# Analysis of Encoding Degradation in Spiking Sensors Due to Spike Delay Variation

Minhao Yang, *Member, IEEE*, Shih-Chii Liu, *Senior Member, IEEE*, and Tobi Delbruck, *Fellow, IEEE*

***Abstract*— Spiking sensors such as the silicon retina and cochlea encode analog signals into massively parallel asynchronous spike train output where the information is contained in the precise spike timing. The variation of the spike timing that arises from spike transmission degrades signal encoding quality. Using the signal-to-distortion ratio (SDR) metric with nonlinear spike train decoding based on frame theory, two particular sources of delay variation including comparison delay $T_{\mathrm{DC}}$ and queueing delay $T_{\mathrm{DQ}}$ are evaluated on two encoding mechanisms which have been used for implementations of silicon array spiking sensors: asynchronous delta modulation and self-timed reset. As specific examples, $T_{\mathrm{DC}}$ is obtained from a 2T current-mode comparator, and $T_{\mathrm{DQ}}$ is obtained from an M/D/1 queue for 1-D sensors like the silicon cochlea and an $\mathrm{M}^{\mathrm{X}}$/D/1 queue for 2-D sensors like the silicon retina. Quantitative relations between the SDR and the circuit and system parameters of spiking sensors are established. The analysis method presented in this work will be useful for future specifications-guided designs of spiking sensors.**

***Index Terms*— Asynchronous communication, comparison delay, delay variation, frame theory, queueing delay, queueing theory, signal-to-distortion ratio (SDR), spike decoding, spike encoding, spiking sensors.**

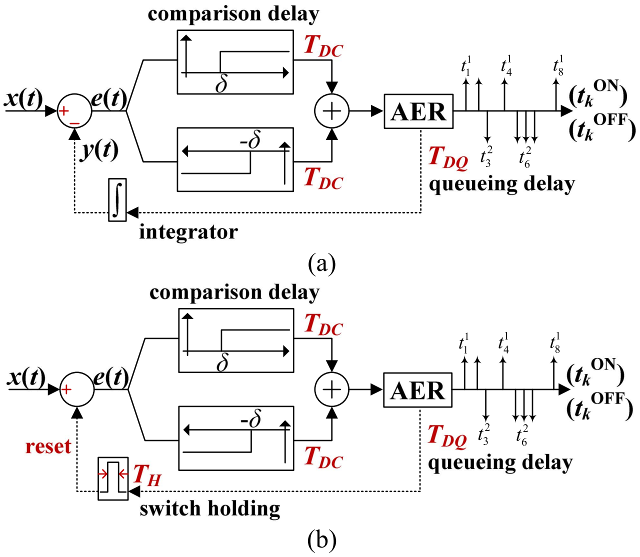


Fig. 1. Spike encoding models of (a) asynchronous delta modulation (ADM), and (b) self-timed reset (STR). The total communication delay $T_D$ is the sum of the comparison delay $T_{DC}$ and the queueing delay $T_{DQ}$. $T_H$ is the switch-holding period in the STR model.

## I. Introduction

TEMPORAL precision of spike timestamps is essential in determining the integrity of signal representation in the spike domain, for both biological sensory systems [1] and artificial spiking sensors [2]–[6]. In the cat's lateral geniculate nucleus, the temporal structure of a neuron's spiking responses has a finer timescale than that of the input stimuli filtered by the neuron's temporal receptive field, and modeling results have shown that large timing jitter with its reciprocal comparable to the firing rate results in much degraded signal reconstruction [1]. This principle of jitter-constrained representation integrity also applies to spiking sensors such as the silicon retina [2]–[4], the silicon cochlea [5], [6], and neural recording arrays [7] where the generated spike trains are asynchronously transmitted. The sources of timing jitter in these spiking sensors can be divided into two categories: the electronic noise and the uncertainty in spike transmission delay. The electronic noise arises from, e.g., the photodiode, transistors and switching kTC, and can be reduced by established means such as high illumination, large transistor size and capacitance. The impact of delay variation in asynchronous spike transmission on the encoding quality, on the other hand, has not yet been quantitatively analyzed to guide the design of spiking sensors. From the application point of view, the interest in quantitative analysis of spike timing jitter lies in its degenerative effects on the system performance of spike-based high-quality signal coding [7]–[9] and high-accuracy pattern recognition [10].

Two commonly used encoding mechanisms that have been implemented in array spiking sensors are the self-timed reset (STR) [2], [3] and asynchronous delta modulation (ADM) [4], [6].With linear decoding (an ideal integrator followed by a sinc filter), the ADM (Fig. 1a) with integrate and subtraction feedback often results in a better signal-to-distortion ratio (SDR) than the STR (Fig. 1 b) with reset feedback, especially at high input frequencies and large input-amplitude-to-threshold ratios [11]. A constant time shift, $T_D$, was used in [11] to model the communication delay caused by in-pixel or in-channel comparison and asynchronous handshake with peripheral circuits in both ADM and STR. A constant $T_H$ was used to model the switch-holding period (corresponding to the refractory period in biological neurons) in STR for the complete reset of the capacitive amplifier and

Manuscript received May 5, 2016; revised August 15, 2016; accepted September 19, 2016. Date of publication December 15, 2016; date of current version January 6, 2017. Minhao Yang is currently supported by SNF Early Postdoc Mobility fellowship. This work was funded by SNF NCCR Robotics. This paper was recommended by Associate Editor T. Serrano-Gotarredona.

M. Yang was with the Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich 8057, Switzerland and is now with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: yangmh.ic@gmail.com).

S.-C. Liu and T. Delbruck are with the Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich 8057, Switzerland (e-mail: shih@ini.uzh.ch; tobi@ini.uzh.ch).

Digital Object Identifier 10.1109/TCSI.2016.2613503

therefore limiting the firing rate of the encoder. The decoded signal can be aligned with the input by applying a constant time shift $T_D$ to the SDR calculation. However, $T_D$ varies in practice for the following two reasons. First, the comparison delay $T_{DC}$ of the in-pixel or in-channel comparator contributes to $T_D$ variation, because this delay is dependent on the slope of input signals at the moment of threshold-crossing [9]. Given a certain bias current of the comparator, a larger input slope leads to a smaller $T_{DC}$. Second, without cheap and reliable 3D integration technologies for massively parallel point-to-point dense connections between chips [12], the address event representation (AER) [13] remains the most viable asynchronous spike communication protocol for chip-to-chip links, where arbitration is employed to prevent spike collision during sequential transmission, and as a result leads to queueing delay $T_{DQ}$ variation. The waiting time distribution $P(W)$ in queueing models can be used to describe $T_{DQ}$ variation. A two-stage Poisson arrival queueing model [14] was applied to estimate the average $T_{DQ}$ of the 2-D burst-mode word-serial AER [13], a spike transmission protocol used in the latest silicon retina chips [4]. However, no $T_{DQ}$ variation can be readily derived from the model due to the lack of an explicit expression of waiting time.

This paper focuses on the analysis of spike encoding degradation in spiking sensors with ADM and STR due to $T_D$ variation caused by the two sources, variation in the comparator delay $T_{DC}$ and variation in the queuing delay $T_{DQ}$. The results of this analysis will give quantitative insights into practical design considerations of circuit and system parameters of spiking sensors. Instead of linear decoding, nonlinear spike decoding based on frame theory [15] is used to evaluate the signal integrity because it can theoretically give a perfect reconstruction. The $T_{DC}$ variation is calculated based on a specific current-mode 2T comparator model, and the $T_{DQ}$ variation is obtained from queueing models that are adapted from the Poisson arrival, deterministic service, single server queue (M/D/1) for 1-D sensors and the bulk Poisson arrival, deterministic service, single server queue ($\mathrm{M}^{\mathrm{X}}$/D/1) for 2-D sensors, respectively. The SDR metric used as a measure of the spike encoding quality in this work, is a direct reflection of the intrinsic signal representation integrity in spike domain. Similar to the SNR metric used for synchronous ADCs, the SDR metric depends on the parameters of a spike encoding system regardless of the subsequent stage implementing spike processing algorithms for different applications.

The rest of the paper is organized as follows: Section II describes the nonlinear decoding algorithms for both ADM and STR spike encoders that are the basis for SDR computation in Section V; Section III derives the relationship between the comparison delay $T_{DC}$ and the input signal slope as the function of comparator parameters. This relationship is used to evaluate the impact of the $T_{DC}$ variation on the decoded signal SDR in Section V; Section IV describes the queueing models for 1-D and 2-D sensors, and shows the mean value and the cumulative distribution function of the queueing delay $T_{DQ}$ as the function of sensor array size, and the results are used to evaluate the impact of the $T_{DQ}$ variation on the decoded signal SDR in Section V; Section V shows the SDR degradation of signals decoded from spike trains due to the variation of $T_{DC}$ and $T_{DQ}$ as the function of several design parameters of the spiking silicon retina and cochlea. Section VI concludes the paper with discussions on implications for future specifications-guided designs of artificial spiking sensory system.

