.3.1. On-device power

An ideal communication scheme would allow for 1000s of motes to communicate with a receiver without bit errors, would stay within system power limits, and would minimize lag. Here, we also considered whether TDMA, a commonly used communication scheme, could be used with IR-powered motes in terms of power consumption. For synchronous communication schemes like TDMA (figure 1(b)), the transmitted data-rate and on-chip clock speed must scale to accommodate greater numbers of motes, requiring increased power. To estimate power consumption, we simulated a ring oscillator clock circuit; at 4 MHz, the clock consumed 1.63 μW, which is above the 1.5 μW limit [26] without accounting for other device modules (ADC, filters, etc) Thus, in this model, TDMA is infeasible for IR-motes. Additionally, TDMA has a lag of 1 time-bin (32 ms) unless a faster clock is used.

PIM, however, is asynchronous and does not necessarily require a faster clock as the number of motes increases. Lim *et al* 2021 previously showed that a full PIM-mote can function with less than 1 μW using an 8 kHz clock.

3.3.2. Receiver feasibility

With PIM communication, additional complexity (and power usage) of 1000-device communication is handed off to the receiver unit. To determine if a receiver could stay within power density limits and accurately receive neural signals, we simulated communication with 1000 PIM motes.

Figure 7(b) shows our proposed receiver design. First, an array of SPADs detect incoming light pulses and each detector output is stored in a local SRAM. SPADs can detect single photons, however, they have a binary output and thus cannot distinguish between multiple motes transmitting simultaneously; with 1000s of motes and wide dispersion of transmitted light (figure 7(a)), packet collisions are likely since each SPAD detects an average of over 100k light pulses per second. Next, to resolve these collisions, for one mote, a temporal-filter matched to the mote's ID is run on the binary outputs of nearby SPADs. If all time samples corresponding to the ID pulse times are high, then the filter output is high. To further reduce the error rate, a spatial-filter combines multiple temporal-filter outputs to reduce packet errors and identify when the mote is transmitting. The receiver simultaneously runs many temporal and spatial filters to recover each mote's packet times.

In simulation, we used real neural data to generate transmitted mote pulses which were probabilistically received at each SPAD according to the light dispersion model shown in figure 7(a). Using SPADs at 300 μm pitch and a 10 MHz sampling rate, after running temporal matched filters each detector had a 19% error rate on average (1% false-positives, 18% false-negatives; sampling rate of 10 MHz; similar to figure 7(c), black traces). Then, the spatial filters recovered the true transmission times with less than 0.5% error rate (0.1% false-positives, 0.3% false-negatives; 32 filters; similar to figure 7(c), blue trace).

An additional offline simulation involving randomly adding packet errors suggests that offline decode correlation depends on the total error rate (false-positives plus false-negatives). This simulation showed that a 1% error rate corresponded to less than 5% increase in MSE and less than 0.01 drop in correlation for velocity-decoding (supplemental figure 2). Thus, for further analyses we aimed to keep the packet error rate below 1%.

Finally, we optimized receiver power consumption by varying SPAD pitch, FOV, and sampling rate. Some tradeoffs emerge when performing this optimization: If the density of SPADs is increased, fewer temporal filters need to be run per SPAD and SPADs are closer to each mote, but the total number of SPADs and SRAM is increased. Another tradeoff is the detector sampling rate (equal to 1/SPAD dead time);

