Quantification of a spike-based winner-take-all VLSI network

Oster, M; Yingxue, W; Douglas, R; Liu, S-C

Quantification of a spike-based winner-take-all VLSI network

Abstract

We describe a formalism for quantifying the performance of spike-based winner-take-all (WTA) VLSI chips. The WTA function non-linearly amplifies the output responses of pixels/neurons dependent on the input magnitudes in a decision or selection task. In this work, we show a theoretical description of this winner-take-all computation which takes into consideration the input statistics, neuron response variance, and output rates. This analysis is tested on a spiking VLSI neuronal network fabricated in a 4-metal, 2-poly 0.35\,μm CMOS process. The measured results of the winner-take-all performance from this chip correspond to the theoretical prediction. This formalism can be applied to any implementation of spike-based neurons.

Quantification of a Spike-Based Winner-Take-All VLSI Network

Matthias Oster, Yingxue Wang, *Member, IEEE*, Rodney Douglas, and Shih-Chii Liu, *Senior Member, IEEE*

***Abstract*—We describe a formalism for quantifying the performance of spike-based winner-take-all (WTA) VLSI chips. The WTA function nonlinearly amplifies the output responses of pixels/neurons dependent on the input magnitudes in a decision or selection task. In this paper, we show a theoretical description of this WTA computation which takes into consideration the input statistics, neuron response variance, and output rates. This analysis is tested on a spiking VLSI neuronal network fabricated in a 4-metal, 2-poly 0.35–μm CMOS process. The measured results of the WTA performance from this chip correspond to the theoretical prediction. This formalism can be applied to any implementation of spike-based neurons.**

***Index Terms*—Analog integrated circuits, event-based systems, neuromorphic engineering, winner-take-all circuits.**

I. Introduction

THE winner-take-all (WTA) function is a commonly used nonlinear function that can be used in any computation that requires a decision. It is a key component in prototypical very-large-scale-integration (VLSI) systems for stereo computation, auditory processing, location of salient objects in space, and also for identification of objects [1]–[9]. The WTA circuits were first implemented using analog VLSI circuits [10]–[14] and are recently also implemented using hybrid analog-digital circuits [15]–[18]. The WTA function is intrinsic to many computational models in neuroscience [19], for example, in modeling attention and recognition processes in the cortex [20]–[22]. It is thought to be a basic function of the cortical microcircuit [23].

The recent construction of various asynchronous event-based multichip VLSI systems where components consist of large-scale networks of spiking neurons and spike-based sensors usually include the WTA operation in their computation [9], [24]–[26]. It is desirable to have an underlying theoretical basis for setting up the connectivity and neuron parameters based on the input statistics so that a network will have the most optimal performance as a WTA. The theoretical constraints of the parameter space for a spiking WTA system have so far only been studied for analog inputs [27].

Manuscript received February 3, 2007; revised August 21, 2007. First published April 18, 2008; current version published November 21, 2008. This work was supported in part by a EU-Grant IST-2001-34124, FP6-2005-015803, and by ETH Research Grant TH-20/04-2. This paper was recommended by Associate Editor A. van Schaik.

The authors are with the Institute of Neuroinformatics, University of Zürich and ETH Zurich, CH-8057 Zurich, Switzerland (e-mail: shih@ini.phys.ethz.ch).

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

Digital Object Identifier 10.1109/TCSI.2008.923430

In this paper, we describe a methodology for setting up the network parameters of a spiking network using spiking inputs dependent on the input statistics [28]. In addition, we also quantify the effect of the intrinsic mismatch across the pixels in a fabricated chip in determining the performance of the WTA network. We introduce a new measure, *the increase factor*, which quantifies the percentage by which the input rate to a neuron should increase, for this neuron to be the winning neuron in a population. This measure indirectly accounts for the pixel circuit variances which lead to variances in the neuron responses. We show different ways of estimating the mean increase factor of a population of neurons on a single chip without resorting to the time-consuming method of measuring the increase factor for every neuron.

We compare our theoretical predictions to experimental measurements performed on a multineuron chip that was fabricated as part of an asynchronous multichip, multilayered, spike-based vision system (CAVIAR) that classifies spatiotemporal trajectories in the scene [9], [29]. The chip receives preprocessed feature map inputs and its role is to identify the locations of the features in each map and to identify the best matching feature in the scene. To perform this function, the chip does a WTA operation in two different dimensions; the pixel space and the feature space [30]. In this paper, we focus only on the WTA analysis.

The chip measurements of the WTA performance are close to the theoretical predictions based on the measured chip variances. The proposed theoretical paradigm is not specific to our chip and can be used for quantifying the performance of any spike-based multineuron network in the WTA operation.

II. WTA Connectivity

We assume a network of integrate-and-fire neurons that receives excitatory or inhibitory spiking input through synaptic connections. To implement a WTA operation, these neurons compete through inhibition. In biological networks, excitation and inhibition are specific to the neuron type. Excitatory neurons make only excitatory connections to other neurons and inhibitory neurons make only inhibitory connections. Inhibition between the array neurons is always mediated by populations of inhibitory interneurons [Fig. 1(a)]. The inhibitory neurons are driven by the excitatory neurons, and in return they inhibit the excitatory neurons.

To adjust the amount of inhibition between the neurons (and thereby the strength of the competition), both types of connections could be modified: the excitatory connections from array neurons to interneurons and the inhibitory connections from interneurons to array neurons. In our analysis, we assume the forward connections between the excitatory and the inhibitory neurons to be strong, so that each spike of an excitatory neuron

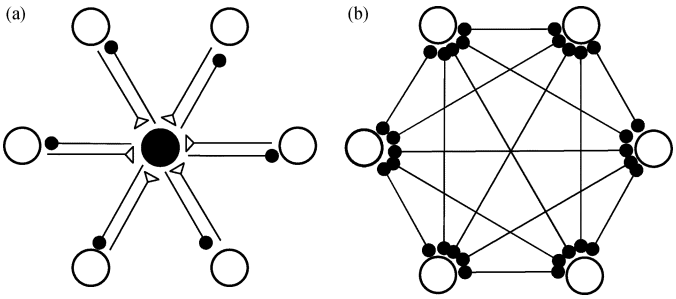

Fig. 1. Simplification of connectivity: (a) with a global inhibitory neuron; (b) with direct inhibitory connections between neurons.

triggers a spike in the global inhibitory neurons. The amount of inhibition between the array neurons is adjusted by tuning the connections from the global inhibitory neurons to the array neurons. This configuration allows the fastest spreading of inhibition through the network and is consistent with findings in biology that the inhibitory interneurons tend to fire at high frequencies.

With this configuration, we can simplify the network by replacing the global inhibitory interneurons with full inhibitory connectivity between the array neurons [Fig. 1(b)]. This simplification is only used during the analysis; the configuration with the global interneurons is simpler for the implemented form.

III. Network Connectivity Constraints for a Hard WTA

One important question is how we would program the synapse and neuron parameters so that the network operates as a WTA network. While the quantitative analysis of the parameter space in regards to analog neurons have been described before, the extension of such analysis to spiking neurons receiving spiking inputs have only recently been described [17], [28], [31]. The constraints on the synaptic connectivity matrix that allow the network to detect the most optimal winner is dependent on the statistics of the input spikes and we summarize the constraints only for regular input frequencies below.

