Statistically Reconstructed Multiplexing for Very Dense, High-Channel-Count Acquisition Systems

David Tsai, *Member, IEEE*, Rafael Yuste, and Kenneth L. Shepard, *Fellow, IEEE*

***Abstract*—Multiplexing is an important strategy in multichannel acquisition systems. The per-channel antialiasing filters needed in the traditional multiplexing architecture limit its scalability for applications requiring high channel density, high channel count, and low noise. A particularly challenging example is multielectrode arrays for recording from neural systems. We show that conventional approaches must tradeoff recording density and noise performance, at a scale far from the ideal goal of one-to-one mapping between neurons and sensors. We present a multiplexing architecture without per-channel antialiasing filters. The sparsely sampled data are recovered through a compressed sensing strategy, involving statistical reconstruction and removal of the undersampled thermal noise. In doing so, we replace large analog components with digital signal processing blocks, which are much more amenable to scaled CMOS implementation. The resulting statistically reconstructed multiplexing architecture recovers input signals at significantly improved signal-to-noise ratios when compared to conventional multiplexing with antialiasing filters at the same per-channel area. We implement the new architecture in a 65 536-channel neural recording system and show that it is able to recover signals with performance comparable to conventional high-performance, single-channel systems, despite a more than four-orders-of-magnitude increase in channel density.**

***Index Terms*—Electrophysiology, interpolation, multielectrode array, multiplexing, sampling.**

I. Introduction

IN TODAY'S big-data systems there are often needs to acquire information from a large number of signal sources within a short period of time. Useful information embedded within these signals are extracted (for example, by band-pass filtering), sampled, and digitized for later retrieval and analyses. Multielectrode electrophysiological recording tools [1] in neuroscience are an important example of such a system, containing an array of electrodes for capturing signals emitted by neurons in brain circuits. In most situations there are orders of magnitude more neurons than electrodes. For instance, there are more than 1 million neurons in the human retina [2], the light-sensing neural tissue at the back of each eye; and there are more than 250 million neurons in the primary visual cortex [3], the brain region devoted to visual processing. In contrast, electrophysiological tools in routine use today contain at most a few hundred electrodes (e.g. [4], [5]); each at best is able to reliably pick up signals from a handful of nearby neurons – on the scale of $\leq 40\,\mu$m [6]. There is, therefore, a desire to increase electrode count, to record from as many neurons as practically possible.

The implementations of these electronic systems are constrained by many factors. For electrophysiology these include: space limitation, heat emission, power budget, sensor-to-signal-source proximity and invasiveness to the biological specimen. Consequently, multiplexing is fundamental to most neural signal acquisition systems (e.g. [7]–[10]). This allows numerous front-end recording pathways, including the electrodes, to share a single back-end. The signals picked up by each electrode are typically very small. Depending on the recording modality, the signal peak-to-peak amplitude is on the order of 10 s of microvolts. Amplification with low noise amplifiers, close to the source, is crucial for preserving signal integrity. These circuits add additional burden on the system design constraints.

We start by outlining the traditional architecture for multiplexed data acquisition, with particular emphasis on its scalability limitations for high-channel-count (greater than a few hundred) applications in Section II. As an approach to tackle this major obstacle, we present a new multiplexed, sampling architecture tailored for ultra-high-channel-count scale up in CMOS-based technologies, referred to as the statistically reconstructed multiplexing architecture (SRMA), in Sections III and IV. In Section V, we then demonstrate an implementation of this architecture in a 65,536-channel neural recording system, followed by concluding remarks in Section VI.

II. The Traditional Approach

In general, we can represent multiplexed signal acquisition systems as shown in Fig. 1(a). The inputs are M continuous-time signals $f_i(t)$, $i \in \{0 \ldots M-1\}$. These signals are first amplified by a preamplifier in each channel. The analog signal for channel i at time t is sampled by the sample-and-hold (SH), amplified further, then digitized by an analog-to-digital converter (ADC) operating at sampling frequency $\geq 2M f_{bw}$,

Manuscript received February 1, 2017; revised April 26, 2017; accepted July 22, 2017. This work was supported in part by the U. S. Army Research Laboratory and the U. S. Army Research Office under Contract W911NF-12-1-0594, by the Defense Advanced Research Project Agency (DARPA) under Contract N66001-17-C-4002, by the National Institutes of Health under Grants U01NS0099697 and U01NS0099726, and by the National Health and Medical Research Council of Australia CJ Martin Fellowship (APP1054058). This paper was recommended by Associate Editor R. Genov. *(Corresponding author: David Tsai.)*

D. Tsai and K. L. Shepard are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: dtsai@ee.columbia.edu; shepard@ee.columbia.edu).

R. Yuste is with the Department of Biological Sciences, Columbia University, New York, NY 10027 USA (e-mail: rmy5@columbia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBCAS.2017.2750484

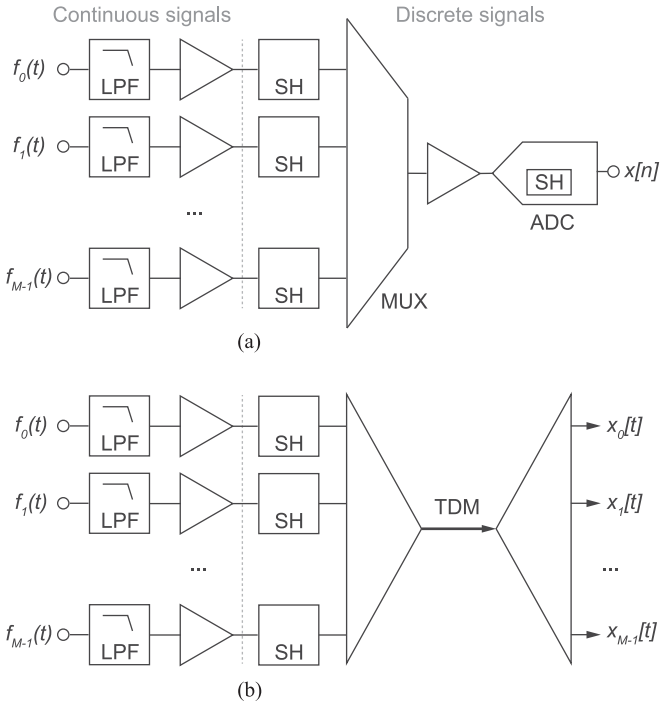

Fig. 1. Conventional multiplexing architecture. (a) Multiplexing with a sample-and-hold element (SH) within each channel requires a low-pass antialiasing filter (LPF) per channel. MUX, multiplexer. ADC, analog-to-digital converter. (b) Conventional multiplexing is equivalent to time division multiplexing (TDM).

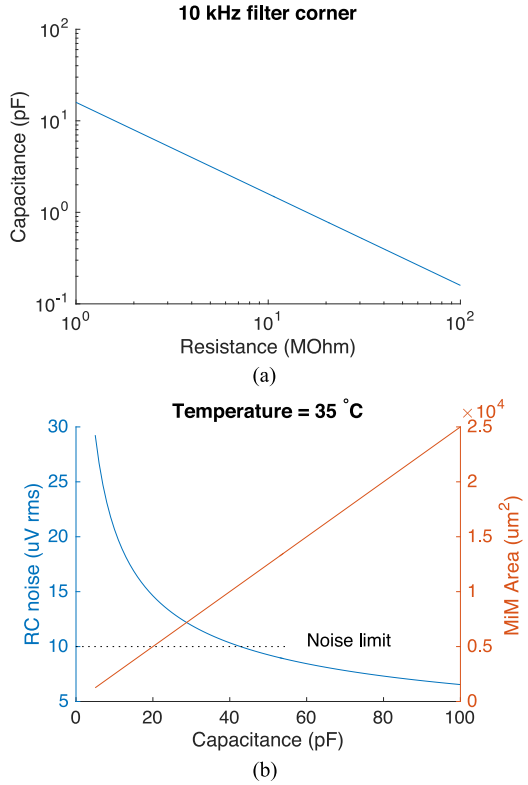

Fig. 2. Low-pass antialiasing filter for conventional multiplexing. (a) Scaling of capacitance and resistance for a 1-pole 10 kHz low-pass filter. (b) Filter noise and area requirements of metal-insulator-metal (MiM) capacitors in typical CMOS microelectronics. "Noise limit" denotes the noise level that should not be exceeded for extracellular electrophysiological recordings.

where f_{bw} is the bandwidth of each input. To prevent aliasing artifacts, each channel must have a low-pass filter (LPF) preceding the SH, with corner frequency f_{bw}. The toggling of the SH, multiplexing (MUX) addressing, and ADC clock are appropriately timed, such that the input to the ADC is sufficiently stabilized for quantization. The ADC output $x[n]$ then contains a sequence of digitized data, organized by MUX addressing and time, in this order. Fig. 1(b) shows how these systems can also be viewed as independent signals time-division multiplexed through a common communication channel (i.e. back-end amplifier and ADC), by means of synchronized switches (i.e. SH toggles, MUX address updates and ADC clocking).