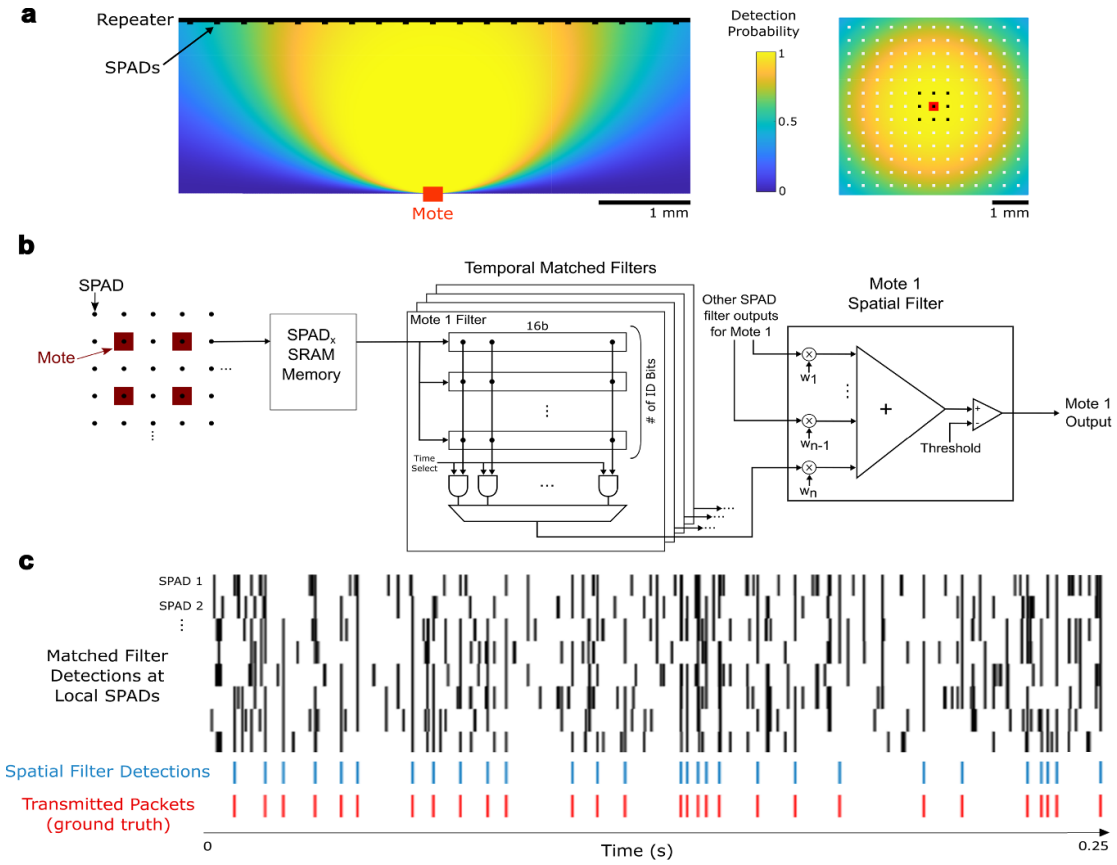

Figure 7. Proposed receiver design and filtering. (a) Probability of IR pulse detection by SPADs. Color indicates the probability of detection. Left—Side view with mote in red (transmitting upwards) and receiver in black. Right—Top-down view with mote in red and SPADs shown as black and white squares. SPADs in black run temporal-filters to 'listen' for the red mote's ID. (b) Receiver filtering pipeline. Binary SPAD samples are stored in a local SRAM and accessed to perform a temporal matched filter for nearby motes. The outputs from multiple temporal filters are combined in a spatial filter and thresholded to produce the output for each mote. (c) An example simulated detection of one mote using temporal and spatial filters. Black traces indicate the temporal-filter detections at SPADs near the mote. The spatial-filter combines the black traces in a weighted sum and threshold operation to reduce erroneous detections (blue traces), and better match the true packet times (red traces).

sampling faster allows for finer temporal resolution and fewer false-positive detections after filtering, but filtering and memory operations must be sped up with higher power usage. Additionally, the SRAM size is proportional to the sampling rate since the mote ID is a fixed time length determined by the mote clock and the receiver must store enough samples to capture the full length of a mote ID. With relatively slow mote clocks, IDs are 2.5 ms in length, requiring an 8 kb SRAM per SPAD when sampling at 2 MHz. Figure 7(b) shows the proposed receiver architecture incorporating SPADs with SRAM memory, followed by temporal and spatial filters.