The membrane potentials $V_i, i = 1 \ldots N$, satisfy the equation of a nonleaky integrate-and-fire neuron model with n excitatory and m nonconductance-based synapses

$$\frac{\mathrm{d}V_i}{\mathrm{d}t} = V_E \sum_n \delta(t - t_i^{(n)}) - V_I \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_m \delta(t - s_j^{(m)}). \quad (1)$$

The membrane resting potential is set to 0. Each neuron receives external excitatory input and inhibitory connections from all other neurons. All inputs to a neuron are spikes and its output is also transmitted as spikes to other neurons. We neglect the dynamics of the synaptic currents and the delay in the transmission of the spikes. Each input spike causes a fixed discontinuous jump in the membrane potential (V_E for the excitatory synapse and V_I for the inhibitory synapse). Each neuron i spikes when $V_i \geq V_{\mathrm{th}}$ and is reset to $V_i = 0$. Immediately afterward, it receives a self-excitation of weight V_{self}. All potentials satisfy $0 \leq V_i \leq V_{\mathrm{th}}$, that is, an inhibitory spike cannot drive the membrane potential below ground. All neurons $i \in 1 \ldots N, i \neq k$, receive excitatory input spike trains of constant frequency r_i. Neuron k receives the highest input frequency ($r_k > r_i \; \forall i \neq k$).

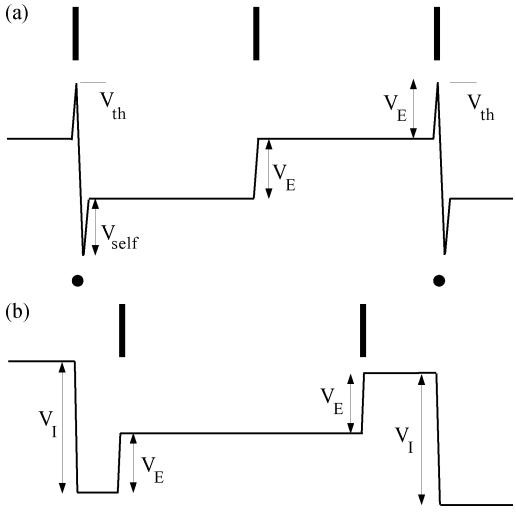

Fig. 2. Membrane potential of the winning neuron k (a) and another neuron in the array (b). Black bars show the times of input spikes. Traces show the changes in the membrane membrane potential caused by the various synaptic weights. Black dots show the times of output spikes of neuron k.

As soon as neuron k spikes once, it has won the computation. Depending on the initial conditions, other neurons can at most have a few transient spikes before the first spike of neuron k. For this hard WTA mode, the network has to fulfill the following constraints (see Fig. 2):

(a) Neuron k (the winning neuron) spikes after receiving $n_k = n$ input spikes that cause its membrane potential to exceed threshold. After every spike, the neuron is reset to V_{self}

$$V_{\mathrm{self}} + n_k V_E \geq V_{\mathrm{th}}. \quad (2)$$

(b) As soon as neuron k spikes once, no other neuron $i \neq k$ can spike because it receives an inhibitory spike from neuron k. Another neuron can receive up to n spikes even if its input spike frequency is lower than that of neuron k because the neuron is reset to V_{self} after a spike, as illustrated in Fig. 2. The resulting membrane voltage has to be smaller than before

$$n_i \cdot V_E \leq n_k \cdot V_E \leq V_I. \quad (3)$$

(c) If a neuron j other than neuron k spikes in the beginning, there will be some time in the future when neuron k spikes and becomes the winning neuron. From then on, the conditions (a) and (b) hold, so a neuron $j \neq k$ can at most have a few transient spikes.

Let us assume that neurons j and k spike with almost the same frequency (but $r_k > r_j$). For the inter-spike intervals $\Delta_i = 1/r_i$ this means $\Delta_j > \Delta_k$. Since the spike trains are not synchronized, an input spike to neuron k has a changing phase offset ϕ from an input spike of neuron j. At every output spike of neuron j, this phase decreases by $\Delta\phi = n_k(\Delta_j - \Delta_k)$ until $\phi < n_k(\Delta_j - \Delta_k)$. When this

happens, neuron k receives $(n_k + 1)$ input spikes before neuron j spikes again and crosses threshold

$$(n_k + 1) \cdot V_E \geq V_{\text{th}}. \tag{4}$$

We can choose $V_{\text{self}} = V_E$ and $V_I = V_{\text{th}}$ to fulfill the inequalities (2)–(3). V_E is adjusted to achieve the desired n_k.

Case (c) happens only under certain initial conditions, for example when $V_k \ll V_j$ or when neuron j initially received a spike train of higher frequency than neuron k. A leaky integrate-and-fire model will ensure that all membrane potentials are discharged ($V_i = 0$) at the onset of a stimulus. The network will then select the winning neuron after receiving a predetermined number of input spikes and this winner will have the first output spike.

Due to the fact that the analysis takes into consideration only spike times, the mechanism of competition is independent of the number of neurons in a network. The network can be scaled to any size, as long as the inhibitory neuron can still completely inhibit the array neurons with one output spike. Other models that exploit firing rate thresholds are normally dependent on the number of neurons in the network.

The performance of the network decreases also in the case of inputs of Poisson statistics instead of inputs of regular rates [28].

In the case of perfect network homogeneity, the network can detect the winner optimally. In Section V, we discuss how variation in the synaptic parameters influences the performance of the network.

IV. Chip Architecture

The WTA network is implemented on a transceiver VLSI array of 32×32 spiking integrate-and-fire neurons that was fabricated in a 4-metal, 2-poly, 0.35-μm CMOS process. The chip can be configured into four populations of 256 neurons (shown in Fig. 3) or a single population of 1024 neurons. This chip is the first component in the CAVIAR multichip vision system to make a decision on the incoming spike inputs [9]. The inputs and outputs of the neurons are communicated using an asynchronous event-based transmission protocol called address-event representation (AER) [24], [32]–[38]. On-chip neurons and synapses are labeled by unique addresses. The AER encoder and decoder blocks encode the spike addresses from the active neurons and decode the input spikes to the neurons, respectively. Two additional signals, /CREQ and /CACK, along with the address bus are used to communicate when the address lines are valid between a sender chip/module and a receiver chip/module.

We have extended the functionality of the AER bus by using the AER address space to carry the bits for 2 on-chip global digital-to-analog converters (DACs). The DACs allow us to program a local weight for each synapse before the synapse is stimulated. The turnaround time for sending an AER address varies from 500 ns to 1 μs compared to the time constant of 1 ms of the neurons thus allowing us to send two AER events for each synaptic stimulation.

We describe the key circuit blocks on the chip excluding the AER encoders and decoders.

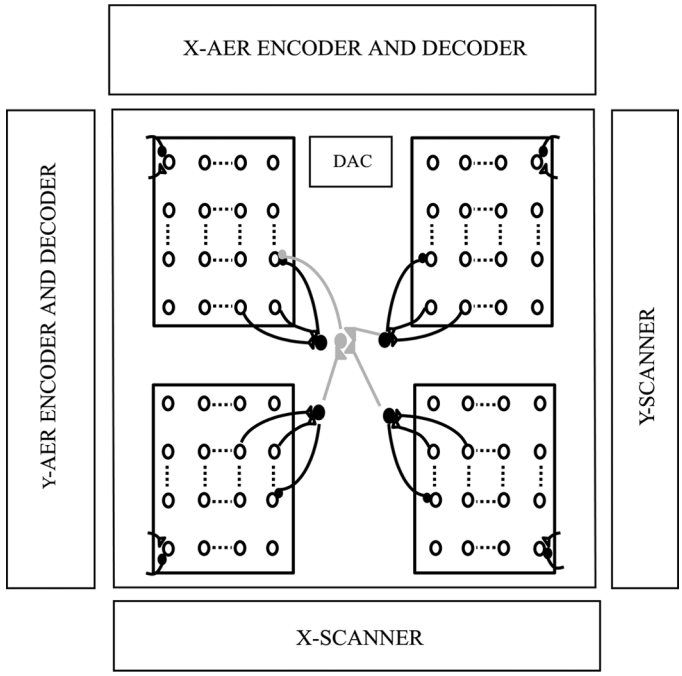

Fig. 3. Architecture of the four populations on the chip configured for competition within a feature map and across feature maps. Each population of neuron consists of 254 excitatory neurons and two inhibitory neurons. In the WTA analysis described here, the four populations are combined as one population of 1024 neurons. The excitatory neurons receive external AER inputs through four input synapses which are either excitatory, depressing, or inhibitory. In addition, each excitatory neuron has two sets of local synapses which form the connections to and from the global inhibitory neuron and a self-excitatory local synapse. These synapses can be activated without using the AER infrastructure. These neurons drive a global inhibitory neuron (solid black circle) which in return inhibits all excitatory neurons. The additional inhibitory neuron colored gray (feature inhibitory neuron) is excited by the first global inhibitory neuron of all populations and in return, it inhibits all excitatory neurons in its own population. This configuration was used in the CAVIAR system to determine the best matching feature in the scene [30]. The DAC block generates a local weight for the stimulated synapse. The spiking activity of the neurons are monitored through the addresses on the AER bus while an on-chip scanner allow us to monitor the membrane potentials of the neurons externally.