The conventional architecture of Fig. 1(a), however, is ill-suited for form-factor-constrained scale-up to thousands of recording channels due to the size of the required antialiasing low-pass filters at the beginning of each signal path. In the simplest form, these filters are first-order networks with a low-pass corner frequency of

$$f_{bw} = \frac{1}{2\pi RC}$$

The frequency of neural signals typically spans dc to approximately 3 kHz, depending on the measurement technique and biological processes under investigation. To realize such a corner frequency, either a large capacitor or a high-valued resistor is needed (Fig. 2(a)). High-resistance pseudo-resistors can be constructed from MOSFETs operating in weak inversion; therefore achieving high R is generally not a problem for applications with small-amplitude inputs. However, input-impedance and thermal noise considerations favor capacitance over resistance. First, a large resistor in series with the source also forms a voltage divider between the signal and the input amplifier, attenuating weak signals and compromising SNR. Second, the thermal noise of this RC network is described as a mean-squared voltage of

$$\overline{V_{noise}^2} = \frac{kT}{C}$$

where k is the Boltzmann constant and T the temperature in Kelvin. To sense signals of a few 10 s of μV peak-to-peak, the network's noise should be no greater than approximately 10 μV rms over the dc to 10 kHz bandwidth. This requires a capacitor in the range of 40 pF at 35 °C (Fig. 2(b)).

High density capacitors in today's commercial microelectronic processes provide approximately 4 fF/μm^2 [11], using metal-insulator-metal (MiM) capacitors. Capacitors equal to, or greater than, 40 pF would require $\geq$10,000 μm^2. To put this on the biological scale, the soma (the neuronal cell body, where recordings are usually made), has typical diameter $\leq$25 μm. A MiM capacitor with adequate noise performance would occupy an area 20 times greater than the neuron from which it senses, a severe limitation on the goal of achieving high-density, high-channel-count recordings in the central nervous system [12].

Additional circuit elements further complicate efforts to increase channel density. In many applications it is desirable to AC couple the biological preparation from the recording circuits, such that neurons are not exposed to the amplifiers'

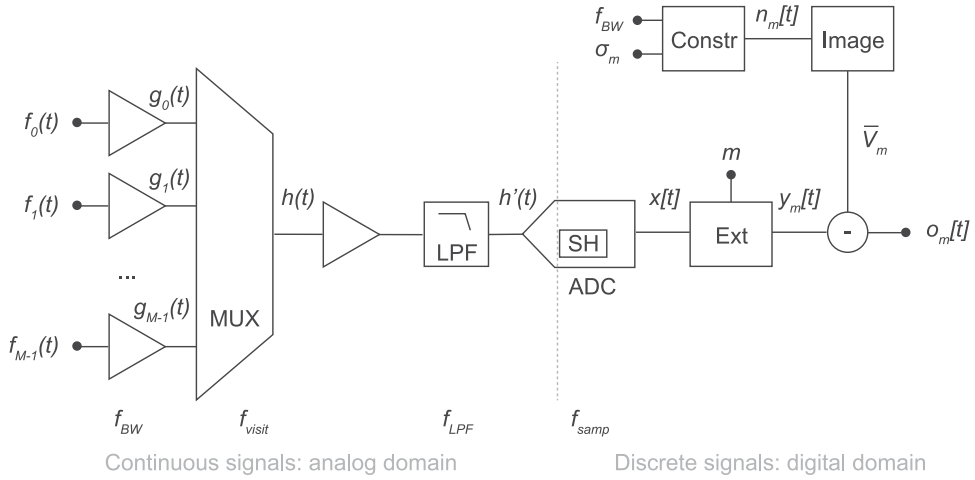

Fig. 3. Overview of the statistically reconstructed multiplexing architecture (SRMA). The input signal $f_i(t)$ from all channels are treated as continuous-time up to the ADC, obviating the need for per-channel antialiasing filters. Extracting the per-channel data from the ADC output causes under-sampling of thermal noise. The spectral contributions of this under-sampled noise are computed and removed from the channel data, in the frequency space, producing aliasing-free outputs.

transistor biasing voltage. This high-pass filter needs a 3-dB corner frequency as low as a few Hz, requiring an additional large capacitor at the beginning of the signal pathway. Finally, noise considerations, in particular 1/f noise, limits minimum transistor sizes [13].

Existing implementations of neural recording systems have, therefore, been limited to low-noise (<5 μV rms in passband), but low-density arrays (approximately 126 simultaneously-recording channels per mm^2) [14], [15] or high-density (approximately 4225 simultaneously-recording channels permm^2), but high-noise (20 to >100 μV rms) arrays [7], [8], [10].

III. A Scalable Multiplexing Architecture

We present a new multiplexed data acquisition architecture that enables high-density, high-channel-count scale up without incurring the noise-to-density trade-offs of traditional multiplexing approaches. The statistically reconstructed multiplexing architecture (SRMA) is illustrated in Fig. 3.

A. Overview of the Architecture

The acquisition process begins by amplifying each signal at the input. A MUX is used to direct several first-level amplifiers to a second, shared amplifier. The output of the MUX, therefore, consists of continuous-time segments of the channels that have been selected. In the traditional approach to such multiplexing (Fig. 1), the sampling operation is placed within each channel, prior to the multiplexer. This necessitates a LPF prior to the SH. As noted previously, space and noise considerations for these analog filters limit system scalability.

Our approach, in contrast, treats all inputs $f_i(t)$ as continuous-time analog signals up to the ADC, where they are sampled and digitized. An antialiasing filter placed before the ADC rejects signals above half the ADC's sampling rate. The digital signal processing of this data stream begins by extracting the per-channel data from the ADC's data stream. As we will describe later in detail, this operation, in conjunction with the lack of per-channel antialiasing filter, causes the channel data to be sampled at a rate substantially lower than the systems' bandwidth. Unless reconstructed and removed from the per-channel data, frequencies between half the per-channel visit rate, by the multiplexer, and the system's bandwidth would be aliased. To prevent this aliasing, we use a compressed sensing strategy to reconstruct and remove the spectral contributions of these under-sampled frequencies from the per-channel data. This is made possible through careful choices of the key system components' operating frequencies (Fig. 3); in particular, f_{visit} (visit rate for a particular channel by the MUX), f_{LPF} (ADC antialiasing low-pass frequency), f_{samp} (ADC sampling rate) and f_{BW} (signal chain bandwidth). We now consider each of these steps in more detail.

B. Signal Multiplexing

When a particular channel m is selected by the MUX, the output $h(t)$ changes continuously according to the pre-amplified $g_m(t)$. Formally, we can view the MUX as mapping continuous time T and channel address A to continuous signal S, which is the voltage at the input channel:

$$h : (T, A) \rightarrow S,\ T \in \mathbb{R}, A \in \{1 \ldots M-1\} \text{ and } S \in \mathbb{R}$$

Importantly, the MUX output is defined for all time $t \in \mathbb{R}$. The parameter A represents the selection address, determined from the set $\{0 \ldots M-1\} \subset \mathbb{N}$, where M is the number of channels in the system.