For each SPAD pitch, FOV, and sampling rate combination, we found the minimum sampling rate and minimum number of temporal filters required to maintain an error rate of less than 1%, and calculated the total receiver power (SPAD power, SRAM power, and filtering power). As seen in figure 8, FOV had the largest impact on the receiver power. At 180° FOV, all motes are visible to each SPAD, requiring 475 mW to resolve packets (500 μm pitch, 10 MHz sampling rate). However, at 45° FOV, only five motes are visible to each SPAD, and receiver power is reduced to 14 mW (600 μm pitch, 2 MHz sampling rate). Across all FOVs, the total power was dominated by the SPADs, accounting for 61%–69% of the total.

Once the FOV is reduced to 30° or less, only one mote on average is visible to each SPAD, significantly reducing packet collisions and removing the need for the spatial filter. With a 30° FOV, the simulated receiver used 2.9 mW (600 μm SPAD pitch, 80 kHz sampling rate) with an error rate of less than 0.1%, for an additional five-fold reduction in receiver power. In this configuration, 66% of power was for SPADs, 33% was for filtering, and 1% was for SRAM operations. Furthermore, narrower FOVs would reduce the effect of vertical mote motion on light dispersion, improving receiver robustness. Narrow FOVs could be achieved through using micro-lenses [46–48] or photonic crystals [49] on the motes or SPADs, or by building small walls around each SPAD to reduce light interference.

In addition to the aforementioned processing power, the receiver also provides downlink power to the motes through IR light. To achieve the required

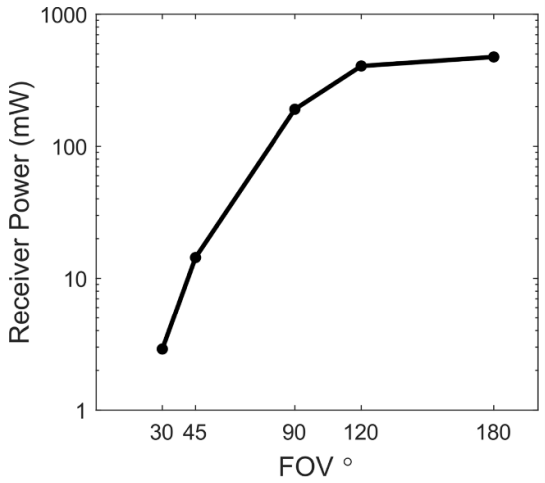

Figure 8. Receiver power consumption for less than 1% packet error rate. Each point represents the optimal parameters across all SPAD pitches and at the given FOV.

0.73 μW of operating power for each mote [22], given a 190 $\times$ 204 μm and 25% efficient photovoltaic cell, and a conservative tissue transmittance of 22%, the receiver needs to produce an illumination intensity of 342 μW mm^{-2}. Given an LED efficiency of 75%, the receiver would use an estimated 456 μW mm^{-2}, or 274 mW for a 2 $\times$ 3 cm craniotomy. Thus, the total receiver power requirement (uplink processing and downlink power) would be approximately 277 mW for a density of 462 μW mm^{-2}, largely dominated by the downlink power. This power density is well below the 1.35 mW mm^{-2} tissue irradiation limit [50], and provides some additional margin for other required circuits.

4. Discussion

Here we evaluated a novel PIM communication scheme that aims to accommodate thousands of wireless IR neural motes while staying within theoretical power constraints. Through offline simulation, we found that asynchronous intervals can efficiently encode SBP using data rates as low as 100 packets per second (1.3 kbps) and maintain strong correlations to both the true underlying firing rate and the SBP. We then simulated the communication scheme in a real-time 96-channel BMI with NHPs and found that despite some information loss at 100 pps, online performance matched the state-of-the-art SBP BMI. While PIM enables reduction of mote power usage, the scheme requires increased power and complexity at the receiver unit. Our simulations of the IR communication channel suggest that communication with 1000s of motes is possible with a total receiver power usage of approximately 3 mW. Thus, an IR PIM mote system could stay within safe power density limits for downlink power [26] and for heating due to receiver processing [51]. We have previously shown successful single-mote communication and signal recovery [23, 26]; future work is required to validate the simulations here with multiple motes and developing a custom receiver ASIC with integrated SPAD array.