A. Programmable Local Synaptic Weights

The DAC block consists of two global current-mode 5-bit DAC. These converters are used to set the weights of individual synapses. This circuit is based on the current-splitter technique [Fig. 4(b)] and uses the masterbias circuit [Fig. 4(a)] which generates bias currents on chip [39], [40]. The DAC output current is generated by the circuit in Fig. 5(a). This circuit is designed to work over five decades of current from 10 pA to 1μA.

The bits to the current splitter in [Fig. 4(b)] are set through the AER input address space. The DAC AER address is sent first and decoded by the DAC before the AER address of the targeted synapse is sent next (Fig. 6). We set the relative time between the DAC AER address and the AER synapse address to be about 5 μs. This time is sufficient for the DAC to settle before the synapse is stimulated with the efficacy that is set by the DAC output. The DAC can be used to both compensate and program the weights of the synapses locally [41], [42] and globally [43].

B. Neuron and Synapses

The DAC output goes to one of two types of synapses: excitatory and inhibitory. The circuit for the excitatory synapse is

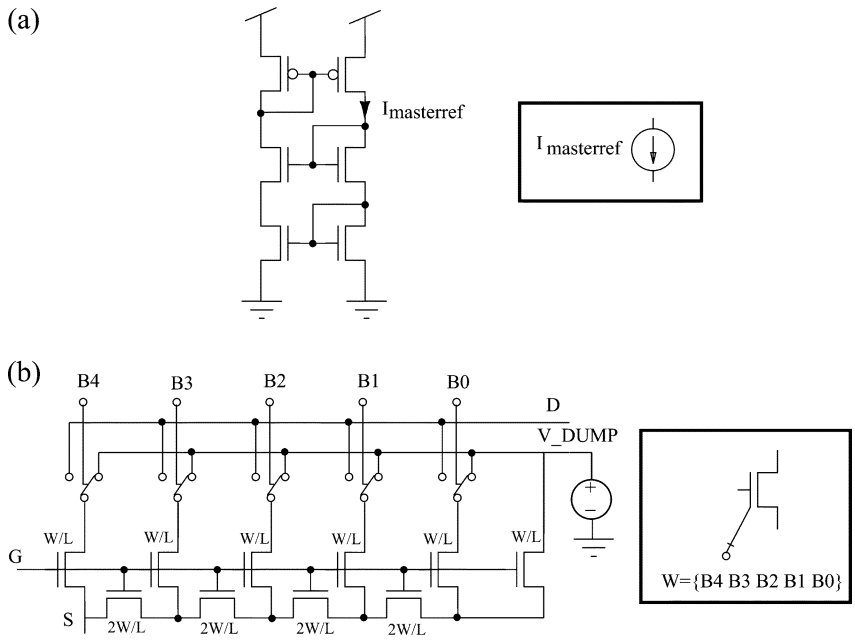

Fig. 4. Circuits for (a) the master reference current generator which is connected to (b) the current splitter. Depending on the value of the bits, the appropriate current is generated for the DAC current output circuit shown in Fig. 5(a). The symbols for these circuits are shown on the right.

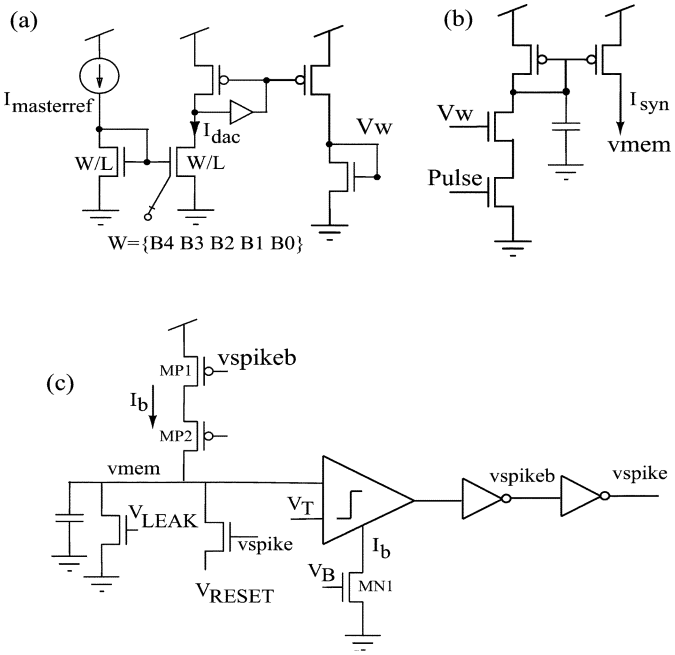

Fig. 5. DAC current output, synapse, and soma circuit. (a) The output of the DAC circuit that generates the local synaptic weight for the targeted synapse. (b) The synapse circuit where its weight is set by Vw. Transistors are 12 λ by 12 λ. (c) Soma circuit where $\mathrm{V_{RESET}}$ is the reset potential. Transistors MP1 and MP2 provide positive feedback current when a spike is generated, that is, $\mathrm{vspikeb}$ goes low and vspike goes high.

shown in Fig. 5(b). In addition, there are two different forms of the excitatory synapse: depressing or nondepressing. These circuits have previously been described in [44], [45] and [34], [46].

On this chip, we implemented an aVLSI integrate-and-fire neuron model which follows the dynamics in (1) [Fig. 5(c)]. In contrast with the theory, our neuron circuit has a linear leak.

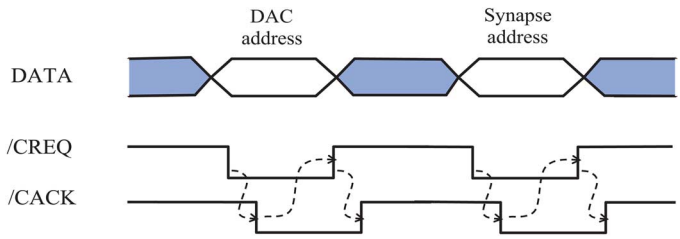

Fig. 6. Timing diagram for AER signals to stimulate a synapse using a 4-phase handshaking. When the DATA is valid, the sender activates the a request signal, /CREQ to the chip. The chip responds by activating the corresponding acknowledge signal /CACK, and the the DAC address defining the synaptic weight is latched internally and decoded. The sender inactivates /CREQ when /CACK is active and removes the data from the bus. It waits for /CACK to be inactivated before placing the synapse address on the bus.

Since this constant leak current does not depend on the membrane voltage, we can neglect it for the analysis of regular input rates.

The controllable parameters in the neuron circuit include the threshold voltage, refractory period, reset potential, spike pulse width, and leak current. We have modified the neuron circuit so we can satisfy the theoretical constraints laid out in Section III. This circuit is based on the original form described in [46] but includes a reset potential. It is derived from other implementations [18], [45]–[47]. Alternative neuron implementations are described in [48]–[50].

The size of each neuron pixel with its AER and local synapses, along with the AER communicating circuits is 86 μm by 140 μm.