It is instructive to contrast the inputs g_m and output h of the MUX to those of a sample-and-hold element. In the latter, the input a is a mapping from continuous-time T to signal S, but the output b maps from discrete-time T to signal S. That is, the output is a sequence of Dirac impulses drawn from the input

$$a : T \rightarrow S,\ T \in \mathbb{R} \text{ and } S \in \mathbb{R}$$

$$b : T \rightarrow S,\ T \in \mathbb{N} \text{ and } S \in \mathbb{R}$$

If we operate an ideal (zero delay and infinite bandwidth) MUX such that the per-channel dwell time is δ and we

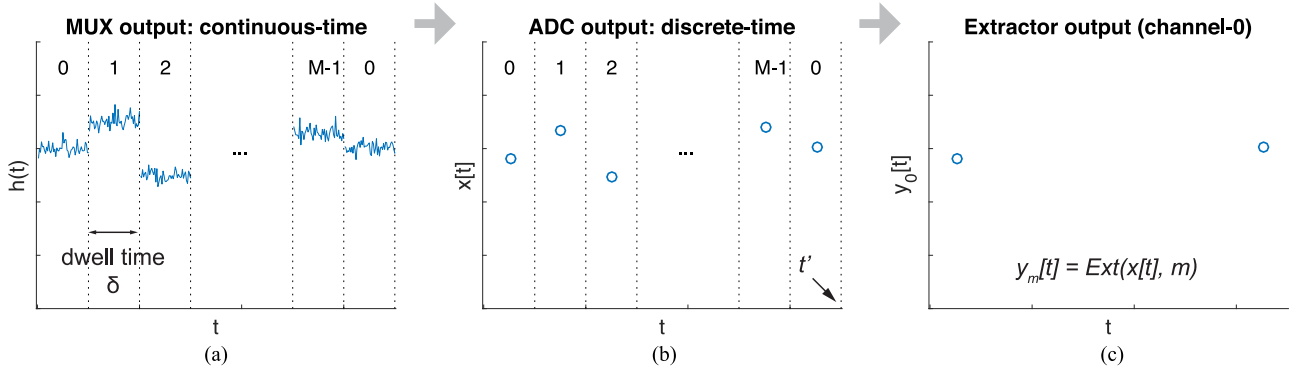

Fig. 4. Extracting channel data from the ADC's output stream. (a) Multiplexing of the analog signals. (b) Periodic sampling of inputs. (c) Extracting per-channel samples.

repeatedly cycle through all inputs sequentially, the output of the analog multiplexer $h(t)$ will be concatenated segments of the continuous signals from the scanned, post-amplified inputs $g_i(t)$, $i \in \{0, \ldots, M-1\}$. Each segment is offset in time with duration δx, $x \in \mathbb{N}^*$, with $\mathbb{N}^*$ denoting the set of natural numbers greater than 0. This is illustrated in Fig. 4(a).

For simplicity we have assumed an ordered cycling set C of channels from lowest to highest address $C = \{0, 1, \ldots, M-1\}$, with equal per-channel dwell time δ. In general, this need not be the case. For example, the scanning sequence can be limited to just the subset of inputs with interesting signals. Furthermore, one or more of the inputs could be scanned multiple times per cycle, if the associated channels contain signal of higher bandwidth than the rest.

C. Digitization

The continuous-time signal $h(t)$ from the MUX is amplified and sampled by the ADC at frequency f_{samp}. A pre-ADC low-pass filter prevents aliasing of contents above half the sampling rate; therefore, it has corner frequency $f_{LPF} = f_{samp}/2$.

It is required that $f_{samp} \geq n f_{visit}, n \in \mathbb{N}^*$, where f_{visit} is the per-channel visit rate by the MUX. This has two effects. First, it ensures at least one conversion by the ADC per MUX address change. Second, this guarantees that MUX address changes are in phase with the ADC conversion, because f_{samp} is divisible by f_{visit}. It is further required that the phase differences, if any, between the MUX address lines and the ADC clock be kept constant to minimize timing jitters in the ADC's data conversion aperture time t_{apt}. It should also be apparent that $t_{apt} < 1/f_{samp}$.

The ADC's output $x[t]$ consists of discrete, time-indexed samples from $h'(t)$, the antialias-filtered version of $h(t)$. Changing the multiplexer address while sampling causes the ADC output $x[t]$ to contain samples from all scanned channels, ordered by the multiplexer's addressing history (Fig. 4(b)).

D. Extracting Single-Channel Data

At some arbitrary time t' (Fig. 4(b)), if we have been keeping a history of the MUX switch positions and the dwell time δ at each position, we can recover data segments for each scanned input from the ADC's output. Given a particular channel m, the canonical ADC output $x[t]$ is processed through an extractor Ext (Fig. 3) to produce a new sequence $y_m[t]$, which is defined as $y : \mathbb{N} \rightarrow \mathbb{R}$

$$y_m[t] = x[t]|_{Addr_{MUX}(t)=m}$$

For example, when the MUX is cycled through the address sequence $\{0, 1, \ldots, M-1\}$, $y_m[t]$ would consists of sampled data from $x[t]$ whenever the MUX is switched to channel m (Fig. 4(c)).

The channel-data extraction procedure Ext creates an output with sampling rate f_{visit}, where $f_{visit} < f_{samp}$ when the number of scanned channels M is >1. This reduced sampling rate, f_{visit}, relative to the ADC rate of f_{samp}, creates two considerations. First, we need to ensure that f_{visit} is sufficient to describe the signal of interest. Second, due to the lack of per-channel, anti-aliasing filters and the large number of channels visited by the MUX in high-channel-count implementations, $f_{visit}/2$ will be, in general, significantly smaller than the bandwidth f_{BW} of each recording channel. Specifically, we expect $f_{visit}/2 \ll f_{BW} \leq f_{samp}/2$. Under such a condition, the content spanning $f_{visit}/2$ to f_{BW} would be under-sampled, thus aliased into the range dc to $f_{visit}/2$.

E. Preserving Signal Bandwidth

We begin by examining the first consideration, that of preserving the signal of interest in the extracted, per-channel data $y_m[t]$, for channel m. In general, we can express the input of a data acquisition system as $f(t) = s(t) + n(t)$, where $s(t)$ represents the signal of interest and $n(t)$ denotes the input-referred noise. If $s(t)$ is bandlimited to the frequency range $(\omega_0 \ldots \omega_0 + \omega)$, the function is completely determined by its values at a set of points with density 2ω [16], [17]. Hence knowing the bandwidth of our signal $s(t)$, we can choose the MUX per-channel visit rate and the ADC sampling frequency appropriately, so that $s(t)$ is completely described by the extracted channel data $y_m[t]$.

Specifically, when multiplexing n channels of a M-channel system, with $n \leq M$, the following conditions must be met to ensure sufficient sampling rate for each signal of interest $s_x(t), x \in \{1 \ldots M-1\}$:

1) The MUX per-channel visit rate, f_{visit}, must be set to $0 < n f_{visit} \leq f_{samp}$, such that the ADC captures at least

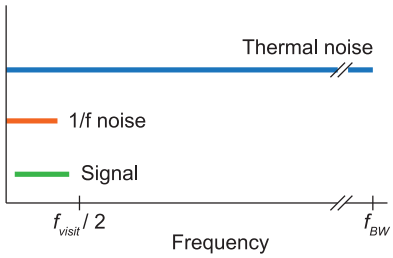

Fig. 5. Spectral range of signal and noise types in electrophysiology.

one complete sample from a channel per conversion period $1/f_{samp}$ for the set of n non-zero input channels.