Surprisingly, we found communication lags as small as 30 ms to have large effects on online, closed-loop BMI performance. The effect of lag is typically lost during offline decoding, unless closed loop control strategies are considered in decoding [52]. Thus, wireless BMIs must consider the lag associated with the communication scheme in addition to other signal processing delays. For TDMA-based communication, lag is dependent on how quickly all motes can be cycled through; if each cycle is 32 ms, then communication lag is 32 ms. For reduced lag, mote clocks must be sped up with stricter precision requirements and increased power usage. In the PIM scheme, however, lag is dependent on the average packet rate, with higher packet rates reducing lag.

These results suggest that current BMI systems, including wired [9, 16, 28] and wireless [21, 53] systems, could achieve similar or better performance with lower data rates and reduced power consumption. While full-fidelity recordings or spike-template matching approaches may be useful for neuroscience contexts, in BMI, performance can be maintained with significantly lower data rates (1 kbps instead of 480 kbps). Mote integrated-chips could achieve state-of-the-art performance using a simple analog front-end with a communication driver, without the need for specific neural-spike detection circuitry.

Additionally, some information loss can be tolerated during online control (similar to the acceptable offline error rate found in [54]). Further information loss could be acceptable with the use of non-linear neural network decoders, allowing for lower data rates and communication power consumption. With greater neural data compression, wireless communication and power transfer is a significantly smaller limitation to the overall BMI. For BMIs with large numbers of tuned channels, one could further optimize power consumption by reducing the signal quality (reducing the packet rate) of poorly-tuned channels and increase or hold-constant the signal quality of strongly-tuned channels, while maintaining strong decode performance.

In light-based mote systems, efforts can be put toward optimizing the power of SPAD detectors and refining methods for narrowing the optical FOV. For FOV reduction, microlens arrays, which are used in camera and microscope applications [46–48], could be integrated on top of the SPAD array. Alternatively, a thin slab photonic crystal consisting of air holes could be used to both narrow the FOV and improve the extraction efficiency of the LED emitter [55], and is a technique commonly used for commercial surface emitting lasers today. Similarly, a multilayer dielectric

stack photonic crystal could be placed in front of the SPAD array to limit the FOV [49]. For example, Shen *et al* [49] created a photonic crystal by stacking layers of SiO_2 and Ta_2O_5 on a fused silica wafer to achieve an 8° optical FOV, which is significantly more restrictive than our 30° target FOV.

Wireless mote-based BMIs may enable stable recording for longer time periods than conventional wired arrays. Unlike monolithic recording systems, mote recording-channels could fail independently of each other, allowing for a longer system lifetime. With 1000 or more intracortical recording channels distributed across the cortex, in addition to uncovering novel neuroscience data, there may be enough information to decode many degrees of freedom with high accuracy, setting the path toward clinically effective BMI.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

We would like to thank Michael Barrow and Eunseong Moon for assistance in developing the IR light model. We thank Eric Kennedy for animal and experimental support. We thank the University of Michigan Unit for Laboratory Animal Medicine for expert surgical and veterinary care. This work was supported by NSF Grant 1926576, the D Dan and Betty Kahn Foundation Grant AWD011321, NSF Grant 2129817, and NIH Grant 1R21EY029452. J T C was supported by NSF GRFP 1841052. S R N was supported by NIH Grant F31HD098804. M S W was supported by NIH Grant T32NS007222.

Ethics statement

This study was carried out in accordance with the recommendations of Guide for the Care and Use of Animals, Office of Laboratory Animal Welfare and the United States Department of Agriculture Animal and Plant Health Inspection Service. The animal care and monitoring protocol was approved by the Institutional Animal Care and Use Committee at the University of Michigan. The study protocol was approved by the Unit for Laboratory Animal Medicine at the University of Michigan.