TABLE I

LIST OF MAIN SYMBOLS USED IN SECTION II

| Symbol | Description |
|---|---|
| $T_{DC}$ | Comparison delay |
| $T_{DQ}$ | Queueing delay |
| $\delta$ | Encoder threshold |
| $\chi$ | Spike sampling function for spike encoding |
| $g$ | Impulse response function of ideal lowpass filters |
| $\eta$ | Representation function for spike decoding |
| $c$ | Weighting coefficient for spike decoding |

## II. NONLINEAR SPIKE DECODING FOR ADM AND STR

Table I summarizes the main symbols used in this section. We first describe the ADM and STR models in Fig. 1. In both cases, we have:

$$T_D = T_{DC} + T_{DQ} \tag{1}$$

*ADM:* In Fig. 1(a), the error signal $e(t)$ is the input signal $x(t)$ subtracted by the feedback signal $y(t)$. Whenever $e(t)$ is above the upper threshold $\delta$ or below the lower threshold $-\delta$, an ON or OFF spike is generated. The generated spike is transmitted by the AER, and is also integrated by an ideal integrator leading to $y(t)$. If a spike is regarded as an ideal delta function, then $y(t)$ has a staircase-like shape.

*STR:* In Fig. 1(b), $e(t)$ starts to follow the incremental change of $x(t)$ from a reset level after each switch-holding period $T_H$. A generated ON or OFF spike due to threshold-crossing triggers a feedback reset operation, i.e., during the period of $T_H$, $e(t)$ is held at the reset level independent of $x(t)$. From the perspective of linear decoding, changes of the input $x(t)$ during $T_H$ is lost [11].

Next we describe the mathematical mapping from input signal amplitude to output spike timing following the $t$-transform [15] for both ADM and STR encoders.

*ADM:* Let $\forall k \epsilon N^+$, and let the delay for the $k^{\text{th}}$ threshold-crossing be denoted as $T_{D,k}$. The $t$-transform of an ADM encoder amounts to

$$x(t_k^1 - T_{D,k}) = \delta \cdot \left( k - 2 \sum_{l \in N^+} 1_{[t_l^2 < t_k^1]} \right) \tag{2}$$

$$x(t_k^2 - T_{D,k}) = -\delta \cdot \left( k - 2 \sum_{l \in N^+} 1_{[t_l^1 < t_k^2]} \right) \tag{3}$$

where $t_k^1/t_k^2$ represents the $k^{\text{th}}$ ON/OFF spike timestamps, $+\delta/-\delta$ is the upper/lower threshold of spike encoders, 1 denotes one count of a single spike, and its subscript $t_l < t_k$ means the condition of the timestamp $t_l$ smaller than the timestamp $t_k$. The inner product form of (2) and (3) can be written as

$$< x, \chi^i_{DA,k} >= q^i_{A,k} \tag{4}$$

where $i = 1,2$, and $q^i_{A,k}$ is the right side of (2) and (3). The sampling function $\chi^i_{DA,k}$ is defined as

$$\chi^i_{DA,k} = g(t - t^i_k + T_{D,k}) \tag{5}$$

with $g(\tau)$ being the impulse response of an ideal low-pass filter (LPF) with the cutoff frequency $\Omega$

$$g(\tau) = \sin(\Omega\tau)/\pi\tau \tag{6}$$

*STR:* Assume that the switch-holding period $T_H$ has a fixed value. The $t$-transform of an STR encoder is

$$x(t^i_k - T_{D,k}) - x(t^j_{k-1} + T_H) = (-1)^{i-1}\delta, \quad (k \geq 2) \tag{7}$$

$$x(t^i_1 - T_{D,1}) = (-1)^{i-1}\delta, \quad (k = 1) \tag{8}$$

where $t^i_k/t^j_k$ represents the timestamps of ON spikes when $i,j = 1$, and OFF spikes when $i,j = 2$. The inner product form of (7) and (8) can be written as

$$< x, \chi^i_{DS,k} > - < x, \chi^j_{HS,k-1} >= q^i_{S,k} \tag{9}$$

where the sampling functions $\chi^i_{DS,k}$ and $\chi^j_{HS,k-1}$ are given below

$$\chi^i_{DS,k} = g(t - t^i_k + T_{D,k}) \tag{10}$$

$$\chi^j_{HS,k-1} = g(t - t^j_{k-1} - T_H), \quad (k \geq 2) \tag{11}$$

$$\chi^j_{HS,0} = 0, \quad (k = 1) \tag{12}$$

and $q^i_{S,k}$ is the right side of either (7) or (8).

The nonlinear decoding algorithms which are based on frame theory are given below for both ADM and STR encoders. The proof is similar to *Proposition* 1 in [16].

*ADM:* The recovered signal $x_{rA}(t)$ from the spike output of an ADM encoder can be written as

$$x_{rA}(t) = \sum_{k\in N^+} c^1_{A,k}\eta^1_{A,k}(t) + \sum_{k\in N^+} c^2_{A,k}\eta^2_{A,k}(t) \tag{13}$$

where the representation functions $\eta^i_{A,k}$ $(i = 1,2)$ are written as

$$\eta^i_{A,k} = g(t - t^i_k) \tag{14}$$

These representation functions form a so-called frame [15]. With $c_{\mathbf{A}} = [\mathbf{c^1_A}; \mathbf{c^2_A}]$, and $[\mathbf{c^i_A}]_k = c^i_{A,k}$, the decoding coefficients $c_{\mathbf{A}}$ can be computed as

$$\mathbf{c_A} = \mathbf{G^+_A q_A} \tag{15}$$

where $q_{\mathbf{A}} = [\mathbf{q^1_A}; \mathbf{q^2_A}]$ with $[\mathbf{q^i_A}]_k = q^i_{A,k}$, and

$$\mathbf{G_A} = \begin{bmatrix} \mathbf{G^{11}_A} & \mathbf{G^{12}_A} \\ \mathbf{G^{21}_A} & \mathbf{G^{22}_A} \end{bmatrix}, \quad [\mathbf{G^{ij}_A}]_{kl} =< \chi^i_{DA,k},\eta^j_{A,l} > \tag{16}$$

for all $i,j = 1,2$, and $k,l\epsilon N^+$. $G^+_{\mathbf{A}}$ is the pseudo-inverse of $G_{\mathbf{A}}$. Using Parseval's formula [17], the elements in matrix $G_{\mathbf{A}}$ can be computed as

$$\begin{aligned}[\mathbf{G^{ij}_A}]_{kl} &= \int_{-\infty}^{\infty} g(t - t^i_k + T_{D,k})g(t - t^j_l)dt \\ &= \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{-i\omega(t^i_k - T_{D,k} - t^j_l)}d\omega = g(t^i_k - T_{D,k} - t^j_l)\end{aligned} \tag{17}$$

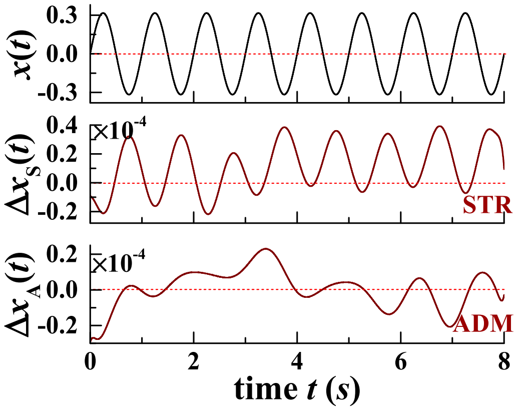


Fig. 2. Example waveforms of an input signal $x(t)$, and the spike reconstruction error $\Delta x(t)$ using both STR and ADM encoders.

*STR:* The recovered signal $x_{rS}(t)$ from the spike output of an STR encoder can be written as

$$x_{rS}(t) = \sum_{k\in N^+} c_{S,k}\eta_{S,k}(t) \tag{18}$$

where the representation functions $\eta_{S,k}$ are written as

$$\eta_{S,k} = g(t - t^i_k) \tag{19}$$

The vector of coefficients $c_{\mathbf{S}}$ with $[c_{\mathbf{S}}]_k = c_{s,k}$ can be computed as

$$\mathbf{c_S = G^+_S q_S} \tag{20}$$

where $[q_{\mathbf{S}}]_k = q^i_{S,k}$, $i = 1,2$, and

$$[\mathbf{G_S}]_{kl} = < \chi^i_{DS,k},\eta_{S,l} > - < \chi^j_{HS,k-1},\eta_{S,l} >, \quad (k \geq 2) \tag{21}$$

$$[\mathbf{G_S}]_{1l} = < \chi^i_{DS,1},\eta_{S,l} >, \quad (k = 1) \tag{22}$$

for all $i,j = 1,2$, and $k,l\epsilon N^+$. $G^+_{\mathbf{S}}$ is the pseudo-inverse of $G_{\mathbf{S}}$. The elements in matrix $G_{\mathbf{S}}$ can be computed as

$$\begin{aligned}[\mathbf{G}]_{kl} &= < \chi^i_{D,k}, \eta_{D,l} > - < \chi^j_{H,k-1}, \eta_{D,l} > \\ &= \mathrm{g}(t^i_k - T_{D,k} - t^m_l) - \mathrm{g}(t^j_{k-1} + T_H - t^n_l), \quad (k \geq 2)\end{aligned} \tag{23}$$

$$[\mathbf{G}]_{1l} = \mathrm{g}(t^i_1 - T_{D,1} - t^m_l), \quad (k = 1) \tag{24}$$

where $m$, $n = 1, 2$.