C. Connectivity Setup

To implement the WTA operation, we activate the local connections between the excitatory neurons and the global inhibitory neuron. The synaptic connectivity for this configuration

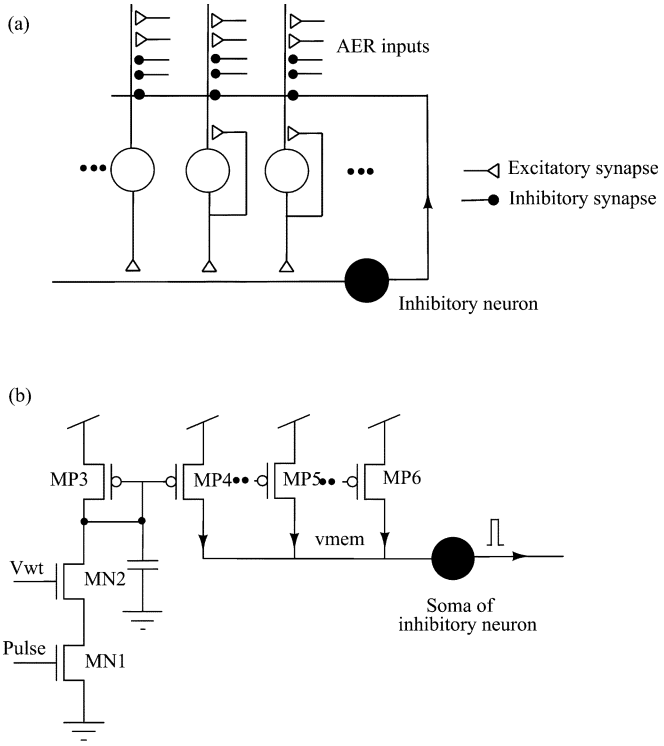

Fig. 7. Local connectivity between excitatory neurons and global inhibitory neuron. In (a), each excitatory neuron (open circle) has an inhibitory connection (small black circles) from the inhibitory neuron and makes an excitatory connection (triangle) to the inhibitory neuron. The schematic in (b) shows the transistor circuit (MN1, MN2, MP3, MP4) for one excitatory synapse. The output transistor of the synapse, MP4, along with the output transistors of the same type of synapse of other neurons (MP5, MP6 represent output synapses of two other neurons) charge the membrane potential of the inhibitory neuron.

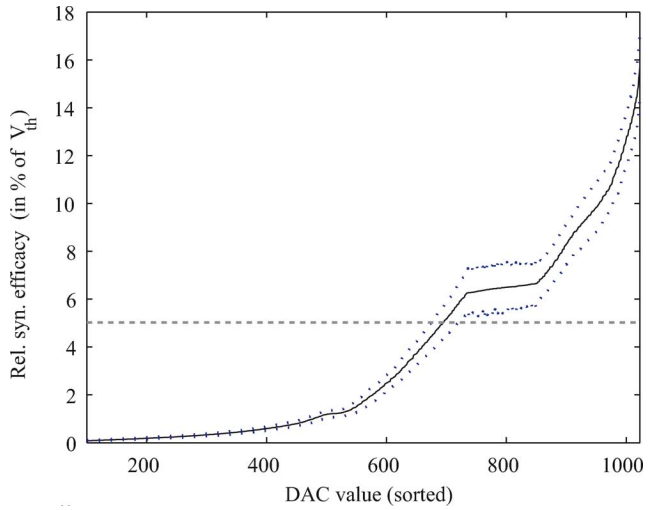

Fig. 8. Mismatch of DACs and synaptic weight of the neuron. For each measured DAC value, the relative synaptic efficacy is shown, that is the effect of one input spike on the membrane potential in percent of the threshold voltage of the neuron. For example, a relative synaptic efficacy of 5% means that a neuron reaches threshold in 20 spikes. We combined both 5-bit DACs into one 10-bit converter and sorted the DAC values by the mean synaptic efficacy to obtain a continuous value range. Every DAC value was measured for all neurons at once for a stimulation time of 90 s using a regular spiking input with frequency of 80 Hz. The graph shows the mean of all neurons with one standard deviation (dotted lines). We showed in [28] that the coefficient of variation for this type of synapse is constant, i.e., the standard deviation scales with the mean of the synaptic weight, as can be seen for DAC values up to about 720. For high relative synaptic weights, e.g., over 10%, the neurons synchronize on the input spikes, since a neuron can only reach threshold with an integer number of spikes. In this range the standard deviation of the synaptic weight, therefore, decreases again.

is shown in Fig. 7. The winner is selected after a predetermined number of input spikes according to the constraints of the connectivity parameters needed for the WTA function [31]. Each excitatory input spike charges the membrane of the postsynaptic neuron until one neuron in the array reaches threshold after the predetermined number of input spikes and is reset. This neuron then drives the inhibitory neuron which in return inhibits all other neurons. Self-excitation of the winning neuron (by making the reset potential higher than the resting potential) facilitates the selection of this neuron as the next winner. The winning neuron and the global inhibitory neuron in each population code the input activity, that is, their output rates are proportional to the input rates.

V. VLSI Chip WTA Performance

The soma and synaptic circuits express a certain amount of heterogeneity because of inherent nonidealities in the fabrication process. This means that the firing rates of neurons on a fabricated chip, even when stimulated with the same input rate, will vary because of transistor mismatch [51]. While the amount of mismatch can be decreased, for example, through increased sizing of the transistors, the larger transistors will also increase the pixel size. This heterogeneity changes the WTA network operation because different neurons will require different factors of increase in their firing rates over the remaining neurons to become the winner. In a previous generation of this chip, we used the spike burst encoding method [17] to decrease the variation from about 20% in the uncalibrated case to an average of 10%. On the present chip, the variance in the uncalibrated case is only 10% primarily due to increased sizing of the transistors.

The on-chip DACs allow us to implement different coefficient of variations (CVs) into the synaptic weights so we can explore the dependence of the WTA performance on the variance of the parameters. In this case, we used the two 5-bit DACs as a single 10-bit DAC and set the reference currents of both converters so that their ranges overlap. In this way, we can sort the DAC values to obtain a continuous range (see Figs. 8 and 9).

To measure the performance of the WTA network we define the *increase factor* c_k. If all neurons of the network receive the same input rate, the increase factor c_k defines the percentage by which the input rate to one neuron k has to be increased for this neuron to be the winner. As we discussed in Section III, the network can theoretically detect the highest input frequency in two spikes. In most applications, the firing rates are not necessarily regular or show some variation due to the preceding computation. In this case, more spikes should be integrated before the neurons reach threshold. The performance on chip is limited by the variation in the network parameters, which we quantify using c_k. We first present measurements of the increase factor for all neurons, and then discuss how the mean increase factor can be estimated from the measured firing rate distribution from all neurons on the chip and from its variance.

With variation in the network parameters and the input spike rates, the performance of the WTA network will actually depend

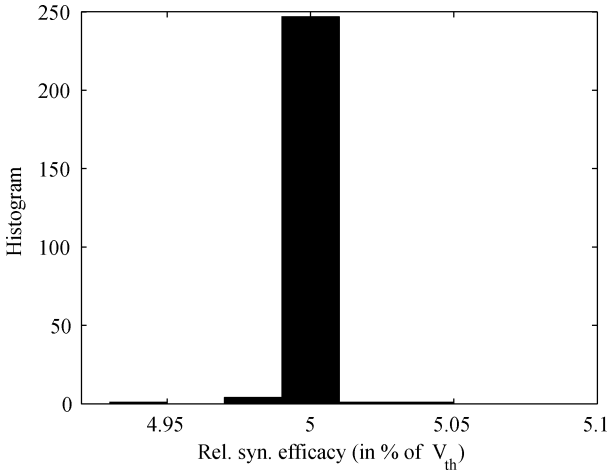

Fig. 9. Example histogram where the synaptic efficacies for all neurons are set to 5%, so every neuron reaches threshold with 20 spikes. For 99% of the neurons, a DAC value can be chosen so that the synaptic efficacy is within 0.1% of the desired value. The maximum difference is 1.1%. Our method of choosing the DAC values is equivalent to selecting the DAC values for the neurons for which the relative synaptic efficacies are on a horizontal line in Fig. 8.