ORCID iDs

Joseph T Costello https://orcid.org/0000-0001-7608-0885
Samuel R Nason-Tomaszewski https://orcid.org/0000-0002-7127-0986
Hyochan An https://orcid.org/0000-0002-6322-025X
Jungho Lee https://orcid.org/0000-0003-4162-3389
Matthew J Mender https://orcid.org/0000-0003-1562-3289
Hisham Temmar https://orcid.org/0000-0002-4464-4911
Dylan M Wallace https://orcid.org/0000-0003-2770-3614
Jongyup Lim https://orcid.org/0000-0003-0306-3966
Matthew S Willsey https://orcid.org/0000-0003-2093-7733
Parag G Patil https://orcid.org/0000-0002-2300-6136
Taekwang Jang https://orcid.org/0000-0002-4651-0677
Jamie D Phillips https://orcid.org/0000-0003-2642-3717
Hun-Seok Kim https://orcid.org/0000-0002-6658-5502
David Blaauw https://orcid.org/0000-0001-6744-7075
Cynthia A Chestek https://orcid.org/0000-0002-9671-7051

References

[1] Hochberg L R *et al* 2012 Reach and grasp by people with tetraplegia using a neurally controlled robotic arm *Nature* **485** 372–5
[2] Collinger J L, Wodlinger B, Downey J E, Wang W, Tyler-Kabara E C, Weber D J, McMorland A J, Velliste M, Boninger M L and Schwartz A B 2013 High-performance neuroprosthetic control by an individual with tetraplegia *Lancet* **381** 557–64
[3] Willett F R, Avansino D T, Hochberg L R, Henderson J M and Shenoy K V 2021 High-performance brain-to-text communication via handwriting *Nature* **593** 249–54
[4] Ajiboye A B *et al* 2017 Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration *Lancet* **389** 1821–30
[5] Nordhausen C T, Maynard E M and Normann R A 1996 Single unit recording capabilities of a 100 microelectrode array *Brain Res.* **726** 129–40
[6] Hochberg L R *et al* 2006 Neuronal ensemble control of prosthetic devices by a human with tetraplegia *Nature* **442** 164–71
[7] Chestek C A *et al* 2011 Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex *J. Neural Eng.* **8** 045005
[8] Barrese J C *et al* 2013 Failure mode analysis of silicon-based intracortical microelectrode arrays in non-human primates *J. Neural Eng.* **10** 066014
[9] Sahasrabuddhe K *et al* 2021 The Argo: a high channel count recording system for neural recording *in vivo J. Neural Eng.* **18** 015002
[10] Steinmetz N A *et al* 2021 Neuropixels 2.0: a miniaturized high-density probe for stable, long-term brain recordings *Science* **372** abf4588
[11] Stark E, Drori R and Abeles M 2009 Motor cortical activity related to movement kinematics exhibits local spatial organization *Cortex* **45** 418–31
[12] Gallego J A, Makin T R and McDougle S D 2022 Going beyond primary motor cortex to improve brain–computer interfaces *Trends Neurosci.* **45** 1–8