As a simple example of adopting the algorithms described above to decode the spike trains encoded by STR and ADM encoders, Fig. 2 shows the waveform of an input signal $x(t)$ and the corresponding reconstruction errors calculated as

$$\Delta x_A(t) = x(t) - x_{rA}(t) \tag{25}$$

$$\Delta x_S(t) = x(t) - x_{rS}(t) \tag{26}$$

$x(t)$ has the same sinusoid parameters as used in [11] with frequency $\Omega = 2\pi$ rad/s, amplitude $A = 0.316$, and threshold $\delta = A/2^3$. All $T_{D,k}$ and $T_H$ values are fixed to 7.8 ms. The simulation has a time support of 8 s and a time step of 3.8 $\mu$s.

As mentioned in the Introduction, we use the SDR metric to determine the fidelity of spike-domain signal representation and it is defined as

$$\mathrm{SDR}_e = \frac{\int_{t_1}^{t_2} (x(t))^2 dt}{\int_{t_1}^{t_2} (\Delta x_e(t))^2 dt} \tag{27}$$

where $e = A,S$, the $\mathrm{SDR}_A$ for the ADM encoder is calculated to be 87 dB, and $\mathrm{SDR}_S$ for the STR encoder is 80 dB. Both values are significantly higher than the values obtained using linear decoding (21 dB for ADM, and 8 dB for STR) [11]. This large SDR values from using nonlinear decoding allows us to evaluate the effects of small spike timing jitter on signal representation integrity in the spike domain. These jitter effects would be completely concealed if linear decoding was used. The residual error in the nonlinear reconstruction is from both the limited time resolution in simulation and finite time support.

## III. Modeling of Comparison Delay Variation $T_{DC}$

Comparators are used for detection of threshold-crossing, and one threshold-crossing event elicits one spike. Because of the finite bandwidth of comparators, a spike is not generated instantaneously the moment one threshold-crossing occurs, and hence there is a comparison delay $T_{DC}$ which is dependent on the input signal slope $s_{xT}$ at the moment of threshold-crossing as well as the comparator parameters. For continuous-time comparators commonly used in asynchronous systems, the relevant parameters are the bias current and the DC gain for a given circuit topology. The following analysis will take the simplest 2T common-source amplifier (CSA) [2] as an example to obtain an analytical expression of $T_{DC}$. Assuming one dominant pole and no slewing, the transfer function of the CSA comparator can be written as [18]

$$H(s) = \frac{A_{DC}}{s\tau_c + 1} \tag{28}$$

where $A_{DC}$ is the DC gain of the CSA, and $\tau_c$ is the time constant associated with the dominant pole. This time constant depends on the bias of the CSA and its output load capacitance. The small input signal change at the moment of threshold crossing is approximated by $s_{xT}\Delta t$, where $\Delta t$ is the time needed for the CSA output to change by $\delta$. The Laplace transform of $s_{xT}\Delta t$ is $s_{xT}/s^2$. The CSA output change in the Laplace domain, $\Delta V_{out}(s)$, is then written as

$$\Delta V_{out}(s) = \frac{s_{xT}}{s^2} \cdot \frac{A_{DC}}{s\tau_c + 1} \tag{29}$$

The time-domain output change $\Delta v_{out}(t)$ is obtained by the inverse Laplace transform of $\Delta V_{out}(s)$

$$\Delta v_{out}(t) = \mathcal{L}^{-1}\{\Delta V_{out}(s)\} = s_{xT}A_{DC}(t - \tau_c + \tau_c e^{-t/\tau_c}) \tag{30}$$

By replacing $t$ with $T_{DC}$ and using Taylor's expansion on the exponential term in (30), the $\Delta v_{out}(t)$ during $T_{DC}$ denoted as

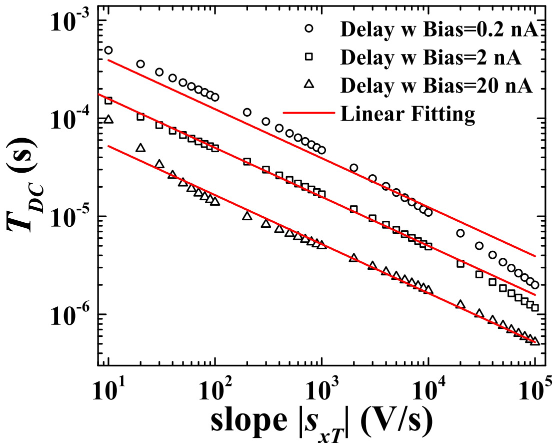


Fig. 3. Simulated comparison delay $T_{DC}$ versus input signal slope $s_{xT}$ at different bias settings of a 2T CSA comparator using Spectre in a 0.18 $\mu$m CMOS process. Equation (32) is used for the linear fits.

$\Delta v_{outT}$ can be formulated as follows assuming $T_{DC} \ll 3\tau_c$

$$\begin{aligned}\Delta v_{outT} &= s_{xT}A_{DC}\left[T_{DC} - \tau_c + \tau_c\left(1 - \frac{T_{DC}}{\tau_c} + \frac{T_{DC}^2}{2\tau_c^2} - O\left(T_{DC}^3\right)\right)\right] \\ &\approx s_{xT}A_{DC}\frac{T_{DC}^2}{2\tau_c}\end{aligned} \tag{31}$$

$T_{DC}$ can be then expressed in terms of $s_{xT}$ as

$$\log_{10} T_{DC} = -sp - \frac{1}{2}\log_{10}|s_{xT}| \tag{32}$$

$$sp = \frac{1}{2}\log_{10}\frac{A_{DC}}{2|\Delta v_{outT}|\tau_c} \tag{33}$$

where $sp$ is used as an indicator of comparison speed. Given a fixed comparator output range, e.g., from 10% to 90% of the power supply rail, $sp$ is a function of both $A_{DC}$ and $\tau_c$. The simulated curves of $T_{DC}$ versus $s_{xT}$ of a CSA comparator biased at 0.2 nA, 2 nA and 20 nA are shown in Fig. 3. The red lines are the linear fits using (32) with $sp = 2.91$, 3.30 and 3.78, respectively.

In spike encoding simulations, we can use (32) to estimate the delay that corresponds to the signal slope $s_{xT}$ at each moment of threshold-crossing. When using nonlinear decoding to evaluate the degradation of spike representation in silicon sensors due to $T_{DC}$ variation, each delay $T_{D,k}$ associated with the spike at time $t_k$, however, cannot be estimated by (32) because of the lack of the $s_{xT}$ information. One simple way of recovering the input $x(t)$ is to estimate the average delay $T_{Davg}$ as a substitute for all $T_{D,k}$ assuming that some prior knowledge of the general characteristics of $x(t)$ such as its bandwidth and power are available. One can estimate the $T_{D,k}$ value iteratively from the slope of the reconstructed signal or adopt optimization methods by treating delay variations as random noise to have a better approximation of $x(t)$. One optimization example is presented in [19] where random noise is intentionally injected into the thresholds of the spike encoders. Nevertheless, the SDR improvement from using these methods is marginal, and the goal here is not to obtain the best reconstruction but rather to examine how the spike encoding quality is affected by circuit and system design parameters.