2) The ADC sampling rate, f_{samp}, must be set to $\geq 2n\omega$, such that the ADC is able to capture the bandwidth of n input channels, each having a signal of interest $s_x(t)$ with bandwidth ω.

F. Constructing Thermal Noise

We now examine the second consideration: under-sampling of the system's bandwidth by the extracted data stream $y_m[t]$, for channel m. The channel input $f_m(t)$ contains neural signal and input-referred noise, as noted previously. The neural signal is completely described by the extractor output $y_m[t]$, which has sampling rate f_{visit}, with $f_{visit} \geq 2\omega$ by virtue of the requirements specified in the previous section. There are two dominant noise types in electrophysiology, 1/f noise and thermal noise. The typical spectral range of the signal and noise types are illustrated in Fig. 5.

1/f noise arises from the transistors. It is particularly prominent in systems with high recording density, where small transistors are used. It is a non-stationary process. The power of this noise decreases with increasing frequency, with a typical corner of a few kHz in CMOS transistors [13]. To properly capture 1/f noise, it is required that $f_{visit}/2$ be greater than the system's 1/f noise corner.

Thermal noise arises from the recording electrodes and the electronics. It has uniform spectral power and its bandwidth is limited by the recording system's bandwidth f_{BW}. Because the half-Nyquist rate of $y_m[t]$, $f_{visit}/2$, is less than f_{BW}, we need to remove the spectral contribution of the under-sampled thermal noise spanning $f_{visit}/2$ to f_{BW} in $y_m[t]$, to prevent aliasing. This is achieved through a compressed sensing strategy [18], where, if one has specific *a priori* knowledge about a signal, it is possible to recover the signal with fewer samples than required by classical Nyquist-Shannon sampling theorem.

Several statistical and spectral characteristics of thermal noise make it possible to reconstruct the effects of aliasing this noise type. It is a stationary random process, with a flat spectrum and a Gaussian time-domain amplitude distribution [19] of zero mean and variance σ^2. The probability density function for such a process is

$$\mathcal{N}(x|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

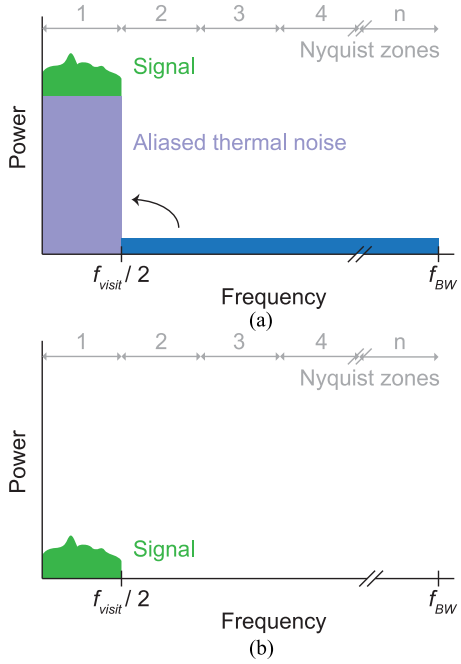

Fig. 6. Removing aliased thermal noise. (a) Sampling at a rate of f_{visit} in a system with bandwidth f_{BW} causes aliasing of contents between $f_{visit}/2$ and f_{BW}, spanning Nyquist zones 2 to n, into the first Nyquist zone. (b) SRMA minimizes the effects of aliasing by calculating the power of these aliased contents in Nyquist zones >1, then removing them from zone 1.

We can easily determine every channel's σ_m^2 for thermal noise calculation by recording each channel without multiplexer interruption (i.e. conventional sampling) at full system bandwidth, thereby completely specifying the channel's thermal noise characteristics up to f_{BW}.

With the thermal noise variance σ_m^2 and bandwidth f_{BW} known for every recording channel, we computationally construct the thermal noise $n_m[t]$ of each multiplexed channel, using the operator $Constr$ (Fig. 3).

G. Removing the Spectral Contribution of Under-Sampled Thermal Noise

Aliasing confers several averaging properties, which greatly simplify the reconstruction, and ultimately the removal, of aliased thermal noise.

First, the power of thermal noise (of infinite length) is uniform across frequencies. Any departure from this ideal, due to the acquired signal's finite length, is averaged out by aliasing, as the contents are folded down into the first Nyquist zone (Fig. 6(a)). Therefore, we can estimate the power contributed by thermal noise aliasing in the under-sampled data, by computing and using the average thermal noise power.

Second, the Fourier space vector angles for thermal noise (of infinite length) has a uniform distribution with zero mean. Again, any departure from this ideal in finite-length signals is averaged out by aliasing, when the contents are folded down into the first Nyquist zone (Fig. 6(a)), causing the angles to converge to zero.

Taking advantage of these properties, we can construct vectors in the frequency space to represent the aliased thermal noise, subtracting these from the aliased data, thereby reversing the effects of aliasing.

The effects of thermal noise aliasing, between $f_{visit}/2$ and f_{BW} can be readily reproduced by decimating the constructed thermal noise $n_m[t]$, for channel m, to a lower rate, f_{visit} (Fig. 6(a)). We denote this aliased sequence a_m:

$$a_m : \mathbf{decimate}(n_m, f_{visit})$$

Next we construct another sequence b_m, a decimated version of $n_m[t]$ without aliasing. This is accomplished by first low-pass filtering $n_m[t]$ at $f_{visit}/2$, followed by decimation to the new rate f_{visit}:

$$b_m : \mathbf{decimate}(\mathbf{lowpass}(n_m, f_{visit}/2), f_{visit})$$

The power contributed by the aliased thermal noise at each frequency, for a system with bandwidth f_{BW} but sampled at only f_{visit}, is, therefore, the difference between the deliberately aliased sequence a_m and the anti-aliased sequence b_m:

$$\rho_m = |\mathcal{F}(a_m)| - |\mathcal{F}(b_m)|$$

where $\mathcal{F}$ denotes Fourier transform.

Because thermal noise is a stochastic process, there will be slight power fluctuations from frequency to frequency for any finite-length segment, and no two finite-length segments n_m are exactly identical. These uncertainties are minimized with increased length for n_m, and by computing ρ_m from the averaged power, which converges to the true value as the number of analysed frequencies increases:

$$\rho'_m = mean(|\mathcal{F}(a_m)|) - mean(|\mathcal{F}(b_m)|)$$

We avoid aliasing by removing the contribution of ρ'_m, at each frequency, in the per-channel data (Fig. 6(b)). This is achieved by building a set of vectors describing the aliased contents in the frequency domain:

$$\vec{V}_m = \rho'_m e^{j \times arg(\mathcal{F}(y_m))}$$

We then remove these aliased contents $\vec{V}_m$ from the per-channel data y_m in the frequency domain. In doing so, we recover the data o_m with the effects of aliasing minimized:

$$o_m = \mathcal{F}^{-1}(\mathcal{F}(y_m) - \vec{V}_m)$$

As a final remark, here we consider how our acquisition methodology relates to compressed sensing.

To recover the signal of interest from the under-sampled (below Nyquist rate) data stream, compressed sensing typically relies on random sampling followed by an optimization-based reconstruction. The latter is an iterative procedure, which generally takes considerable processing time. This is particularly problematic for at-scale electrophysiology, involving thousands of recording channels or more. Taking advantage of the spectral characteristics of thermal noise aliasing, we can instead compute and remove the aliased thermal noise from the per-channel data in constant time, thereby recovering the signal despite sub-Nyquist-rate sampling - the hallmark of compressed sensing. Furthermore, compressed sensing relies on incoherent (random) measurements to spread the spectral power of the under-sampled (aliased) contents evenly across frequencies. This is notionally similar to our acquisition strategy, where we took advantage of the averaging properties of thermal noise aliasing in the frequency domain.

IV. Performance Evaluation

A. Testing Strategy

To quantify the performance of SRMA, it is imperative that we know the exact signal and noise entering the system, such that we may compare outputs against ground truth. These conditions can only be realized computationally, as shown in Fig. 7(a).