[13] Xie C, Liu J, Fu T M, Dai X, Zhou W and Lieber C M 2015 Three-dimensional macroporous nanoelectronic networks as minimally invasive brain probes *Nat. Mater.* **14** 1286–92
[14] Chiang C H *et al* 2020 Development of a neural interface for high-definition, long-term recording in rodents and nonhuman primates *Sci. Transl. Med.* **12** 1–13
[15] Scholten K, Larson C E, Xu H, Song D and Meng E 2020 A 512-channel multi-layer polymer-based neural probe array *J. Microelectromech. Syst.* **29** 1054–8
[16] Musk E 2019 An integrated brain-machine interface platform with thousands of channels *J. Med. Internet Res.* **21** 0–11
[17] Hanson T L, Diaz-Botia C A, Kharazia V, Maharbiz M M and Sabes P N 2019 The 'sewing machine' for minimally invasive neural recording *bioRxiv* p 578542 (available at: www.biorxiv.org/content/10.1101/578542v1%0Awww.biorxiv.org/content/10.1101/578542v1.abstract)
[18] Seo D, Carmena J M, Rabaey J M, Alon E and Maharbiz M M 2013 Neural dust: an ultrasonic, low power solution for chronic brain-machine interfaces http://arxiv.org/abs/1307.2196
[19] Seo D, Carmena J M, Rabaey J M, Maharbiz M M and Alon E 2015 Model validation of untethered, ultrasonic neural dust motes for cortical recording *J. Neurosci. Methods* **244** 114–22
[20] Ghanbari M M *et al* 2019 A Sub-mm^3 ultrasonic free-floating implant for multi-mote neural recording *IEEE J. Solid State Circuits* **54** 3017–30
[21] Lee J *et al* 2021 Neural recording and stimulation using wireless networks of microimplants *Nat. Electron.* **4** 604–14
[22] Lim J *et al* 2022 A light-tolerant wireless neural recording IC for motor prediction with near-infrared-based power and data telemetry *IEEE J. Solid State Circuits* **57** 1061–74
[23] Lim J *et al* 2020 A 0.19 × 0.17 mm^2 wireless neural recording IC for motor prediction with near-infrared based power and data telemetry *2020 IEEE Int. Solid- State Circuits Conf.* vol 24 pp 470–1
[24] Lee S, Cortese A J, Gandhi A P, Agger E R, McEuen P L and Molnar A C 2018 A 250 μm × 57 μm microscale opto-electronically transduced electrodes (MOTEs) for neural recording *IEEE Trans. Biomed. Circuits Syst.* **12** 1256–66
[25] Singer A and Robinson J T 2021 Wireless power delivery techniques for miniature implantable bioelectronics *Adv. Healthcare Mater.* **10** 2100664
[26] Moon E *et al* 2021 Bridging the 'Last Millimeter' gap of brain-machine interfaces via near-infrared wireless power transfer and data communications *ACS Photonics* **8** 1430–8
[27] Nason S R *et al* 2020 A low-power band of neuronal spiking activity dominated by local single units improves the performance of brain–machine interfaces *Nat. Biomed. Eng.* **4** 973–83
[28] Yoon D Y, Pinto S, Chung S W, Merolla P, Koh T W and Seo D 2021 A 1024-channel simultaneous recording neural SoC with stimulation and real-time spike detection *IEEE Symp. VLSI Circuits, Dig. Tech. Pap.* vol 2021 June pp 2020–1
[29] Nason S R *et al* 2021 Real-time linear prediction of simultaneous and independent movements of two finger groups using an intracortical brain-machine interface *Neuron* **109** 3164–3177.e8
[30] Irwin Z T *et al* 2017 Neural control of finger movement via intracortical brain-machine interface *J. Neural Eng.* **14** 066004
[31] Vaskov A K *et al* 2018 Cortical decoding of individual finger group motions using ReFIT Kalman filter *Front. Neurosci.* **12** 00751
[32] Hemmati H, Biswas A and Djordjevic I B 2011 Deep-space optical communications: future perspectives and applications *Proc. IEEE* **99** 2020–39
[33] Lempka S F, Johnson M D, Moffitt M A, Otto K J, Kipke D R and McIntyre C C 2011 Theoretical analysis of intracortical microelectrode recordings *J. Neural Eng.* **8** 045006
[34] Lim W, Jang T, Lee I, Kim H S, Sylvester D and Blaauw D 2016 A 380 pW dual mode optical wake-up receiver with ambient noise cancellation *IEEE Symp. VLSI Circuits, Dig. Tech. Pap.* vol 2016 *(September)* pp 5–6
[35] Haas H, Chen C and O'Brien D 2017 A guide to wireless networking by light *Prog. Quantum Electron.* **55** 88–111
[36] Niclass C and Charbon E 2005 A single photon detector array with 64 × 64 resolution and millimetric depth accuracy for 3D imaging *Dig. Tech. Pap.—IEEE Int. Solid-State Circuits Conf.* vol 48 pp 364–6
[37] Perenzoni M, Perenzoni D and Stoppa D 2017 A 64 × 64-pixels digital silicon photomultiplier direct TOF sensor with 100-MPhotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing *IEEE J. Solid State Circuits* **52** 151–60
[38] Al Abbas T, Dutton N A W, Pellegrini S, Rae B and Henderson R K 2017 8.25 μm pitch 66% fill factor global shared well SPAD image sensor in 40 nm CMOS FSI technology *2017 Int. Image Sens. Work.* pp 97–100 (available at: www.imagesensors.org/)
[39] Ceccarelli F, Acconcia G, Gulinatti A, Ghioni M and Rech I 2019 Fully integrated active quenching circuit driving custom-technology SPADs With 6.2-ns dead time *IEEE Photonics Technol. Lett.* **31** 102–5
[40] Eisele A and Henderson R 2011 185 MHz count rate, 139 dB dynamic range single-photon avalanche diode with active quenching circuit in 130 nm CMOS technology *Proc. Int. Image Sensor Workshop* pp 6–8 (available at: www.imagesensors.org/PastWorkshops/2011Workshop/2011Papers/R43_Eisele_SPAD139dB.pdf.)
[41] Burri S, Maruyama Y, Michalet X, Regazzoni F, Bruschini C and Charbon E 2014 Architecture and applications of a high resolution gated SPAD image sensor *Opt. Express* **22** 17573
[42] Niclass C and Soga M 2010 A miniature actively recharged single-photon detector free of afterpulsing effects with 6 ns dead time in a 0.18 μm CMOS technology *Tech. Dig.—Int. Electron Devices Meet. IEDM* pp 340–3
[43] Ghioni M, Gulinatti A, Rech I, Zappa F and Cova S 2007 Progress in silicon single-photon avalanche diodes *IEEE J. Sel. Top. Quantum Electron.* **13** 852–62
[44] Pandarinath C *et al* 2017 High performance communication by people with paralysis using an intracortical brain-computer interface *Elife* **6** e18554
[45] Cunningham J P, Gilja V, Ryu S I and Shenoy K V 2009 Methods for estimating neural firing rates, and their application to brain-machine interfaces *Neural Netw.* **22** 1235–46
[46] Yuan W, Li L H, Lee W B and Chan C Y 2018 Fabrication of microlens array and its application: a review *Chin. J. Mech. Eng.* **31** 16
[47] Liu X Q *et al* 2019 Optical nanofabrication of concave microlens arrays *Laser Photonics Rev.* **13** 1800272
[48] Wei Y *et al* 2018 Fabrication of high integrated microlens arrays on a glass substrate for 3D micro-optical systems *Appl. Surf. Sci.* **457** 1202–7
[49] Shen Y, Ye D, Celanovic I, Johnson S G, Joannopoulos J D and Soljačić M 2014 Optical broadband angular selectivity *Science* **343** 1499–501
[50] American National Standard Institute 2014 American National Standard for safe use of lasers (ANSI Z136.1–2014) (American National Standards Institute, Inc)
[51] Kim S, Tathireddy P, Normann R A and Solzbacher F 2007 Thermal impact of an active 3D microelectrode array implanted in the brain *IEEE Trans. Neural Syst. Rehabil. Eng.* **15** 493–501

[52] Willett F R et al 2017 Feedback control policies employed by people using intracortical brain-computer interfaces J. Neural Eng. 14 016001
[53] Simeral J D et al 2021 Home use of a percutaneous wireless intracortical brain-computer interface by individuals with tetraplegia IEEE Trans. Biomed. Eng. 68 2313–25
[54] Even-Chen N et al 2020 Power-saving design opportunities for wireless intracortical brain–computer interfaces Nat. Biomed. Eng. 4 984–96
[55] Fan S, Villeneuve P R, Joannopoulos J D and Schubert E F 1997 High extraction efficiency of spontaneous emission from slabs of photonic crystals Phys. Rev. Lett. 78 3294–7