To estimate $T_{Davg}$, the average slope of $x(t)$ needs to be obtained first. The characteristics of $x(t)$ differ significantly in various sensing application scenarios. Here without preference for a specific application, we assume random Gaussian noise as the input $x(t)$ for the rest of the paper. The MATLAB codes that we developed for simulations in this work are provided in the supplementary materials online in which the Gaussian input and the related computation of average slope and delay can be modified to accommodate other signal types. The joint probability density function of Gaussian noise signal with amplitude $a_x$ and slope $s_x$ has been derived in [20]

$$p(a_x, s_x) = \frac{(-\psi_0\psi_0'')^{-1/2}}{2\pi}\exp\left(-\frac{a_x^2}{2\psi_0}+\frac{s_x^2}{2\psi_0''}\right) \tag{34}$$

where $\psi_0$ and $\psi_0''$ are the correlation function $\psi(\tau)$ and the second derivative of $\psi(\tau)$ at $\tau = 0$, respectively. The probability density function of $s_x$ is

$$p(s_x) = \int_{-\infty}^{\infty} p(a_x, s_x)da_x = \frac{(-\psi_0'')^{-1/2}}{\sqrt{2\pi}}\exp\left(\frac{s_x^2}{2\psi_0''}\right) \tag{35}$$

The average absolute slope $s_{xavg}$ of $x(t)$ is

$$s_{xavg} = 2\int_0^{\infty} |s_x|p(|s_x|)d|s_x| = \sqrt{\frac{2}{\pi}}(-\psi_0'')^{1/2} \tag{36}$$

$\psi''(\tau)$ is related to $\omega(f)$, the power spectrum of $x(t)$, by [21]

$$\psi''(\tau) = -\int_0^{\infty} (2\pi f)^2\omega(f)\cos(2\pi f\tau)df \tag{37}$$

For bandlimited white noise with zero mean, $\sigma^2$ power and $f_n$ bandwidth, $\psi_0''$ is calculated as

$$\psi_0'' = -\frac{4\pi^2\sigma^2}{f_n}\int_0^{f_n} f^2 df = -\frac{4}{3}(\pi\sigma f_n)^2 \tag{38}$$

$s_{xavg}$ can hence be calculated as

$$s_{xavg} = 2\sqrt{\frac{2\pi}{3}}\sigma f_n \tag{39}$$

Using (32), $T_{Davg}$ can be estimated as

$$T_{Davg} = 10^{-sp}s_{xavg}^{-1/2} \tag{40}$$

Equations (39) and (40) show that the parameters needed to compute $T_{Davg}$ are the power $\sigma^2$ and bandwidth $f_n$ of the noise signal, and the comparison speed indicator $sp$. Equations (32) and (40) are used later in Section V to calculate the exact and average comparison delay, respectively.

## IV. Modeling of Queueing Delay Variation $T_{DQ}$

Arbitration of spikes during transmission is of particular importance for sparse spike encoding schemes like ADM and STR, in contrast to the inefficient pulse-frequency modulation (PFM) scheme used in some early spiking sensors [22], [23]. In PFM, the amplitude of input signal is linearly transformed to spike frequency. Therefore, several missing spikes due to collision in an unfettered communication channel only have negligible impact on the average frequency within a time window. On the other hand, the SDR of reconstructed signals from spike trains encoded by ADM and STR directly relies on the presence and accurate timestamps of each generated spike. Arbitration in AER guarantees the transmission of every spike at the cost of compromised accuracy of timestamps. Because spikes are only recorded or processed on the receiving end, the timestamps are intact if the generated spikes pass through the AER without waiting. Otherwise the timestamps are skewed if the spikes need to wait until the transmission of earlier generated spikes is completed. This kind of system can be studied by queueing models, and the variation of queueing delay $T_{DQ}$ can be characterized by the waiting time distribution $P(W)$.

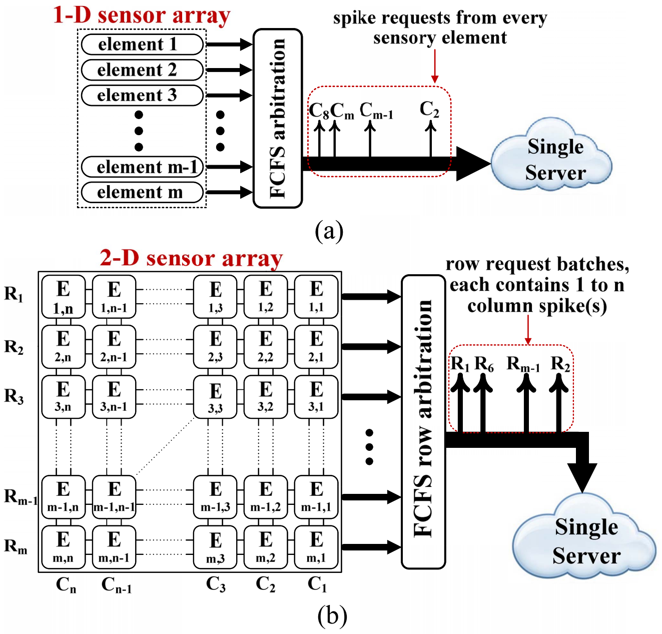

1-D sensor array
element 1
element 2
element 3
element m-1
element m
FCFS arbitration
spike requests from every sensory element
$C_8$ $C_m$ $C_{m-1}$ $C_2$
Single Server
(a)
2-D sensor array
$R_1$ $R_2$ $R_3$ $R_{m-1}$ $R_m$
E 1,n E 1,n-1 E 1,3 E 1,2 E 1,1
E 2,n E 2,n-1 E 2,3 E 2,2 E 2,1
E 3,n E 3,n-1 E 3,3 E 3,2 E 3,1
E m-1,n E m-1,n-1 E m-1,3 E m-1,2 E m-1,1
E m,n E m,n-1 E m,3 E m,2 E m,1
$C_n$ $C_{n-1}$ $C_3$ $C_2$ $C_1$
FCFS row arbitration
row request batches, each contains 1 to n column spike(s)
$R_1$ $R_6$ $R_{m-1}$ $R_2$
Single Server
(b)

Fig. 4. The queueing models of (a) 1-D sensors with $m$ sensing elements like a silicon cochlea, and (b) 2-D sensors with $m \times n$ sensing elements like a 2-D silicon retina.

The geometric form of integrated CMOS sensors can be categorized as either 1-D sensors like the silicon cochlea [6] and optical line sensors [24], or 2-D sensors like the silicon retina [2], [4] and neural recording arrays [25]. For 1-D spiking sensors as shown in Fig. 4(a), the spike encoder in each sensing element generates a stream of spike trains in response to the input, and the spike trains from all $m$ sensing elements are arbitrated according to the rule of first-come-first-serve (FCFS), which can be implemented as the 1-D version of the fair AER arbiter circuits [26]. Assuming an ideally fair arbiter, namely the arbiter does not distinguish the identities of different sensing elements and assign priorities, all spikes enter one queueing line abiding by the time order of their generation. All spikes are served by a single server. The service time is normally fixed when the server is a synchronous module, like an off-chip FPGA or an on-chip time-to-digital converter. When a sensor is exposed to a real-world stimulus, the spike arrival for the queue may form a particular stochastic process. In general, the queueing model G/D/1 can be used [27], where G indicates that the inter-spike interval of the arrival spike train has an arbitrary distribution, D means the deterministic service time, and 1 means a single server. Here we consider the Poisson arrival case to demonstrate the effect of spike queueing on encoding degradation. The corresponding queueing model is thus called M/D/1 where M stands for Poisson arrival.

To obtain the waiting time distribution $P(W)$ of an M/D/1 queue, two parameters, i.e. the mean spike arrival rate $\lambda_{q1D}$ and the service rate $\mu_q$ need to be determined. $\mu_q$ is the reciprocal of the service time which is the sum of the intrinsic delay of the AER circuit and the fixed time assigned for the spike address and timestamp registration. $\lambda_{q1D}$ depends on a specific 1-D sensor model and the input sensory stimulus. Taking a silicon cochlea with $m$ total sensing elements as the example, each sensing element or cochlea channel, contains a bandpass filter (BPF) with a central frequency of $f_i$ $(i \epsilon N^+ \cap i \epsilon [1, m])$ followed by a spike encoder with a threshold of $\delta_i$. The $f_i$ of all the BPFs are geometrically scaled with a ratio of $r$ between neighboring channels. With the pre-defined frequency range from $f_1 = f_L$ to $f_m = f_H$, $r$ can be calculated as

$$r = (f_H/f_L)^{\frac{1}{m-1}} \tag{41}$$

If the mean spike rate of channel $i$ is $\lambda_i$, $\lambda_{q1D}$ is the sum of all $\lambda_i$

$$\lambda_{q1D} = \sum_{i=1}^{m} \lambda_i \tag{42}$$

To calculate $\lambda_i$ of the bandpass-filtered white Gaussian noise input, its $\psi_0''$ is first calculated by using (37)

$$\begin{aligned}\psi_0'' &= -\frac{4\pi^2\sigma_i^2}{f_{Hi} - f_{Li}} \int_{f_{Li}}^{f_{Hi}} f^2 df \\ &= -\frac{4}{3}(\pi\sigma_i)^2(f_{Hi}^2 + f_{Hi}f_{Li} + f_{Li}^2)\end{aligned} \tag{43}$$

where $\sigma_i$ is the rms amplitude and $f_{Hi}$ and $f_{Li}$ are the highpass and lowpass corner frequencies of the BPF in channel $i$. $f_{Hi}$ and $f_{Li}$ satisfy the equations below

$$f_{Hi} - f_{Li} = f_i/Q_i, \quad f_{Hi} \cdot f_{Li} = f_i^2 \tag{44}$$

where $Q_i$ is the quality factor of the BPF in channel $i$. Together with (36), $\lambda_i$ of an ADM encoder [28] can be calculated as