First, a sine wave $s[t]$ with frequency ω is generated at a rate of f_{samp}, with $f_{samp} \gg \omega$. We get $\lambda[t]$ by down-sampling $s[t]$ to f_{csamp}, with $2\omega < f_{csamp} \ll f_{samp}$. Because $s[t]$ is down-sampled to a new frequency above its Nyquist rate, $\lambda[t]$ is effectively the outcome of a perfect sampling system.

Simulating thermal noise generated by electrodes immersed in physiological media [20], [21], we add to $s[t]$ a bandlimited Gaussian noise $n[t]$. This noise is limited to the bandwidth of the system under test ($f_{samp}/2$), to emulate finite bandwidth of real acquisition systems. We tested two types of multiplexed sampling systems: SRMA – our new architecture (Fig. 3), and convM – the conventional multiplexed sampling system (Fig. 1). Their outputs are $\alpha[t]$ and $\beta[t]$, respectively. The convM block takes an additional noise source $m[t]$, to account for the antialiasing RC network noise (Fig. 2), due to limited space for a large-value capacitor in high-density, high-channel-count implementations.

The test signal $s[t]$ is a 1 kHz sine wave ($\omega = 1$ kHz). All three outputs, $\alpha[t]$, $\beta[t]$ and $\lambda[t]$, have identical rate, f_{csamp}. We used $f_{samp} = 5$ MHz and $f_{csamp} = 10$ kHz, to emulate 500:1 multiplexing. Accounting for the per-channel antialiasing filter in convM, we low-pass filter its input ($s[t] + n[t]$) at 3 kHz. This is above 2ω and below $f_{csamp}/2$. For fair comparisons between convM and SRMA, we also low-pass filter the output of SRMA at 3 kHz, so that both $\alpha[t]$ and $\beta[t]$ are band-limited to 3 kHz.

We now define the variance for the signal $s[t]$, the thermal noise $n[t]$ and the filter noise $m[t]$. Recalling that signal-to-noise ratio (SNR) can be expressed in terms of amplitude variance (σ^2) for zero-mean time series: $\mathrm{SNR} = \sigma^2_{signal}/\sigma^2_{noise}$, with the variance of $n[t]$ at unity, we set the variance of $s[t]$ to 15, for a recording SNR of 15. Using a MiM density of 4 $\mathrm{fF}/\mu\mathrm{m}^2$ (Fig. 2), and a layout area of 25 μm by 25 μm, to record from neurons with approximately 25 μm diameter, we would achieve a capacitance of 2.5 pF. A low-pass RC filter with this capacitance has rms noise

$$\sqrt{\sigma^2_m} = \sqrt{\frac{kT}{C}} = \sqrt{\frac{1.38e^{-23} \times 310}{2.5\,\mathrm{pF}}} = 41.4\ \mu\mathrm{V\ rms}$$

Typical extracellular microelectrodes in physiological media, such as the cerebrospinal fluid, have thermal noise of approximately $10\,\mu$V rms. We can therefore write the filter noise

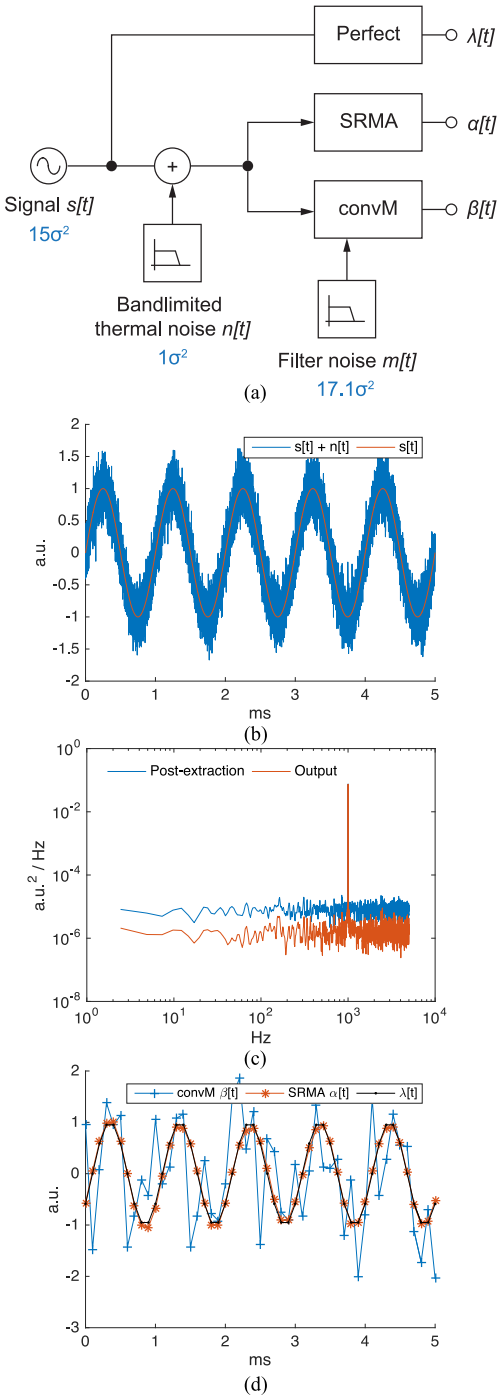

Fig. 7. Testing the SRMA architecture. (a) Test setup for comparing SRMA against a perfect sampler and conventional multiplexing with limited space for RC filter capacitance (convM). (b) Representative signal ($s[t]$) and noise contaminated signal ($s[t] + n[t]$). (c) Power spectral density of signal before and after removal of aliased thermal noise by SRMA. (d) Output of the perfect sampler (λ), SRMA (α) and conventional multiplexing (β).

variance σ_m^2 in terms of σ_n^2

$$\frac{\sqrt{\sigma_m^2}}{\sqrt{\sigma_n^2}} = \frac{41.4}{10}$$

$$\left(\frac{\sqrt{\sigma_m^2}}{\sqrt{\sigma_n^2}}\right)^2 = 4.14^2$$

$$\frac{\sigma_m^2}{\sigma_n^2} \approx 17.1$$

$$\sigma_m^2 = 17.1\sigma_n^2$$

For the purpose of comparing multiplexer performance, the SRMA and convM blocks are noiseless in Fig. 7(a), such that σ_m^2 and σ_n^2 together account for all the noise in the test setup.

B. SRMA Outperforms Conventional Multiplexing

Fig. 7(b) illustrates a noise-contaminated input $s[t] + n[t]$ for SRMA and conventional multiplexing with a small-capacitance RC filter (convM). The true signal $s[t]$ is also plotted for comparison. Fig. 7(c) shows the power spectral density estimate for SRMA's extracted per-channel data (blue, corresponding to $y_m[t]$ in Fig. 3) and its final output (orange, corresponding to $o_m[t]$ in Fig. 3). Due to the 500-fold reduction in sampling rate, thermal noise is aliased in the per-channel data. As desired, SRMA uniformly remove the power of the aliased thermal noise across the entire output bandwidth of $f_{csamp}/2 = 5$ kHz.

Fig. 7(d) compares the output of SRMA and convM to the perfect sampler's output $\lambda[t]$, given the same noise-contaminated input in Fig. 7(b). The SRMA output closely resembles $\lambda[t]$. In contrast, the output of the conventional multiplexing scheme with small-capacitance RC deviates significantly from $\lambda[t]$, due to the additional noise contributed by the RC network. Finally, by virtue of the SRMA operational procedures, its output is statistically and spectrally indistinguishable from that of the perfect sampler as the data length approaches infinity.