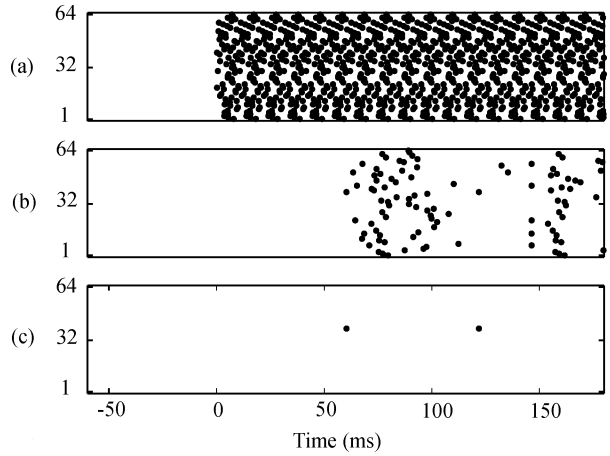

Fig. 10. Example raster plot of the spike trains to and from the neurons. (a) Input: starting from 0 ms, the neurons are stimulated with spike trains of a regular frequency of 100 Hz, but randomized phase. Neuron number 42 receives an input spike train with an increased frequency of 120 Hz. (b) Output without WTA connectivity: after an adjustable number of input spikes, the neurons start to fire with a regular output frequency. The output frequencies of the neurons are slightly different due to mismatch in the synaptic efficacies. Neuron 42 has the highest output frequency since it receives the strongest input. (c) Output with WTA connectivity: only neuron 42 with the strongest input fires, all other neurons are suppressed. The example plot was taken with a previous version of the chip with 64 neurons per array. The activity of the inhibitory neurons is not shown here. Figure is reproduced from [31, Fig. 4].

on the precise distribution of input firing rates to the chip. Since this input distribution varies for every application (for example, in the CAVIAR network described later the input can be approximated by a Gaussian peak of activity traveling across the chip), one input pattern has to be defined for characterization. The increase factor assumes that all neurons receive the same input spike rate, and that only the firing rate to the winner is increased. This is the worst-case input distribution, since in an application not all neurons will receive the same maximum firing rate. The mean increase factor as we present it here gives, therefore, a worst-case boundary for the performance of the chip, and in applications with natural input distributions the performance can be assumed to be significantly better. In [28], we analyze how the case of spike trains with regular frequency can be generalized to spike trains of Poisson statistics.

Since measuring the increase factor for every neuron is time-consuming, the latter methods provide a quick estimation of the WTA performance. A comparison between the measured increase factors and the expected mean value allows us also to incorporate effects from mismatch different than the synaptic weights as discussed before.

A. WTA Measurements

We first demonstrate the WTA operation using the analysis described in Section III. The network behavior is illustrated using a spike raster plot in Fig. 10. At time $t = 0$, the neurons receive inputs with the same regular firing frequency of 100 Hz except for one neuron which receives a higher input frequency of 120 Hz. The connectivity was configured in this experiment so that the neuron reaches threshold in five to six input spikes, after which the network selects the neuron with the strongest input as the winner. In the experiments, we have set the synaptic time constants to be short so that the currents coming in for each presynaptic spike act almost as impulse currents. This setting also reduces the impact of the variations in the time constants.

B. Measurement of Increase Factors

A group of 254 excitatory neurons were stimulated with an input rate of 80 Hz using their excitatory synapse whose weight is set by the global DACs. We used spike trains of regular frequency, but shuffled the phases to avoid synchronization effects. The synaptic weights were adjusted so that the mean output firing rate of all neurons was about 4 Hz, that means the neurons reach threshold with 20 spikes.

For each neuron k we increased its input frequency until the neuron was selected as the winner. We describe a neuron as the winner if at least 90% of the output spikes of the network comes from this neuron. Each measurement was taken for a stimulation period of 60 s.

The measured increase factors are shown in Fig. 11 and in a cumulative graph in Fig. 12 (circles). In these data, the excitatory synapses of all neurons were set to a DAC value of 700. The mean measured increase factor is about 38.7%.

C. Estimation From Firing Rate Distribution

If all neurons receive the same input spike frequency, the neuron with the highest synaptic weight will have the highest output firing rate $r_{\max}$ and will win the competition. To select a different neuron k as the winner, the input rate to this neuron has to be increased until its output firing rate exceeds that of the winner. We, therefore, estimate the increase factor for each neuron k as the normalized difference of its output rate from the maximum output firing rate of the array $r_{\max}$. To select neuron k as winner, its input rate has to be increased by at least

$$c_k = \frac{r_{\max} - r_k}{r_k} \cdot 100\%. \tag{5}$$

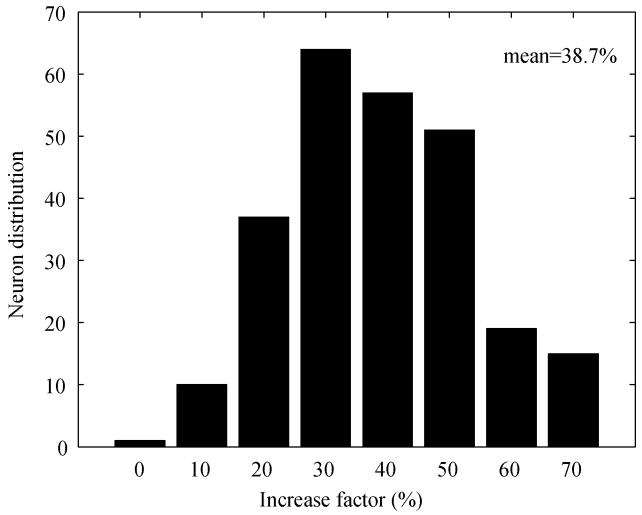

Fig. 11. Discrimination capability of the WTA network. The increase factor is the percentage by which the input firing rate to a neuron has to be increased for this neuron to be selected as the winner, assuming that the remaining neurons receive the same input rate. A neuron is considered the winner when more than 90% of the output spikes of the network originate from that neuron. Data was obtained from a group of neurons on the chip, where all excitatory synapses of all neurons were set to a DAC value of 700. Stimulation with regular input trains with frequency of 80 Hz, measurement time was 60 s.

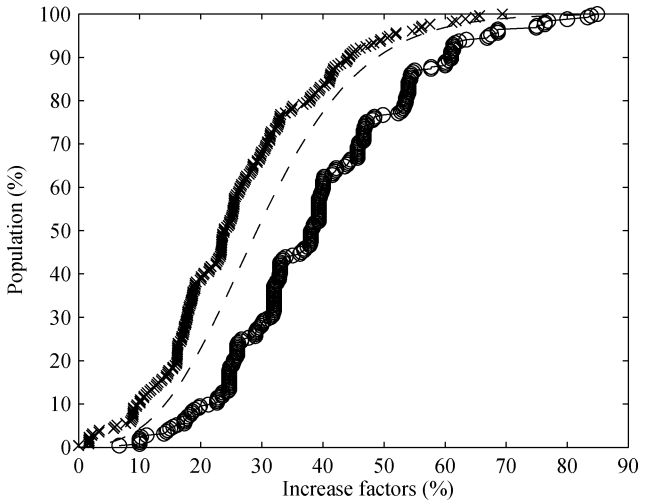

Fig. 12. Performance of chip WTA computation. We show the measured increase factors (circles), the increase factors estimated from the firing rate distribution of the chip (crosses) and from the variation of firing rates assuming a normal distribution (dashed line) as described in Sections V-C and -D, respectively. Shown is the cumulative graph of increase factors, with the percentage of the population (y axis) where the neurons will be the winner as a function of the increase factor (x axis). The difference to the maximum firing rate (crosses) shows the minimum boundary increase factor on which additional effects are added, so the curve of the measured increase factors (circles) is shifted to the right. Both curves fit quite well to the assumption of a normal distribution of synaptic weights (dashed line). The staircase shape of the measured increase factors is caused by the measurement method, since we gradually increased the input frequency to the test neuron until it was selected as the winner.

We measured the distribution of output firing rates using the same setting values as described in Section V-B (see Fig. 13). The measured increase factors in Section IV-B are on average 14% higher than the ones calculated here from the maximum

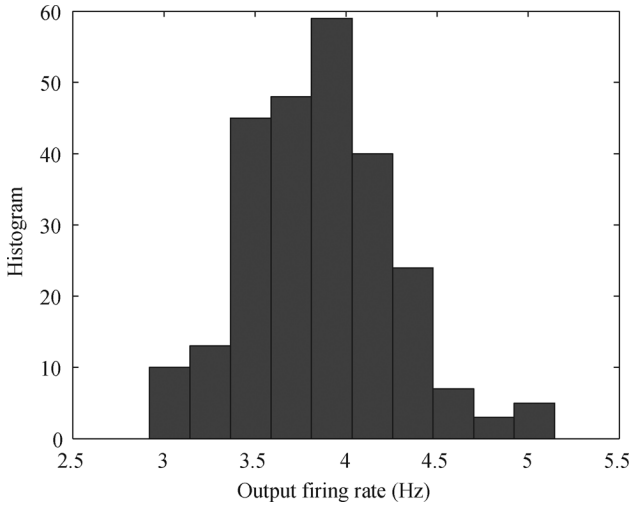

Fig. 13. Distribution of output firing rates. Data from one group of neurons on the chip. Each neuron was driven by the input through its excitatory synapse whose weight was set to a DAC value of 700. Stimulation with inputs of regular frequency of 80 Hz, measurement time 60 s.

firing rate of the output rate distribution. There are several reasons for this. First, the neurons have to overcome the maximum firing rate $r_{\max}$ which means the increase factor has to be higher than the calculated one. Second, the measurement of output firing rates only to determine the WTA operation does not consider mismatch from other circuits, for example, in the connections to and from the inhibitory neuron, the mismatch in the refractory period of the neurons, and the length of inhibition.

D. Estimation From Variation

The maximum firing rate can also be estimated from the variation of the measured firing rate distribution. Let $g(r)$ describe the distribution of firing rates of the neuron population (or equivalently, the distribution of synaptic efficacies). We draw N samples from this distribution. For one sample, the expected mean is obtained by integrating the rate r times the probability of this rate: $\int_{-\infty}^{\infty} g(r) r \, \mathrm{d}r$. We are not looking for the mean, but for the maximum. For one sample to be the maximum, all other $N-1$ samples have to be smaller than its value $r_{\max}$. The probability for this is

$$\left(\int_{-\infty}^{r_{\max}} g(r) \mathrm{d}r \right)^{N-1}. \tag{6}$$

Taken together we get

$$\langle r_{\max} \rangle = N \int_{-\infty}^{\infty} \left(\int_{-\infty}^{r} g(r) \right)^{N-1} g(r)\, r \, \mathrm{d}r. \tag{7}$$

Multiplying by N considers that each of the drawn samples can become the maximum.

We evaluate the expected maximum with the assumption of normal distributed firing rates for $g(r)$, see Fig. 14. Replacing r_k in (5) with the mean output firing rate results in the expected mean increase factor based on the variation of the firing rates. We show the increase factors for a normal distribution in Fig. 12 (dashed line). The expected mean increase factor is about 26%.

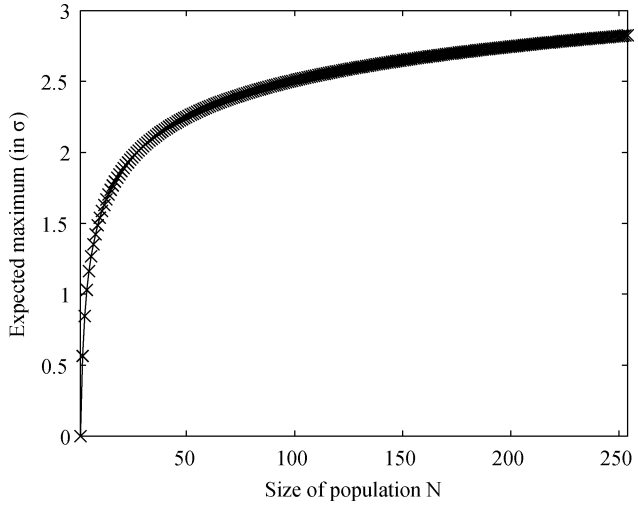

Fig. 14. Expected maximum value versus population size N. We assume the firing rates **r** to be Gaussian distributed. Shown is the difference of the maximum firing rate $r_{\max}$ and the mean firing rate μ_r of the population, normed to the standard deviation of the distribution. For example, for $N = 254$, the expected maximum firing rate is $\langle r_{\max} \rangle = \mu_r + 2.824\sigma$.

Estimating the mean increase factor directly from the variation of the firing rate distribution has the advantage that no time-consuming measurements are necessary on a particular chip. The coefficient of variation can be assumed to be equal for all samples of a chip run. The performance of the WTA can then be estimated from the characterization of one reference sample.

We conclude that the VLSI spiking WTA network performs according to the discussed constraints. The comparison we show here provides a good test case for quantifying effects from different mismatch sources. Since only the timing information of the spike trains is used, the results can be extended to a wide range of input frequencies other than the 80 Hz used in the presented measurements.

VI. Discussion

Demonstrations of multichip asynchronous spike-based systems in performing brain-like tasks like orientation selectivity, saliency detection, and pattern classification show the promise of asynchronous brain-like computing technology [7], [9], [25], [52]. The WTA function is usually a key component in these systems or an intrinsic operation in the multineuron modules.

To program such systems for optimal performance in a task, a formal description of each module's performance based on the input statistics, network parameters, and the output statistics is necessary. We described the technical constraints on the connectivity of a spiking neuronal network in order to achieve the most optimal performance based on regular spiking input statistics in this paper. The results for Poisson statistics are described elsewhere [28]. Our theoretical analysis of the network parameter setting does not assume a regular sampling system and does not consider spike rates. It is centered on spike numbers and spike times thus making this analysis extendable for any size network.

These technical constraints guided us in setting the network parameters of this chip in the CAVIAR system, a multichip spike-based vision system that classifies spatiotemporal trajectories [9], [42], [53], [54].

To assemble such novel spike-based systems for optimal performance of a particular task, specifications that describe the performance of the individual modules are necessary. Here, we define a new measure for measuring the WTA performance which takes into account the variance of the responses of the neurons due to circuit mismatch. We show two different ways of estimating the mean increase factor and compare the measured increase factors against the estimated mean values. The estimated increase factors provide a quick way of measuring the WTA performance of all chips of a particular design and from a particular fabrication run without having to resort to the time-consuming step of testing all neurons on all chips.

Our analysis considered the worst case input to the network, in which all neurons of the network receive a high regular input firing rate. In real-world applications, the input is not necessarily present to all neurons at the same time; for example, the inputs to the CAVIAR WTA module turned out to be a spatially moving Gaussian blob [29]. In this case, the network will show better performance, that is, it will select the winner with higher probability since fewer neurons receive input. Our methodology for setting the network connectivity and our proposed quantification of the WTA performance using theoretical increase factors, can be applied towards the performance estimation of any spike-based WTA implementation.

VII. Conclusion

We describe a methodology for quantifying the WTA performance of a spike-based aVLSI network and the theoretical constraints for the network parameters so that we obtain the most optimal performance from the network in terms of the input statistics and the output rates. We introduce a new measure, the increase factor, which quantifies the measured WTA performance and which takes into account the variances in the neurons responses. The measured performance of a chip is close to the theoretical prediction from the measured CV of the firing rate distribution derived from measuring one chip sample. This analysis can be extended to any spike-based multineuron chip used in a WTA computation.