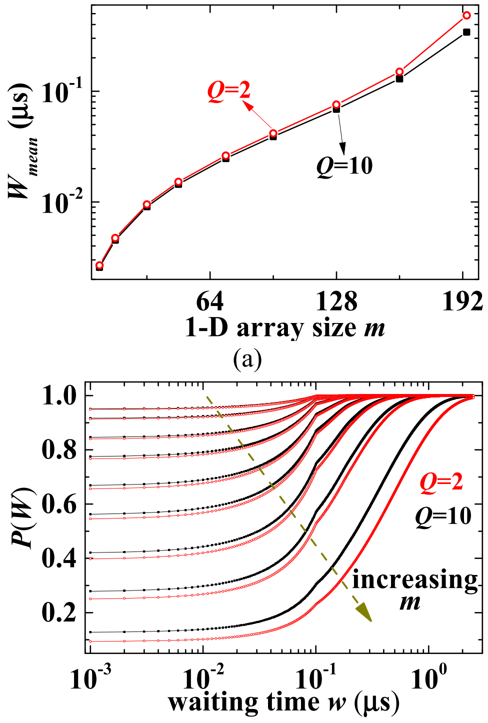
$$\lambda_i = 2\pi\frac{\sigma_i f_i}{\delta_i}\left[\frac{2}{3\pi}\left(3 + \frac{1}{Q_i^2}\right)\right]^{1/2} \tag{45}$$

Note that $\lambda_i$ of an STR encoder is usually smaller than that of an ADM encoder because of the signal loss during reset and switch-holding period [11], and the difference of these two $\lambda_i$ depends on the values of $T_{D,k}$ and $T_H$. For simplicity, the STR $\lambda_i$ is assumed to take the same value as the ADM with the consequence of underestimated reconstruction SDR. Further simplifying assumptions of identical $\sigma_i$, $\delta_i$, and $Q_i$ among all channels denoted as $\sigma$, $\delta$ and $Q$ give the explicit $\lambda_{q1D}$ expression

$$\lambda_{q1D} = \frac{2\sigma}{\delta}\left[\frac{2\pi}{3}\left(3 + \frac{1}{Q_i^2}\right)\right]^{1/2} \frac{f_m(1 - r^{-m})}{1 - r^{-1}} \tag{46}$$

TABLE II
PARAMETER VALUE LIST FOR COMPUTING $W_{mean}$ AND $P(W)$ IN SECTION IV

| Symbol | Description | Value |
|---|---|---|
| $\sigma$ | Signal rms amplitude | 0.382 |
| $\delta$ | Encoder threshold | 0.125 |
| $f_H$ | Highest frequency for cochlea | 20 kHz |
| $f_L$ | Lowest frequency for cochlea | 20 Hz |
| $f_{vis}$ | Pixel bandwidth for retina | 20 Hz |
| $\mu_q$ | Service rate or spike departure rate | $10^7$/s |
| $p$ | Probability of a retina pixel being active | 0.16 |



(b)

Fig. 5. (a) The mean waiting time $W_{mean}$ as a function of the channel number $m$ in the 1-D silicon cochlea and (b) the waiting time distribution $P(W)$ of the M/D/1 queue.

Knowing both $\lambda_{q1D}$ and $\mu_q$, the mean waiting time $W_{mean}$ [29] and the waiting time distribution $P(W)$ [30] of the M/D/1 queue can be computed by

$$W_{mean} = \frac{\rho}{2\mu_q(1-\rho)} \tag{47}$$

$$P(W \le w) = (1-\rho)\sum_{i=0}^{N} \frac{\left[-\lambda_{q1D}\left(w - \frac{i}{\mu_q}\right)\right]^i}{i!} e^{\lambda_{q1D}\left(w - \frac{i}{\mu_q}\right)}$$

$$\frac{N}{\mu_q} \le w < \frac{N+1}{\mu_q} \tag{48}$$

where $\rho = \lambda_{q1D}/\mu_q$ is the traffic intensity. With the parameter values given in Table II, $W_{mean}$ and $P(W)$ as a function of the number of channels $m$ are plotted in Fig. 5. As $m$ increases, the mean waiting time $W_{mean}$ increases exponentially, the sigmoid waiting time distribution $P(W)$ shifts towards larger waiting time with increasing variance, and the probability of immediate service with no delay decreases. Both $W_{mean}$ and $P(W)$ are slightly dependent on $Q$, and they will be later used for calculating the reconstruction SDR of a 1-D cochlea. The reason why a smaller $Q$ gives a larger delay can be seen from (46).

The latest 2-D spiking sensors with in-pixel ADM or STR encoders mostly use the 2-D burst-mode word-serial AER. As illustrated in Fig. 4(b), each sensing element in the $m \times n$ array is a pixel in the case of a silicon retina. Whenever one or several pixels in a row $R_i$ $(i \epsilon N^+ \cap i \epsilon [1, m])$ initiate spike transmission, $R_i$ enters the queue of row requests waiting for its row acknowledge from the single server. The fair arbitration of row requests also follows the FCFS rule. By the time the $R_i$ request is served, i.e. the row address and the row request timestamp are recorded, all the 'active' pixels in $R_i$ ('active' pixels are pixels which have generated a spike before the $R_i$ request is served) start to transmit their column addresses in a burst from column $C_1$ to $C_n$ and they all get the same timestamp as the $R_i$ request. After the column address of the last active pixel in $R_i$ is registered, the next row request waiting in the queue will get served, and the process repeats. Because an active pixel is not allowed to generate another spike before the current column request is acknowledged, the number of active columns in one row request cannot be larger than $n$.

The arrival pattern of row requests described above is called bulk arrival in queueing theory because each row request contains 1 to $n$ column requests, and the number of active columns $n_a$ is drawn from another independent distribution. The closest queueing model to describe such a 2-D sensor system is the $\mathrm{G}^{\mathrm{X}}$/D/1 where X denotes the bulk arrival. This model assumes that the delay for all the active columns is equal to the delay of the row request, which is only true if all the active columns of that row generate spikes at the same time. In practice, it is possible that the moment when a pixel becomes active is right before the burst column transmission, which may be later than its row request due to queue waiting, and consequently the delay of this spike is overestimated. In this sense, the $\mathrm{G}^{\mathrm{X}}$/D/1 model gives an upper bound of $T_{DQ}$ variation. In the specific case of a 2-D silicon retina, the $\mathrm{G}^{\mathrm{X}}$/D/1 model is more accurate because most visual stimuli are spatiotemporally correlated, and consequently the spikes in a batch tend to be elicited simultaneously and share the same timestamp. In the following analysis we consider the Poisson arrival of row requests, and the corresponding queueing model is $\mathrm{M}^{\mathrm{X}}$/D/1 with compound Poisson arrival [27].

Let $\lambda_{q2D}$ be the mean arrival rate of row requests in a 2-D sensor. Assuming that the spike generation and transmission in a 2-D sensor form a stationary queue, the equation below holds in light of conservation of spikes generated

$$\lambda_{q2D} \cdot n_{avg} = \lambda_{pixel} \cdot m \cdot n \cdot p \tag{49}$$

where $n_{avg}$ is the mean number of active pixels in a row request, $\lambda_{pixel}$ is the mean spike rate of a stimulated pixel, and $p$ is the probability of one pixel being stimulated. $n_{avg}$ was shown in [14] to be positively dependent on the traffic intensity because the spikes in a row request can be accumulated during its queue waiting. However, it will be shown later that this dependence is negligible in the specific case of a silicon retina where $\lambda_{pixel}$ and $p$ are small so that the spike accumulation is negligible. With the assumption of a binomial distribution of $n_a$ in a row request, the probability mass function of effective number of spikes in a row request, $p_a$, is written as

$$p_a = \binom{n-1}{a-3} p^{a-1}(1-p)^{n-a+2} \tag{50}$$

where $a$ has the support of $a \epsilon N^+ \cap a \epsilon [3, n+2]$. Note that $a$ is not from 1 to $n$ because the time needed to process a row request also needs to be considered and is about twice as the time needed to process a column request in the latest design [31]. $n_{avg}$ can now be written as

$$n_{avg} = \sum_{a=3}^{n+2} (a-2) p_a \tag{51}$$

For a bandlimited white Gaussian noise input, $\lambda_{pixel}$ of an ADM is derived from (39)

$$\lambda_{pixel} = \frac{s_{xavg}}{\delta} = 2\sqrt{\frac{2\pi}{3}} \frac{\sigma f_{vis}}{\delta} \tag{52}$$

where $f_{vis}$ is the visual signal bandwidth. $\lambda_{q2D}$ can thus be calculated with the knowledge of $n_{avg}$ and $\lambda_{pixel}$.