We can quantify each multiplexing scheme's error magnitude over a closed range of sample points $(a \ldots b)$ by comparing its output sum-of-squares error (SSE) against $\lambda[t]$, normalize by the range length $L = (b - a + 1)$

$$SSE/L : \frac{1}{L}\sum_{t=a}^{b}(x[t] - \lambda[t])^2$$

We generated 30 sets of $s[t] + n[t]$ and tested them on convM and SRMA. In every case SRMA performed much better than convM (Fig. 8; Paired t-test, $p < 0.0001$). The length-normalized sum of squares errors (SSE/L) of SRMA were significantly lower than those of convM for all tested datasets (Fig. 8), indicating that SRMA's outputs are much closer to that of the perfect sampler than convM.

C. SRMA is Highly Robust in the Presence of Noise

SRMA operates well in poor SNR conditions. As an example, the test data $s[t] + n[t]$ in Fig. 9(a) has a SNR of 2. SRMA produced an output closely matching that of the perfect sampler (Fig. 9(b)). In contrast, the capacitance-limited conventional scheme, convM, failed to produce useful output, due to the additional noise contributed by the small-capacitance RC network. To examine how well SRMA operates at different noise levels, we generated test data across a range of SNRs and compared the length-normalized SSE of SRMA and convM (Fig. 9(c)). SRMA consistently out-performed convM. This was statistically

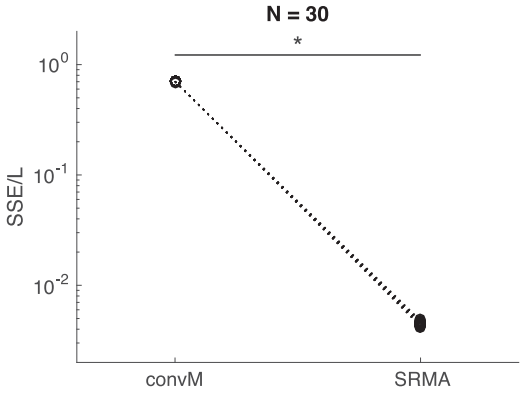

Fig. 8. Deviation (length-normalized sum of squares error; SSE/L) from perfect sampling for conventional multiplexing with limited RC filter capacitance (convM) and SRMA. SRMA performed significantly better than convM. The symbol '*' denotes statistical significance.

analyzed by two-way ANOVA (analysis of variance; [22]). The difference between SRMA and convM was highly significant, with a p-value of <0.0001.

V. System Implementation

We implemented the SRMA architecture in a 65,536-channel, multiplexed, electrophysiology system (Fig. 10(a)). It consists of a custom CMOS IC fabricated in a 1.8 V/3.3 V CMOS process (Fig. 10(b)). The IC has an array of 256 by 256 front-end sensors, each occupying $25.5 \times 25.5\ \mu\text{m}^2$. The tight pitch gives us the ability to achieve one-to-one mapping between sensors and neurons. In-house post-processing [23] of the ICs allows us to interface them directly with neurons. In particular, we deposited 6 nm of HfO_2, a high-K dielectric, on top of each electrode. This provides a capacitance of 5.8 pF over each $14\ \mu\text{m} \times 14\ \mu\text{m}$ electrode. A pseudo-resistor, constructed from a p-type MOSFET operating in weak inversion and placed in parallel with the foregoing capacitor, forms a high-pass filter for the input signal. The corner frequency is user-configurable, by setting the gate voltage of the pseudo-resistor.

Every front-end sensor contains an amplifier, microstimulator, control logic and a multiplexer switch (Fig. 10(c)). The array is partitioned into 16 banks, with 4096 front-end elements each. These front-ends are multiplexed into a shared back-end circuit within the IC, containing a band-pass filter and additional amplifiers. The last amplifier in the back-end (Fig. 10(e)) has user-selectable gain, ranging from $1\times$ to $5\times$, to cater for input amplitude variations between different biological specimen.

The full-differential output from each IC back-end circuit is connected to a digitization circuit implemented on the PCB using discrete components (Fig. 10(f)). This board-level circuit is comprised of a Sallen-Key band-pass filter, a 12-bit ADC, digital buffers for the ADC outputs and a Xilinx Spartan-6 FPGA. Each FPGA handles the outputs from four ADCs and transmits the data to the host PC via USB3. The SRMA digital signal processing steps are implemented on the PC in C++. Each SRMA instance handles data from one of the sixteen banks in the IC array.

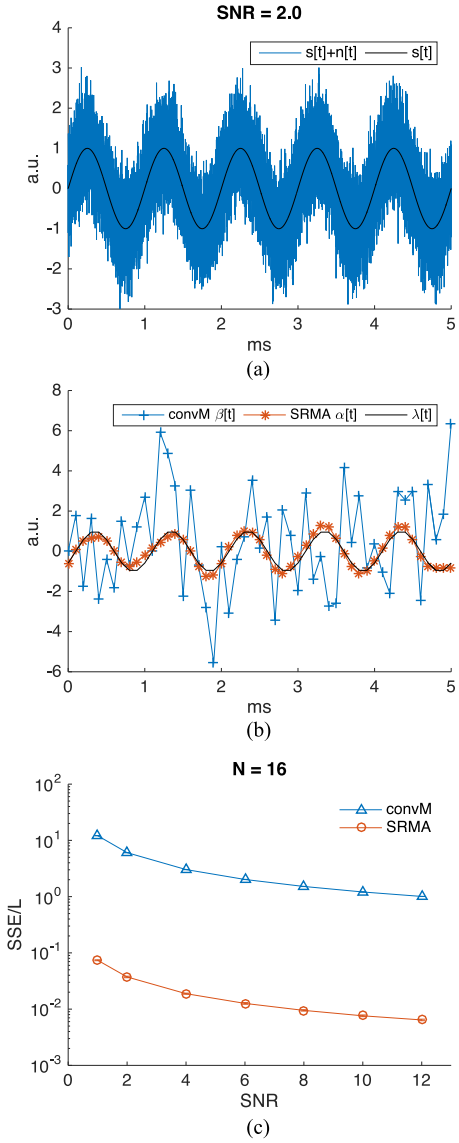

Fig. 9. Performance of SRMA with low-SNR inputs. (a) Test data with SNR of 2. (b) The SRMA output closely matched that of a perfect sampler (λ), while conventional multiplexing with limited RC filter capacitance (convM) performed poorly. (c) SRMA outperformed convM for all test data, across all SNR values. We repeated the test with sixteen different inputs at each SNR.

Fig. 11(a) and (b) show the normalized bandwidth of the IC front-end and back-end circuits, respectively. These values were determined by applying sine waves, of different frequencies, at test points built into the IC, while recording the applied signals using the system's hardware and software. The power spectral density for the input-referred noise, when measured in physiological saline and after SRMA processing, is approximately 10 μV rms over the 100–3 kHz bandwidth (Fig. 11(c)). The entire system uses about 24.7 W during operation (6 V supply). The power consumption is dominated by the four Xilinx Spartan-6 FPGAs, and to a lesser extent, the board-level Sallen-Key filters. The IC consumes less than 0.61 % of the power budget.