Acknowledgment

The authors acknowledge members of the Institute of Neuroinformatics hardware group especially P. Lichtsteiner and T. Delbrück, the University of Sevilla, for the development of the software and hardware infrastructure, and the Institute of Microelectronics, Sevilla, especially B. Linares-Barranco and R. Serrano-Gotarredona. They also thank J. Rolfe for the discussion on the expected maximum firing rate.

References

[1] M. Mahowald and T. Delbrück, "An analog VLSI implementation of the Marr-Poggio stereo correspondence algorithm," in *Proc. Int. Neural Netw. Soc.*, Boston, MA, 1988.

[2] C. Mead, X. Arreguit, and J. Lazzaro, "Analog VLSI model of binaural hearing," *IEEE Trans. Neural Netw.*, vol. 2, pp. 230–236, 1991.

[3] M. Mahowald, *An Analog VLSI System for Stereoscopic Vision*. Boston, MA: Kluwer, 1994.

[4] T. Morris, T. Horiuchi, and S. DeWeerth, "Object-based selection within an analog VLSI visual attention system," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 45, no. 12, pp. 1564–1572, 1998.

[5] T. Horiuchi and E. Niebur, "Conjunction search using a 1-D, analog VLSI-based attentional search/tracking chip," in *Conf. Adv. Res. VLSI*, Atlanta, GA, Mar. 20–24, 1999.

[6] G. Indiveri, R. Mürer, and J. Kramer, "Active vision using an analog VLSI model of selective attention," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 48, no. 5, pp. 492–500, 2001.

[7] R. J. Vogelstein, U. Mallik, G. Cauwenberghs, E. Culurciello, and R. Etienne-Cummings, "A multichip neuromorphic system for spike-based visual information processing," *Neural Comput.*, vol. 19, pp. 2281–2300, 2007.

[8] T. Horiuchi, "A neural model for sonar-based navigation in obstacle fields," in *Proc. 2006 IEEE Int. Symp. Circuits Syst.*, Island of Kos, Greece, May 21–24, 2006, ISCAS 2006, pp. 4543–4546.

[9] R. Serrano-Gotarredona, M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gomez-Rodriguez, H. Riis, T. Delbrück, S.-C. Liu, S. Zahnd, A. M. Whatley, R. Douglas, P. Häfliger, G. Jimenez-Moreno, A. Civit, T. Serrano-Gotarredona, A. Acosta-Jimenez, and B. Linares-Barranco, "AER building blocks for multilayer multichip neuromorphic vision systems," in *Adv. Neural Inf. Process. Syst.*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, vol. 18, pp. 1217–1224.

[10] J. Lazzaro, S. Ryckebusch, M. A. Mahowald, and C. A. Mead, "Winner-take-all networks of $O(n)$ complexity," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1989, vol. 1, pp. 703–711.

[11] A. Andreou, K. Boahen, P. Pouliquen, A. Pavasovic, R. E. Jenkins, and K. Strohbehn, "Current-mode subthreshold MOS circuits for analog VLSI neural systems," *IEEE Trans. Neural Netw.*, vol. 2, pp. 205–213, 1991.

[12] J. Starzyk and X. Fang, "CMOS current mode winner-take-all circuit with both excitatory and inhibitory feedback," *Electron. Lett.*, vol. 29, no. 10, pp. 908–910, 1993.

[13] G. Indiveri, "Modeling selective attention using a neuromorphic analog VLSI device," *Neural Comput.*, vol. 12, no. 12, pp. 2857–2880, 2000.

[14] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, 2000.

[15] G. Indiveri, T. Horiuchi, E. Niebur, and R. Douglas, F. Rattay, Ed., "A competitive network of spiking VLSI neurons," in *Proc. World Congress on Neuroinformatics*, Vienna, Austria, Sep. 24–29, 2001, ARGESIM Rep., pp. 443–455.

[16] J. P. Abrahamsen, P. Häfliger, and T. S. Lande, "A time domain winner-take-all network of integrate-and-fire neurons," in *IEEE Int. Symp. Circuits Syst.*, Vancouver, Canada, May 25–28, 2004, vol. V, ISCAS '04, pp. 361–364.

[17] M. Oster and S.-C. Liu, "A winner-take-all spiking network with spiking inputs," in *Proc. 11th IEEE Int. Conf. Electron., Circuits Syst. (ICECS '04)*, Tel Aviv, Israel, Dec. 13–15, 2004.

[18] G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Trans. Neural Netw.*, vol. 17, pp. 211–221, 2006.

[19] S. Amari and M. Arbib, "Competition and cooperation in neural nets," in *Systems Neuroscience*, J. Metzler, Ed. New York: Academic, 1977, pp. 119–165.

[20] C. Itti, E. Niebur, and C. Koch, "A model of saliency-based fast visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, Apr. 1998.

[21] D. Lee, C. Itti, C. Koch, and J. Braun, "Attention activates winner-take-all competition among visual filters," *Nature Neuroscience*, vol. 2, pp. 375–381, Apr. 1999.

[22] M. Riesenhuber and T. Poggio, "Models of object recognition," *Nature Neurosci.*, vol. 3, pp. 1199–1204, 2000.

[23] R. Douglas and K. Martin, "Cortical microcircuits," *Ann. Rev. Neurosci.*, vol. 27, no. 1f, pp. 419–451, 2004.

[24] M. Mahowald, "VLSI analogs of neuronal visual processing: A synthesis of form and function," Ph.D., Calif. Inst. Technol., Neural Syst. Dep., Pasadena, 1994.

[25] T. Y. W. Choi, P. Merolla, J. Arthur, K. Boahen, and B. E. Shi, "Neuromorphic implementation of orientation hypercolumns," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 52, no. 6, pp. 1049–1060, 2005.

[26] U. Mallik, R. Vogelstein, E. Culurciello, R. Etienne-Cummings, and G. Cauwenberghs, "A real-time spike-domain sensory information processing system," in *Proc. 2005 IEEE Int. Symp. Circuits Syst. (ISCAS 2005)*, May 23–26, 2005, vol. III, pp. 1919–1923.

[27] D. Z. Jin and H. S. Seung, "Fast computation with spikes in a recurrent neural network," *Phys. Rev. E*, vol. 65, no. 5, pp. 051 922–1–051 922–4, 2002.

[28] M. Oster, "Spike-based winner-take-all computation in a multi-chip vision system," Ph.D., ETH Zurich, Zurich, Switzerland, 2006.

[29] M. Oster, R. Douglas, and S.-C. Liu, "Quantifying input and output statistics of a winner-take-all network in a vision system," in *Proc. 2007 IEEE Int. Symp. Circuits Syst. (ISCAS '07)*, New Orleans, LA, May 27–34, 2007, pp. 853–856.

[30] S.-C. Liu and M. Oster, "Feature competition in a spike-based winner-take-all VLSI network," in *Proc. 2006 IEEE Int. Symp. Circuits Syst. (ISCAS '06)*, Island of Kos, Greece, May 21–24, 2006, pp. 3634–3637.

[31] M. Oster and S.-C. Liu, "Spiking inputs to a winner-take-all network," in *Advances in Neural Information Processing Systems*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, vol. 18, pp. 1051–1058.

[32] M. A. Sivilotti, "Wiring considerations in analog VLSI systems, with application to field-programmable networks," Ph.D., Calif. Inst. Technol., Neural Syst. Dep., Pasadena, CA, 1991.

[33] A. Mortara, P. Venier, and E. A. Vittoz, "A communication scheme for analog VLSI perceptive systems," *IEEE J. Solid-State Circuits*, vol. 30, pp. 660–669, 1995.

[34] K. A. Boahen, "Retinomorphic vision systems: Reverse engineering the vertebrate retina," Ph.D., Calif. Inst. Technol., Pasadena, 1997.

[35] K. A. Boahen, "Communicating neuronal ensembles between neuromorphic chips," in *Neuromorphic Syst. Eng.*, T. Lande, Ed. Boston, MA: Kluwer Academic, 1998, ch. 11, pp. 229–259.

[36] S. R. Deiss, R. J. Douglas, and A. M. Whatley, "A pulse-coded communications infrastructure for neuromorphic systems," in *Pulsed Neural Netw.*, W. Maass and C. M. Bishop, Eds. Boston, MA: MIT Press, 1999, ch. 6, pp. 157–178.

[37] K. A. Boahen, "A burst-mode word-serial address-event link – I: Transmitter design," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 51, no. 7, pp. 1269–1280, 2004.

[38] K. A. Boahen, "A burst-mode word-serial address-event link – II: Receiver design," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 51, no. 7, pp. 1281–1291, 2004.

[39] A. Linares-Barranco, T. Serrano-Gotarredona, and R. Serrano-Gotarredona, "Compact low-power calibration mini-DACS for neural arrays with programmable weights," *IEEE Trans. Neural Netw.*, vol. 14, pp. 1207–1216, 2003.

[40] T. Delbrück and A. van Schaik, "Bias current generators with wide dynamic range," *Analog Integr. Circuits Signal Process.*, vol. 43, no. 3, pp. 247–268, 2005.

[41] R. Serrano-Gotarredona, B. Linares-Barranco, and T. Serrano-Gotarredona, "A new charge-packet driven mismatch-calibrated integrate-and-fire neuron for processing positive and negative signals in AER-based systems," in *Proc. 2004 IEEE Int. Symp. Circuits Syst. (ISCAS '04)*, Vancouver, Canada, May 25–28, 2004, vol. V, pp. 23–26.

[42] R. Serrano-Gotarredona, T. Serrano-Gotarredona, A. Acosta-Jimenez, and B. Linares-Barranco, "A neuromorphic cortical-layer microchip for spike-based event processing vision systems," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 53, pp. 2548–2566, 2006.

[43] X. Wang and S.-C. Liu, "Programmable synaptic weights for an aVLSI network of spiking neurons," in *Proc. 2006 IEEE Int. Symp. Circuits Syst. (ISCAS '06)*, Island of Kos, Greece, May 21–24, 2006, pp. 4531–4534.

[44] C. Rasche and R. Hahnloser, "Silicon synaptic depression," *Biolog. Cybern.*, vol. 84, no. 1, pp. 57–62, 2001.

[45] M. Boegerhausen, P. Suter, and S.-C. Liu, "Modeling short-term synaptic depression in silicon," *Neural Comput.*, vol. 15, no. 2, pp. 331–348, 2003.

[46] S.-C. Liu, J. Kramer, G. Indiveri, T. Delbrück, T. Burg, and R. Douglas, "Orientation-selective aVLSI spiking neurons," *Neural Networks: Special Issue on Spiking Neurons in Neurosci. Technol.*, vol. 14, no. 6/7, pp. 629–643, 2001.

[47] A. van Schaik, "Building blocks for electronic spiking neural networks," *Neural Networks: Special Issue on Spiking Neurons in Neurosci. Technol.*, vol. 14, no. 6/7, pp. 617–628, 2001.

[48] P. Merolla and K. Boahen, "Recurrent model of orientation maps with simple and complex cells," in *Adv. Neural Inf. Process. Syst.*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004, vol. 16, pp. 995–1002.

[49] E. Farquhar and P. Hasler, "A bio-physically inspired silicon neuron," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 52, no. 3, pp. 477–488, 2005.
[50] K. Wittig, J. Arthur, and K. Boahen, "Silicon neurons that phase-lock," in *Proc. 2006 IEEE Int. Symp. Circuits Syst. (ISCAS'06)*, Island of Kos, Greece, May 21–24, 2006, pp. 4535–4538.
[51] T. Serrano-Gotarredona and B. Linares-Barranco, "Systematic width-and-length dependent CMOS transistor mismatch characterization and simulation," *Analog Integr. Circuits Signal Process.*, vol. 21, pp. 271–296, 1999.
[52] E. Chicca, A. M. Whatley, P. Lichtsteiner, V. Dante, T. Delbrück, P. Del Giudice, R. J. Douglas, and G. Indiveri, "A multi-chip pulse-based neuromorphic infrastructure and its application to a model of orientation selectivity," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 5, pp. 981–993, 2007.
[53] P. Lichtsteiner, C. Posch, and T. Delbrück, "A 128 × 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.
[54] P. Häfliger, "Adaptive WTA with an analog VLSI neuromorphic learning chip," *IEEE Trans. Neural Netw.*, vol. 18, pp. 551–572, 2007.

Matthias Oster received the M.S. degree in electrical engineering and information technologies from the Technical University of Munich (TUM), Munich, Germany, in 2002, and the Ph.D. degree in November 2006, under the supervision of S.-C. Liu and R. Douglas, from the Institute of Neuroinformatics, University/ETH Zurich, Zurich, Switzerland.

During his Ph.D. studies, he explored spike-based winner-take-all computation integrated in the EU-project CAVIAR building a biologically inspired high-speed image-processing system. In a follow-up postdoctoral project until March 2007, he explored the embedding of syntactic constraints in hardware neuronal networks in the scope of the DAISY project. His research interests include biologically inspired computation algorithms and their implementation in hardware VLSI systems, asynchronous communications and information theory, and the field of vision processing. Since May 2007, he has been affiliated with Varian Medical Systems, Baden, Switzerland.

Dr. Oster was a member of the German National Academic Foundation and the 'deutsche Schülerakademie' and received travel grants from the ICANN conference in 2005, the Institute of Neuromorphic Engineering, MD, in 2004, and the University of Sevilla, Seville, Spain, in 2005.

Yingxue Wang (S'06-M'08) received the B.E. degree from the Beijing University of Technology, Beijing, China, and the M.S. degree from the Royal Institute of Technology, Sweden, both in electrical engineering, in 2002 and 2005, respectively.

From 2002 to 2003, she was a software engineer with Beijing Pattek Limited Co. (Institution of Automation, Chinese Academy of Sciences). She is currently working toward the Ph.D. degree in electrical engineering at the Swiss Federal Institute of Technology, Zurich, Switzerland. Her research interests include models of spiking neural network, neurons with dendritic structure, networks with oscillatory behavior, and VLSI implementation of those models.

Rodney Douglas received the M.S. degree in science and medicine, and the Ph.D. degree in neuroscience from the University of Cape Town, South Africa.

He is Professor of Neuroinformatics, and Co-Director of the Institute of Neuroinformatics of the Swiss Federal Institute and the University of Zürich, Switzerland. Thereafter, he joined the Anatomical Neuropharmacology Unit, Oxford, U.K., where he continued his research on the anatomy and biophysics of neocortical circuits together with K. Martin. As a Visiting Associate, and then Visiting Professor at the California Institute of Technology (Caltech), Pasadena, he extended his research interests in neuronal computation to the modeling of cortical circuits using digital methods (together with C. Koch), and also by the fabrication of analog VLSI circuits (together with M. Mahowald). In 1996, he and K. Martin moved to Zurich to establish the Institute of Neuroinformatics.

Dr. Douglas was awarded the Körber Foundation prize for European Science in 2000.

Shih-Chii Liu (M'02–SM'07) received the B.S. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, and the M.S. degree in electrical engineering from University of California, Los Angeles. She received the Ph.D. degree from the Computation and Neural Systems Program, California Institute of Technology (Caltech), Pasadena, in 1997.

She is currently an Assistant Professor with the Institute of Neuroinformatics, University of Zurich/ETH Zuric, Zurich, Switzerland. She was with Gould American Microsystems from 1983 to 1985, and with LSI Logic from 1985 to 1988. She worked at Rockwell International Research Labs from 1988 to 1997. Her research interests include neuromorphic neuronal modeling of vision and cortical processing, networks for behavior generation, and hybrid analog/digital signal processing.