The mean waiting time $W_{mean}$ [32] and waiting time distribution $P(W)$ [33] of row requests of the $\mathrm{M}^{\mathrm{X}}$/D/1 queue can be computed as

$$W_{mean} = \frac{\rho}{2\mu_q(1-\rho)} \left( \frac{\sum_{a=3}^{n+2} a(a-1)p_a}{\sum_{a=3}^{n+2} a p_a} + 1 \right) \tag{53}$$

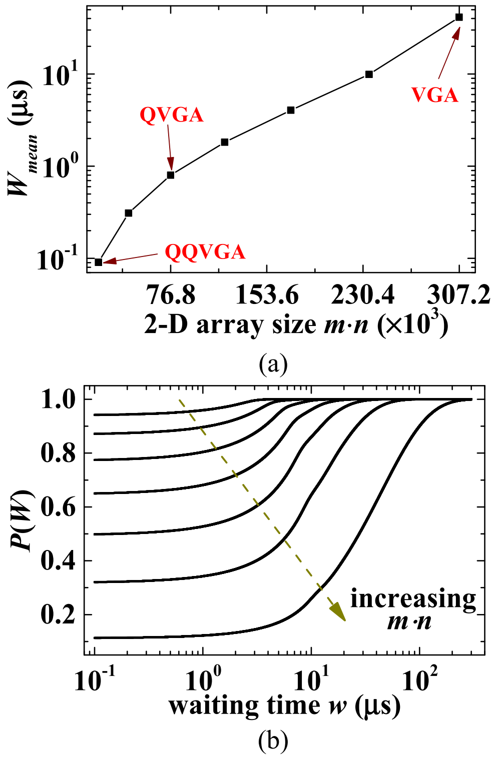
$$P(W \le w) = \sum_{k=0}^{N} \sum_{i=0}^{N+1-k} f_i h_k \left( \frac{N+1}{\mu_q} - w \right), \ \frac{N}{\mu_q} \le w < \frac{N+1}{\mu_q} \tag{54}$$

where $\rho$ is the traffic intensity defined as

$$\rho = \frac{\lambda_{q2D}}{\mu_q} \sum_{a=3}^{n+2} a p_a \tag{55}$$

$f_i$ denotes the probability of the stationary distribution with $i$ row requests being held in the system, and $h_k(t)$ denotes the probability that exactly $k$ new row requests enter the queue during an arbitrary time interval of length $t$ (see Appendix for the computation of $f_i$ and $h_k(t)$). With the parameter values given in Table II, $W_{mean}$ and $P(W)$ as a function of the total pixel number $m \cdot n$ are plotted in Fig. 6. An aspect ratio of $m$:$n$ = 3:4 is used. The popular video graphics array (VGA) resolution and its scaled versions are labeled in Fig. 6(a). Similar to the 1-D case, the mean waiting time $W_{mean}$ increases exponentially as the array size $m \cdot n$ increases, and so does the variance of the waiting time distribution $P(W)$. However, the maximum $W_{mean}$ value is about ×100 larger than the 1-D case with comparable traffic intensities (e.g., both around 0.88), which is attributed to bulk arrival. Note that a decreased $m/n$ aspect ratio would result in increased $W_{mean}$ even though $m \cdot n$ is kept constant.

We now verify an earlier statement below (49) that the dependence of $n_{avg}$ on the traffic intensity is negligible at



(a)

(b)

Fig. 6. (a) The mean waiting time $W_{mean}$ as a function of the number of pixels $m \cdot n$ in the 2-D silicon retina and (b) the waiting time distribution $P(W)$ of the $M^X$/D/1 queue.

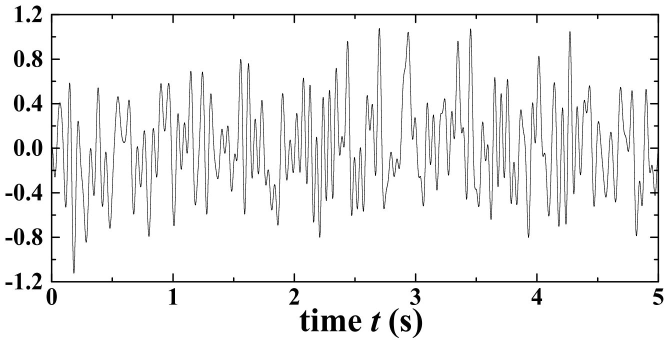


Fig. 7. White Gaussian noise used as the input $x(t)$ for evaluating encoding degradation caused by spike delay variation.

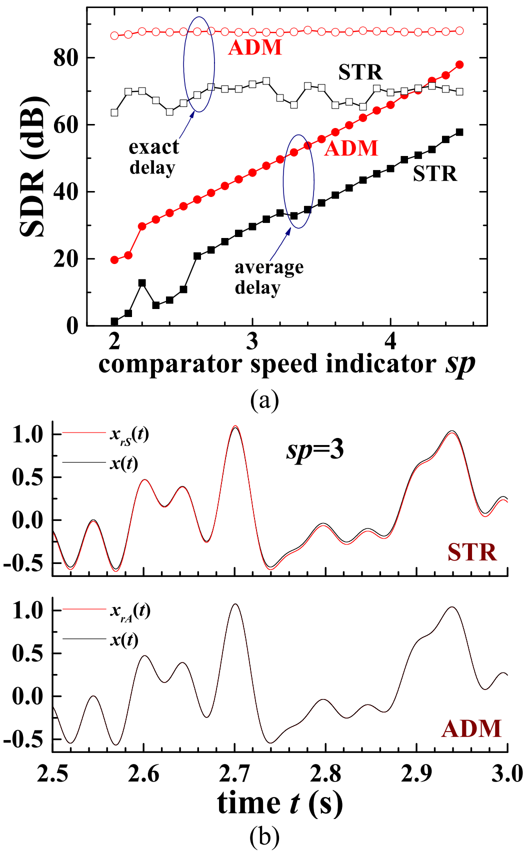
least in this silicon retina case we considered. The average accumulated spike number $N_{ac}$ can be calculated as

$$N_{ac} = \lambda_{pixel} \cdot n \cdot p \cdot W_D \tag{56}$$

where $W_D$ is the waiting delay before the requesting row gets served. In the case of a VGA array, the largest array size we have considered, we take a large $W_D$ of 200 $\mu$s which already has a very small chance to happen according to Fig. 6(b), and $N_{ac}$ is calculated to be 3.6. For VGA, $n_{avg}$ is calculated to be 103, much larger than $N_{ac}$.

## V. Encoding Degradation Due to $T_{DC}$ and $T_{DQ}$ Variation

Fig. 7 shows the white Gaussian noise waveform used as the input $x(t)$ to study the degradation of encoding quality of ADM and STR encoders due to $T_D$ variation. This input has zero mean, 0.382 variance, the same as the $\sigma$ in Table II, and 20 Hz bandwidth.



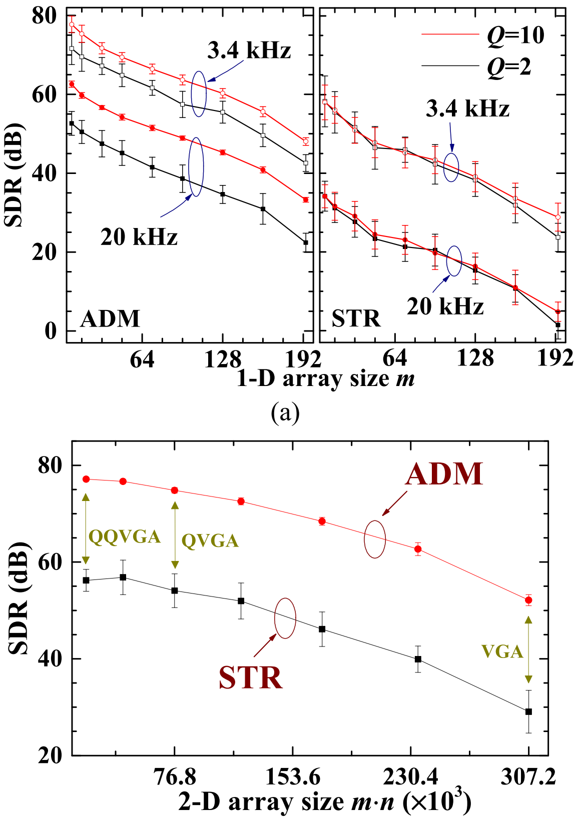
(a)

(b)

Fig. 8. (a) Reconstruction SDR of STR and ADM as the function of comparison speed indicator $sp$. Open and solid symbols represent reconstruction using $T_{D,k}$ and $T_{Davg}$, respectively; (b) example reconstructed signals at $sp = 3$ together with the original signal within the time window of [2.5s 3.0s]. Note that in the ADM case, the difference between the original and the recovered signals is indiscernible.

First, the effect of $T_{DC}$ variation is considered. The time resolution is set to 1 $\mu$s, and $T_H$ for STR is fixed at 1 ms in the simulations. The exact delay $T_{D,k}$ at each threshold-crossing is generated according to (32). The average delay $T_{Davg}$ is calculated using (40). Both $T_{D,k}$ and $T_{Davg}$ are used for reconstruction with the algorithms presented in Section II, and the reconstruction SDRs are compared as shown in Fig. 8(a). Both ADM and STR retain a high level of reconstruction SDR (88 dB and 69 dB in average) regardless of the value of the comparison speed indicator $sp$ when $T_{D,k}$ is used. When $T_{Davg}$ is used in practice with no prior knowledge of the slope of the input signal, both SDR values decrease by about 20 dB as $sp$ increases by 1, i.e. either the comparator time constant increases or the DC gain decreases by a factor of 100. Fig. 8(b) shows the example reconstructed waveforms. For ADM, the difference between the input and the reconstructed waveforms is indiscernible.

To study the effect of $T_{DQ}$ variation on the SDR degradation in the 1-D silicon cochlea, the waveform in Fig. 7 is scaled to a 5-ms time length with a bandwidth of 20 kHz as the input to the BPF bank in order to match with the human hearing frequency range. As an example, the input is filtered by two BPFs with central frequencies at $f_{c1} = 3.4$ kHz (telephony voice frequency band) and $f_{c2} = 20$ kHz (upper



(a)

(b)

Fig. 9. SDR degradation in (a) 1-D silicon cochlea as the function of cochlea channel number $m$ and (b) 2-D silicon retina as the function of retina array size $m \cdot n$.

limit of human hearing) and with different $Q$ of 2 and 10. The filtered signal is rescaled to have an amplitude variance of 0.382, the same as the $\sigma$ in Table II, before being encoded into spike train. The time resolution in simulation is 1 ns, and $T_H$ for STR is 5.9 $\mu$s for $f_{c1}$ and 1 $\mu$s for $f_{c2}$. The SDR degradation as a function of the total number of cochlea channels $m$ is plotted in Fig. 9(a). The error bar represents the SDR variance over 20 runs with the $T_{D,k}$ vector used in encoding sampled according to the $P(W)$ in Fig. 5(b) in each run. The $T_{Davg}$ used for reconstruction is obtained from Fig. 5(a). With $m$ increasing from 8 to 194, the SDR degrades by about 30 dB in all cases because of the increased variation in the waiting time distribution as evident in Fig. 5(b). The $f_{c2}$ case always has a worse SDR than the $f_{c1}$ case because signals with higher frequency components result in finer timescale of the encoded spike train and are therefore more susceptible to time quantization [15]. This implies that the time resolution in real-time spike timestamp registration using FPGAs or on-chip time-to-digital converters needs to be set carefully so that it does not become the limiting factor of the integrity of signal representation in spike domain. The high $Q$ case gives a higher SDR in ADM than the low $Q$ case, most likely due to the slightly smaller $T_{DQ}$ variation in light of Fig. 5(b). The less obvious benefit of a higher $Q$ in STR is the result of a much lower reconstruction SDR than that in ADM which results in masking the effect of the small difference in $T_{DQ}$ variation.

The same white Gaussian noise waveform as in Fig. 7 is used as the input to the spike encoders in the 2-D silicon retina case. The time resolution is 0.1 $\mu$s, and $T_H$ is 1 ms in the simulations. The $T_{D,k}$ vector used in encoding is drawn from the $P(W)$ in Fig. 6(b). The $T_{Davg}$ used for reconstruction is obtained from Fig. 6(a). The SDR degradation as a function of the array size $m \cdot n$ is plotted in Fig. 9(b). As $m \cdot n$ increases from the size of QQVGA to VGA, the SDR in both ADM and STR degrades by about 25 dB because of the increased variation in waiting time distribution as evident in Fig. 6(b). Note that in both the 1-D and 2-D cases, the traffic intensity is below 0.9. It is well-known in queueing theory that as the traffic intensity approaches 1, the delay continues to increase without bounds. Any spiking sensor should avoid operating in that regime for the sake of encoding integrity.

## VI. Conclusion and Discussions

This paper presents the first steps toward a quantitative analysis of the relationship between circuit and system design parameters and the performance of two different spike encoders used in spiking sensors. In particular, the effects of comparator speed and spike queueing on the encoding quality measured by the SDR metric with ADM and STR spike encoders are studied. The comparator speed is determined by its DC gain and bias current. The traffic intensity of spike queueing is directly related to the spike service time and the spike arrival rate. More specifically, besides the time needed for registration of spike addresses and timestamps, the spike service time is limited by the latency of the AER circuits. The spike arrival rate is determined by multiple factors, including the size of the sensor array as analyzed in Section V, the amplitude and bandwidth of the input signal, the threshold of the spike encoder, and the sparsity of the array activity indicated by the parameter $p$ in Table II. The analysis presented in this paper has implications in future designs of spiking sensors, especially in the context of internet of things which requires ultra-low-power sensing. The comparator speed and the communication channel bandwidth can be reduced to the minimal level where the required encoding quality specifications (e.g., measured by SDR) can still be satisfied or the system performance is limited by the noise of front-end analog circuits, and therefore the bias current of the comparators and the supply voltage of the AER circuits can be lowered to a certain degree to save system power.

The SDR metric used in this paper is directly relevant to faithful recording applications like optical neuroimaging [4] and electrical neural signal acquisition [25] where a small encoding error is of paramount importance. Compared to traditional clocked Nyquist sampling, asynchronous spike encoding schemes like ADM and STR have the advantage of reduced data redundancy in recording sparse signals, which is essential in minimizing RF transmission power for wireless sensors. For emerging smart sensing systems with low-power embedded processing for within-sensor classification and recognition [34], SDR may not be the best measure and further study is needed to establish the link between signal encoding quality and system performance in terms of classification or recognition accuracy. Note that even though the signal rms is only about three times the encoder threshold (Table II), the nonlinear decoding can still deliver high reconstruction SDR in both ADM and STR (>50 dB with small timing jitter). This suggests that information is preserved in the precise timestamps of sparse spike trains with low average spike rate, and further reduction in data redundancy might be possible in contrast to high-amplitude-resolution sampling, which is beneficial for both wireless signal acquisition and spike-based processing regarding system power consumption and processing latency.

Even though the results in this paper are obtained based on white Gaussian noise input, quantitative analysis with different types of input signals that closely model the statistics of real-world signals in various sensing application scenarios will be pursued in the future using the presented methods.

## Appendix

This section describes the method of computing $f_i$ and $h_k(t)$ in (54) based on the fully probabilistic analysis described in [33]. Recall that $h_k(t)$ denotes the probability that exactly $k$ new row requests enter the queue during an arbitrary time interval of length $t$. If $j$ row requests arrive within $t$, the probability $\varphi_k(j)$ of these $j$ row requests containing $k$ spikes can be derived by recursion (note that one row request is equivalent to two column spikes because twice the handshake time is needed)

$$\varphi_k(1) = p_k(k \in N^+ \cap k \in [3, n+2]) \qquad (57)$$

$$\varphi_k(j) = \sum_{a=3}^{k-3(j-1)} p_a \varphi_{k-a}(j-1), \quad (k \geq 3j) \qquad (58)$$

$$\varphi_k(j) = 0, \quad (k < 3j) \qquad (59)$$

where $p_k$ is the probability mass function of spike number in a row request. Assuming that the number of row request arrival within time $t$ is Poisson distributed, $h_k(t)$ can be written as

$$h_k(t) = \sum_{j=1}^{\infty} \varphi_k(j) \frac{(\lambda_{q2D} t)^j}{j!} e^{-\lambda_{q2D} t} \qquad (60)$$

Let $f_i(t_0)$ denote the probability of the system holding $i$ row requests at time $t_0$. By conditioning on the number of spikes present at $t_0$, the equation below holds

$$f_i(t_0 + D) = \sum_{l=0}^{1} f_l(t_0) h_i(D) + \sum_{l=2}^{i+1} f_l(t_0) h_{i+1-l}(D), \quad (i \in N_0) \qquad (61)$$

where $D = 1/\mu_q$ is the deterministic service time. The stationary distribution $f_i$ is found by letting $t_0 \to \infty$

$$f_i = \sum_{l=0}^{1} f_l h_i(D) + \sum_{l=2}^{i+1} f_l h_{i+1-l}(D), \quad (i \in N_0) \qquad (62)$$

To solve $f_i$ in the equation above, the geometric tail method described in [35] is used. For $i \geq M$ where $M$ is a sufficiently large positive integer, $f_i$ is approximated as

$$f_i = f_M \varepsilon^{M-i}, \quad (i \geq M) \qquad (63)$$

The scaling factor $\varepsilon$ can be solved by setting the denominator of the probability generating function of the $M^X$/D/1 queue [36] to 0

$$1 - \varepsilon \cdot \exp\left[-\frac{\lambda_{q2D}}{\mu_q}\left(\sum_{a=3}^{n+2} p_a \varepsilon^a - 1\right)\right] = 0 \qquad (64)$$

Equation (62) can then be written as

$$f_i = \sum_{l=0}^{1} f_l h_i(D) + \sum_{l=2}^{i+1} f_l h_{i+1-l}(D), \quad (i < M) \qquad (65)$$

which is an $M$-dimensional linear equation system. The normalization equation is written as

$$\sum_{i=0}^{M-1} f_i + f_M \sum_{i=M}^{\infty} \varepsilon^{M-i} = 1 \Rightarrow \sum_{i=0}^{M-1} f_i + \frac{f_M}{1 - \varepsilon^{-1}} = 1 \qquad (66)$$

Now an $(M+1)$-dimensional linear equation system is complete for $M+1$ variables $f_i$, $i \in [0, M]$. In the numerical simulations to obtain the data to plot Fig. 6(b), $M = 200$ is used.

## Acknowledgement

The authors acknowledge Prof. Aurel Lazar for helpful discussions on the $M^X$/D/1 queue, Dr. Yiyin Zhou for comments on the manuscript, and the anonymous reviewers for valuable suggestions.

## References

[1] D. A. Butts *et al.*, "Temporal precision in the neural code and the timescales of natural vision," *Nature*, vol. 449, no. 7158, pp. 92–95, 2007.

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

[3] T. Serrano-Gotarredona and B. Linares-Barranco, "A 128 × 128 1.5% contrast sensitivity 0.9% FPN 3$\mu$s latency 4 mW asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers," *IEEE J. Solid-State Circuits*, vol. 48, no. 3, pp. 827–838, Mar. 2013.

[4] M. Yang, S.-C. Liu, and T. Delbruck, "A dynamic vision sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding," *IEEE J. Solid-State Circuits*, vol. 50, no. 9, pp. 2149–2160, Sep. 2015.

[5] S.-C. Liu, A. Van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous binaural spatial audition sensor with 2 × 64 × 4 channel output," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 4, pp. 453–464, Aug. 2014.

[6] M. Yang, C.-H. Chien, T. Delbruck, and S.-C. Liu, "A 0.5 V 55$\mu$W 64 × 2-channel binaural silicon cochlea for event-driven stereo-audio sensing," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb. 2016, pp. 388–389.

[7] W. Tang *et al.*, "Continuous time level crossing sampling ADC for bio-potential recording systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 6, pp. 1407–1418, Jun. 2013.

[8] L. C. Gouveia, T. J. Koickal, and A. Hamilton, "An asynchronous spike event coding scheme for programmable analog arrays," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 4, pp. 791–799, Apr. 2011.

[9] C. Weltin-Wu and Y. Tsividis, "An event-driven clockless level-crossing ADC with signal-dependent adaptive resolution," *IEEE J. Solid-State Circuits*, vol. 48, no. 9, pp. 2180–2190, Sep. 2013.

[10] H. Akolkar *et al.*, "What can neuromorphic event-driven precise timing add to spike-based pattern recognition?" *Neural Comput.*, vol. 27, no. 3, pp. 561–593, Jan. 2015.

[11] M. Yang, S.-C. Liu, and T. Delbruck, "Comparison of spike encoding schemes in asynchronous vision sensors: Modeling and design," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2014, pp. 2632–2635.

[12] M. Koyanagi *et al.*, "Neuromorphic vision chip fabricated using three-dimensional integration technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2001, pp. 270–271.

[13] K. A. Boahen, "A burst-mode word-serial address-event link—I: Transmitter design," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 7, pp. 1269–1280, Jul. 2004.

[14] K. A. Boahen, "A burst-mode word-serial address-event link—III: Analysis and test results," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 7, pp. 1292–1300, Jul. 2004.

[15] A. A. Lazar and L. T. Toth, "Perfect recovery and sensitivity analysis of time encoded bandlimited signals," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 10, pp. 2060–2073, Oct. 2004.

[16] A. A. Lazar and E. A. Pnevmatikakis, "Video time encoding machines," *IEEE Trans. Neural Netw.*, vol. 22, no. 3, pp. 461–473, Mar. 2011.

[17] A. Papoulis, *Signal Analysis*. New York, NY, USA: McGraw-Hill, 1977.

[18] P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design*, 3rd ed. Oxford, U.K.: Oxford Univ. Press, 2011.

[19] A. A. Lazar, E. A. Pnevmatikakis, and Y. Zhou, "Encoding natural scenes with neural circuits with random thresholds," *Vis. Res.*, vol. 50, no. 22, pp. 2200–2212, Oct. 2010.

[20] S. O. Rice, "Mathematical analysis of random noise," *Bell Syst. Tech. J.*, vol. 24, no. 1, pp. 46–156, 1945.

[21] S. O. Rice, "Mathematical analysis of random noise," *Bell Syst. Tech. J.*, vol. 23, no. 3, pp. 282–332, 1944.

[22] A. Mortara and E. A. Vittoz, "A communication architecture tailored for analog VLSI artificial neural networks: Intrinsic performance and limitations," *IEEE Trans. Neural Netw.*, vol. 5, no. 3, pp. 459–466, May 1994.

[23] E. Culurciello, R. Etienne-Cummings, and K. A. Boahen, "A biomorphic digital image sensor," *IEEE J. Solid-State Circuits*, vol. 38, no. 2, pp. 281–294, Feb. 2003.

[24] C. Posch *et al.*, "A dual-line optical transient sensor with on-chip precision time-stamp generation," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 500–501.

[25] J. N. Y. Aziz *et al.*, "256-channel neural recording and delta compression microsystem with 3D electrodes," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 995–1005, Mar. 2009.

[26] K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 47, no. 5, pp. 416–434, May 2000.

[27] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. New York, NY, USA: McGraw-Hill, 2002.

[28] R. Steele, *Delta Modulation Systems*. London, U.K.: Pentech Press Ltd., 1975.

[29] R. Cahn, *Wide Area Network Design: Concepts and Tools for Optimization*. San Francisco, CA, USA: Morgan Kaufmann, 1998.

[30] G. J. Franx, "A simple solution for the M/D/c waiting time distribution," *Oper. Res. Lett.*, vol. 29, no. 5, pp. 221–229, Dec. 2001.

[31] R. Berner, "Building-blocks for event-based vision sensors," Ph.D. dissertation, Dept. Elect. Eng. Inf. Technol., ETH Zurich, Zürich, Switzerland, 2011.

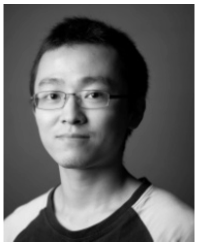
[32] M. L. Chaudhry and U. C. Gupta, "Exact computational analysis of waiting-time distributions of single-server bulk-arrival queues: $\mathrm{M}^X$/G/1," *Eur. J. Oper. Res.*, vol. 63, no. 3, pp. 445–462, Dec. 1992.

[33] G. J. Franx, "The $\mathrm{M}^X$/D/c batch arrival queue," *Probab. Eng. Inf. Sci.*, vol. 19, no. 3, pp. 345–349, Jul. 2005.

[34] K. M. H. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "A 90 nm CMOS, 6 $\mu$W power-proportional acoustic sensing frontend for voice activity detection," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 291–302, Jan. 2016.

[35] H. C. Tijms, *Stochastic Models: An Algorithmic Approach*. New York, NY, USA: Wiley, 1995.

[36] M. L. Chaudhry and J. G. C. Templeton, *First Course in Bulk Queues*. New York, NY, USA: Wiley, 1983.



**Minhao Yang** (S'11-M'16) received the Ph.D. degree in physics from ETH Zurich, Switzerland, in 2015. He is currently a Postdoctoral Fellow with Columbia University, New York, NY, USA, funded by SNF Early Postdoc Mobility Fellowship. His current research interests include spike coding and processing, low-power spiking sensors with embedded processing, and silicon retina and cochlea.

**Shih-Chii Liu** (M'02–SM'07) received the B.S. degree in electrical engineering and the Ph.D. degree in the computation and neural systems program from the California Institute of Technology, Pasadena, CA, USA, in 1997.

She worked at various companies, including Gould American Microsystems, LSI Logic, and Rockwell International Research Labs. She is currently a group leader at the Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland. Her current research interests include neuromorphic visual and auditory sensors, cortical processing circuits, and event-driven circuits, networks, and algorithms.

Dr. Liu was the Chair of the IEEE CAS Sensory Systems and Neural Systems and Applications Technical Committees. She is currently the Chair of the IEEE Swiss CAS/ED Society and an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMSand of *Neural Networks*.

**Tobi Delbruck** (M'99–SM'06–F'13) received the B.Sc. degree in physics and applied mathematics from the University of California, San Diego, CA, USA, and the Ph.D. degree from the California Institute of Technology, Pasadena, CA, USA, in 1986 and 1993, respectively. He has been a Professor of Physics with the Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland, since 1998. His group focuses on neuromorphic sensory processing. He worked on electronic imaging at Arithmos, Synaptics, National Semiconductor, and Foveon.

Dr. Delbruck has co-organized the Telluride Neuromorphic Cognition Engineering summer workshop and the live demonstration sessions at International Symposium on Circuits and Systems. He is also co-founder of iniLabs and Insightness. He was the Chair of the IEEE CAS Sensory Systems Technical Committee, is current Secretary of the IEEE Swiss CAS/ED Society, and an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS. He has received 9 IEEE awards.