We next tested SRMA-based recordings in this system by applying a 1 kHz sine wave, through a pair of silver-silver

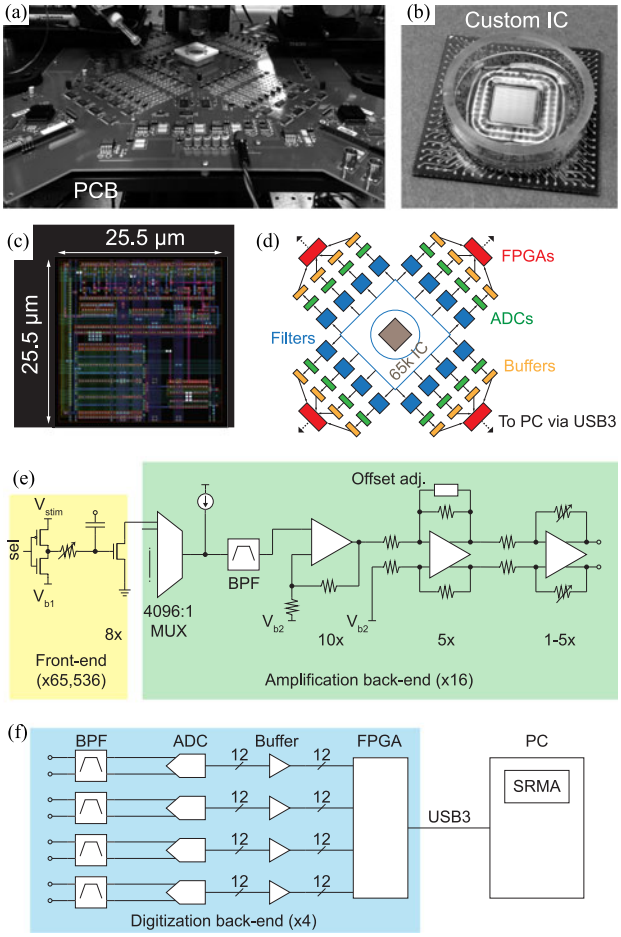

Fig. 10. SRMA implementation on a 65,536-channel, multiplexed, electrophysiology system. (a) Photo of the complete system, consisting of circuit board and (b) custom integrated circuit. (c) Layout of the front-end sensors. (d) Overview of the system. The 65,536 front-end channels are divided into 16 banks, each handled by a separate back-end, implemented within the IC and with board-level discrete components. (e) Overview of the IC recording circuit. BPF, band-pass filter. (f) Overview of the PCB digitization circuit.

chloride electrodes, into the chamber above the recording array, as depicted in Fig. 11(d). The chamber was filled with conductive physiological phosphate buffered saline, to mimic conditions similar to those in biological experiments. We reduced the sine wave to typical electrophysiological signal amplitude of 100 μV using attenuators, op-amp buffers and isolation transformers.

The SRMA readout for one of the electrodes is shown in Fig. 11(e) (blue trace). Patch clamp recordings have been the gold-standard in electrophysiology [25], [26]. This non-multiplexed, low-noise recording technique is carried out with a commercial, purpose-built amplifier (Molecular Devices MultiClamp 700B). To verify the SRMA output, we performed patch clamp recordings within 50 μm above the custom IC front-end electrode from which SRMA acquired the test signal (Fig. 11(e), red trace). One notices the close correspondence between the SRMA output and the patch clamp recording.

A number of at-scale, CMOS-based recording arrays, with thousands to tens-of-thousands of recording electrodes have

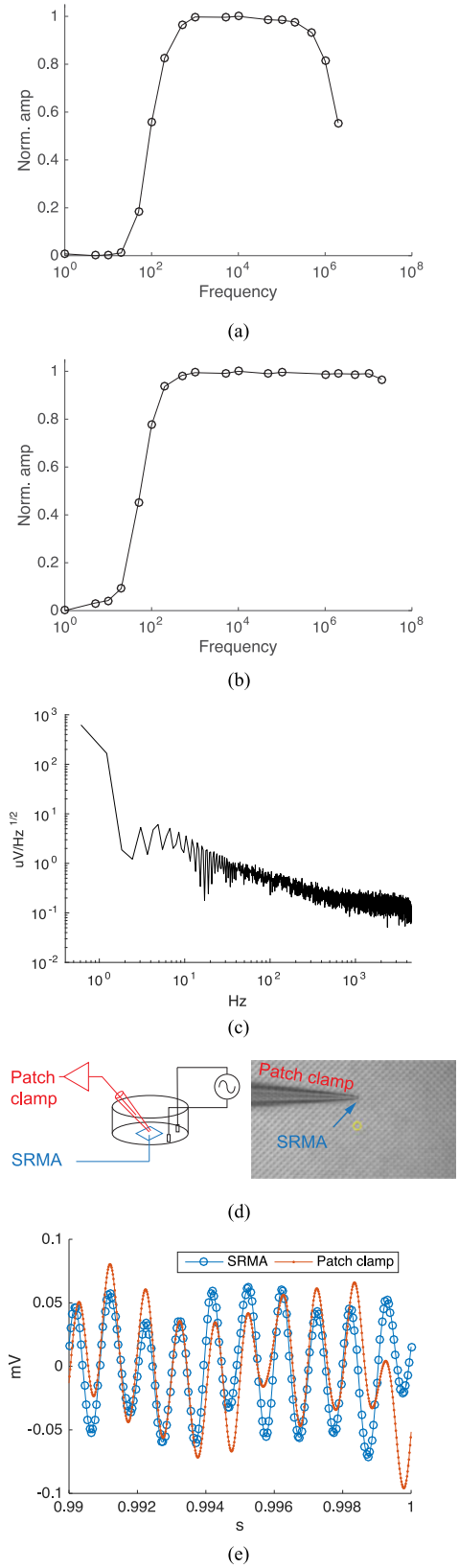

Fig. 11. Characterizing the IC and SRMA. (a) Normalized bandwidth of the front-end circuit (65,536 elements). (b) Normalized bandwidth of the back-end circuit (16 elements). (c) Power spectral density of input-referred noise for one channel. (d) Setup for comparison to patch clamp recordings. (e) Example outputs from SRMA and patch clamp recording. Both traces have been band-pass filtered between 300 and 3 k Hz for clarity.

TABLE I
SUMMARY OF AT-SCALE, CMOS-BASED RECORDING ARRAYS

	[7]	[8]	[14]	[24]	[10]	[15]	This work
Process (μm)	0.5	0.35	0.6	0.6	0.18	0.35	0.18
Voltage (V)	5	3.3	3.3/5	5	1.8/3.3	3.3	1.8/3.3
Sensing area (mm^2)	2.6	7.13	3.5	73.73	1.08/4.33	8.09	42.61
Electrodes	32,768	4,096	11,016	81,920	4,225	26,400	65,536
Pitch (μm)	8.775	42	17.8	30	16/32	17.5	25.5
Parallel recordings	32,768	4,096	126	1,024	4,225	1,024	65,536
Noise (μV rms)	$>$50	26	2.4 (1–10 kHz)	5	$>$44 (300–10 kHz)	2.4 (300–10 kHz)	10 (100–10 kHz)
IC Power (mW)	4,000	132	135	175	–	75	153
Stimulation	No	No	Tes	No	Yes	Yes	Yes

been reported to date. Their key performance metrics are summarized in Table I. In several designs, the number of simultaneously recording channels is a small fraction of the available electrodes. Furthermore, it has not been previously possible to achieve better than 26 μV rms input-referred noise over the spike bandwidth with arrays having more than approximately four thousand simultaneously recording channels.

VI. CONCLUSION

In this paper we show that traditional multiplexing approaches are not scalable for high-density, high-channel-count electrophysiology. As the per-channel antialiasing filters are made smaller, the thermal noise of these circuit elements increases, to the extent that recordings of typical neural signals on the order of 100 μV peak-to-peak is no longer possible. As a solution, we developed a new multiplexing scheme (statistically reconstructed multiplexing architecture, SRMA) without the need for these per-channel antialiasing filters. The spectral power contributed by the under-sampled thermal noise is calculated by statistical reconstruction, then removed from the per-channel data, thereby preventing aliasing.

We quantified the SNR performance improvements of SRMA over that of traditional multiplexing with area-limitations expected in high-density applications, and showed that SRMA is able to extract signals with significantly better accuracy. Furthermore, we implemented SRMA on a 65,536-channel, multiplexed, electrophysiological recording system. The new architecture is able to acquire test signals in a physiological environment with outputs comparable to single-channel, low noise patch clamp recordings.

REFERENCES

[1] M. E. J. Obien, K. Deligkaris, T. Bullmann, D. J. Bakkum, and U. Frey, "Revealing neuronal function through microelectrode array recordings," *Frontiers Neurosci.*, vol. 8, 2015, Art. no. 423.

[2] C. A. Curcio and K. A. Allen, "Topography of galion cells in human retina," *J. Comparative Neurol.*, vol. 300, pp. 5–25, 1990.

[3] G. Leuba and R. Kraftsik, "Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age," *Anatomy Embryol.*, vol. 190, pp. 351–366, 1994.

[4] G. T. Einevoll, F. Franke, E. Hagen, C. Pouzat, and K. D. Harris, "Towards reliable spike-train recordings from thousands of neurons with multielectrodes," *Current Opinion Neurobiol.*, vol. 22, pp. 11–17, 2012.

[5] G. Buzsáki *et al.*, "Tools for probing local circuits: High-density silicon probes combined with optogenetics," *Neuron*, vol. 86, pp. 92–105, 2015.

[6] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents - EEG, ECoG, LFP and spikes," *Nature Neurosci.*, vol. 13, pp. 407–420, 2012.

[7] B. Eversmann *et al.*, "A 128 $\times$ 128 CMOS biosensor array for extracellular recording of neural activity," *IEEE J. Solid-State Circuits*, vol. 38, no. 12, pp. 2306–2317, Dec. 2003.

[8] L. Berdondini *et al.*, "Active pixel sensor array for high spatio-temporal resolution electrophysiological recordings from single cell to large scale neuronal networks," *Lab Chip*, vol. 9, pp. 2644–2651, 2009.

[9] J. Du, T. J. Blanche, R. R. Harrison, H. A. Lester, and S. C. Masmanidis, "Multiplexed, high density electrophysiology with nanofabricated neural probes," *PLoS ONE*, vol. 6, 2011, Art. no. e26204.

[10] G. Bertotti *et al.*, "A CMOS-based sensor array for in-vitro neural tissue interfacing with 4225 recording sites and 1024 stimulation sites," in *Proc. 9th Int. Meeting Substrate-Integrated Microelectrode Arrays*, 2014, pp. 247–250

[11] *CMOS7RF (CMRF7SF) Design Manual*, IBM Microelectron. Div., Armonk, NY, USA, 2008.

[12] A. P. Alivisatos *et al.*, "The brain activity map project and the challenge of functional connectomics," *Neuron*, vol. 74, pp. 970–974, 2012.

[13] J. Chang, A. A. Abidi, and C. R. Viswanathan, "Flicker noise in CMOS transistors from subthreshold to strong inversion at various temperatures," *IEEE Trans. Electron Devices*, vol. 41, no. 11, pp. 1965–1971, Nov. 1994.

[14] U. Frey, U. Egert, F. Heer, S. Hafizovic, and A. Hierlemann, "Microelectronic system for high-resolution mapping of extracellular electric fields applied to brain slices," *Biosensors Bioelectron.*, vol. 24, pp. 2191–2198, 2009.

[15] M. Ballini *et al.*, "A 1024-channel CMOS microelectrode array with 26,400 electrodes for recording and stimulation of electrogenic cells in vitro," *IEEE J. Solid-State Circuits*, vol. 49, pp. 2705–2719, 2015.

[16] A. Kohlenberg, "Exact interpolation of band limited functions," *J. Appl. Phys.*, vol. 24, pp. 1432–1436, 1953.

[17] A. Aldroubi and K. Gröchenig, "Nonuniform sampling and reconstruction in shift-invariant spaces," *SIAM Rev.*, vol. 43, pp. 585–620, 2001.

[18] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, pp. 1207–1223, 2006.

[19] J. R. Barry, E. A. Lee, and D. G. Messerschmitt, *Digital Communication*, 3rd ed. Norwell, MA, USA: Kluwer, 2003.

[20] S. F. Cogan, "Neural stimulation and recording electrodes," *Annu. Rev. Biomed. Eng.*, vol. 10, pp. 275–309, 2008.

[21] N. P. Aryan, H. Kaim, and A. Rothermel, *Stimulation and Recording Electrodes for Neural Prostheses*. Berlin, Germany: Springer, 2015.

[22] D. C. Montgomery, G. C. Runger, and N. F. Hubele, *Engineering Statistics*, 5th ed. Hoboken, NJ, USA: Wiley, 2010.

[23] D. Tsai, E. John, T. Chari, R. Yuste, and K. Shepard, "High-channel-count, high-density microelectrode array for closed-loop investigation of neuronal networks," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2015, pp. 7510–7513.

[24] L. J. Johnson, E. Cohen, D. Ilg, R. Klein, P. Skeath, and D. A. Scribner, "A novel high electrode count spike recording array using an 81,920 pixel transimpedance amplifier-based imaging chip," *J. Neurosci. Methods*, vol. 205, pp. 223–232, 2012.

[25] G. J. Stuart, H. U. Dodt, and B. Sakmann, "Patch-clamp recordings from the soma and dendrites of neurons in brain slices using infrared video microscopy," *Pflügers Archiv*, vol. 423, pp. 511–518, 1993.

[26] W. Walz, *Patch-Clamp Analysis Advanced Techniques*. New York, NY, USA: Humana Press, 2007.

David Tsai (M'03) received the B.E. degree in software engineering (first class Hons.), the Master's degree in biomedical engineering from the University of New South Wales, Sydney, NSW, Australia, in 2007, and the Ph.D. degree in biomedical engineering from the University of New South Wales, in 2012, on electrical stimulation strategies for retinal implants to restore sight in the blind. He worked as an embedded system Engineer and on formal systems verification, including the L4 microkernel at National ICT Australia. It has been shipped in more than a billion Qualcomm wireless chips, then later in Apple iOS devices. He is currently a NHMRC CJ Martin Postdoctoral Research Fellow, between Electrical Engineering and Biological Sciences of Columbia University, New York, NY, USA. He has 6 granted and pending patents, and has published more than 25 journal articles and peer-reviewed proceedings. His research interests include neural interfaces, implantable bionics, computational neuroscience, and neurobiology of the visual system.

Rafael Yuste was born in Madrid, Spain. He received the Medical degree from the Universidad Autónoma, Madrid, Spain. He was engaged in Ph.D. study with Larry Katz in Torsten Wiesel's Laboratory, Rockefeller University, New York, NY, USA. He is currently a Professor of Biological Sciences and Neuroscience at Columbia University, New York, NY, USA. He was a Postdoctoral student of David Tank at Bell Labs. He is interested in understanding the function and pathology of the cerebral cortex, using calcium imaging and optogenetics to "break the code" and decipher the communication between groups of neurons. He also led the researchers who proposed the Brain Activity Map, precursor to the BRAIN initiative, and is currently a member of the NIH BRAIN-I advisory committee. He then joined the Sydney Brenner's Laboratory, Cambridge, U.K. In 2005, he became a Howard Hughes Medical Institute investigator and the Co-Director of the Kavli Institute for Brain Circuits. Since 2014, he has been serving as the Director of the Neurotechnology Center, Columbia. He is a member of Spain's Royal Academies of Science and Medicine. He has received many awards for his work, including those from the New York City Mayor, the Society for Neuroscience, and the National Institutes of Health's Director.

Kenneth L. Shepard (M'91–SM'03–F'08) received the B.S.E. degree from Princeton University, Princeton, NJ, USA, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1988 and 1992, respectively. From 1992 to 1997, he was a Research Staff Member and the Manager with the VLSI Design Department, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA, where he was responsible for the design methodology for IBMs G4 S/390 microprocessors. He was the Chief Technology Officer with CadMOS Design Technology, San Jose, CA, USA, until its acquisition by Cadence Design Systems in 2001. Since 1997, he has been with Columbia University, New York, NY, USA, where he is currently the Lau Family Professor of Electrical Engineering and Biomedical Engineering and the Co-Founder and the Chairman of the Board of Ferric, Inc., New York, NY, USA, which is a commercializing technology for integrated voltage regulators. His current research interests include power electronics, carbon-based devices and circuits, and CMOS bioelectronics. He has been an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE-SCALE INTEGRATION SYSTEMS, the IEEE JOURNAL OF SOLID-STATE CIRCUITS, and the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS.