(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 July 2020 (16.07.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/144555 A1

(51) International Patent Classification:
G06N 3/063 (2006.01) *G06N 3/04* (2006.01)

(21) International Application Number:
PCT/IB2020/050058

(22) International Filing Date:
06 January 2020 (06.01.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
19150781.3 08 January 2019 (08.01.2019) EP

(71) Applicant: **UNIVERSITÄT ZÜRICH** [CH/CH]; Prorektorat Forschung, Rämistrasse 71, CH-8006 Zürich (CH).

(72) Inventors: **INDIVERI, Giacomo**; Mattackerstrasse 33, 8052 Zürich (CH). **NAIR, Manu Vijayagopalan**; Querstrasse 6, 8600 Dübendorf (CH).

(74) Agent: **LUMI IP LLC**; Rodtmattstrasse 45, 3014 Bern (CH).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

WO 2020/144555 A1

(54) **Title:** RANDOM WEIGHT GENERATING CIRCUIT

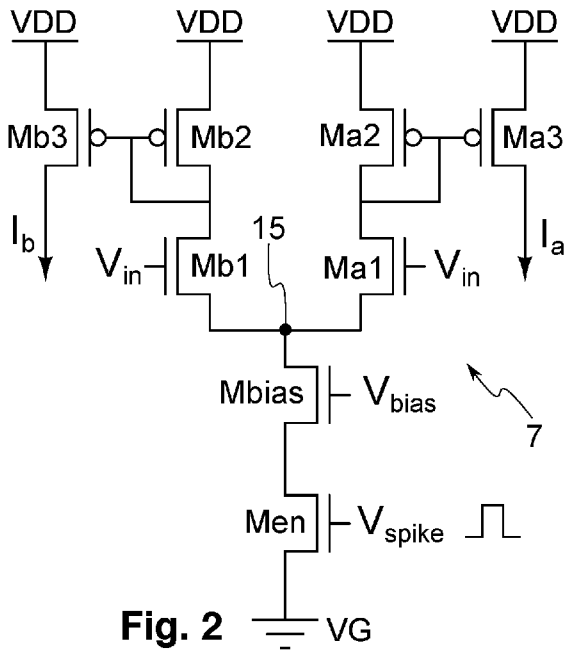


Fig. 2

(57) **Abstract:** The present invention relates to a circuit for generating random weights for neuromorphic processors, for example. The circuit comprises a first voltage node (VDD) for providing a supply voltage for the circuit; a second voltage node (VG) at a given electric potential; a first circuit element (Ma1) with a first electric current carrier concentration for outputting a first circuit element output signal; and a second circuit element (Mb1) with a second electric current carrier concentration for outputting a second circuit element output signal. The first and second circuit elements are located between the first and second voltage nodes, which have a given voltage difference between them. The first circuit element output signal and the second circuit element output signal are different due to the first and second electric carrier concentrations being mismatched. The circuit further comprises a subtraction unit configured to generate a respective random weight, which is represented by a difference between the first circuit element output signal and the

[Continued on next page]

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

RANDOM WEIGHT GENERATING CIRCUIT

TECHNICAL FIELD

The present invention relates to a random weight generating circuit to be used for instance in analogue or mixed-signal (analogue-digital) very-large-scale integration (VLSI) implementations of artificial neural networks. The circuit may comprise a compact synapse memory cell for use in neuromorphic processors. More specifically, the proposed circuit takes advantage of a mismatch between a given property of two devices in the circuit. The invention also relates to a method of operating the circuit.

BACKGROUND OF THE INVENTION

Recurrent neural networks (RNNs) are powerful algorithms for processing temporal information. The effectiveness of RNNs in processing temporal data makes it compelling to use them in ultra-low power applications, such as biomedical implants or energy-harvesting smart sensors. However, these algorithms are computationally expensive in classical von Neumann architectures. In-memory computing architectures address this problem by performing analogue or digital compute operations within the memory itself instead of moving the data to the processor and back. This approach has been demonstrated for various applications, using neuromorphic processors which comprise synaptic memory cells containing circuits that implement scaling and filtering functions.

A mixed-signal neuromorphic processor typically uses low-precision and mismatch-prone synaptic weights. Conventional RNNs typically use backpropagation through time (BPTT) to train the network. However, this is difficult to implement on-chip in neuromorphic platforms. This problem can be by-passed by resorting to an alternative RNN computing paradigm based on “reservoir computing”, where the difficult-to-train parameters that define the recurrent connectivity are not trained. Instead, the recurrent synaptic weights are initialised randomly with an eigen-spread chosen such that they have the ”echo-state property”. This property ensures that in the absence of an input signal, the network activity becomes zero in the steady state. To achieve this, the randomly generated network weights are re-scaled such that the largest eigen-value of the recurrent matrix is made sufficiently small. In this state, the recurrent network acts as a dynamic system that projects the temporal and spatial properties of the incoming data into a high-dimensional space. The projections in this high-dimensional space can then be classified using a shallow feed-forward network, which can be trained online in a neuromorphic platform.

Reservoir RNNs typically do not perform as well as BPTT trained RNNs in complex tasks, such as speech recognition or translation, but are adequate for simpler tasks, such as object tracking, motion prediction or event detection. Reservoir computing is a framework for computation that may be viewed as a subset of artificial neural networks. Typically, an input signal is fed into a fixed (and random) dynamic system called a reservoir and the dynamics of the reservoir map the input to a higher dimension. A simple readout mechanism is trained to read the state of the reservoir and map it to the desired output. The main benefit is that training is carried out only at the readout stage and the reservoir is fixed. Liquid-state machines and echo state networks are two common types of reservoir computing solutions. Spiking neuromorphic reservoirs can implement such networks in a very energy-efficient manner. However, neuromorphic reservoirs require scalable random synaptic weights for the recurrent connectivity matrix to ensure that the network operates in the “echo-state” regime. However, currently known circuits for generating random weights are rather complex and they consist of many components and it is thus very challenging to make them compact. The synaptic weights require a sufficiently high resolution to ensure that the distribution of the sampled weights is representative of a random distribution. Achieving high resolution by allocating several bits of memory per cell costs too much area and therefore, does not scale well.

US8384569 (B2) discloses a stochastic signal generation circuit including a signal output circuit and a signal processing circuit connected with the signal output circuit. The signal output circuit includes two matching semiconductor components, wherein the signal output circuit detects a slight mismatch between the two matching semiconductor components, converts the detected slight mismatch into a corresponding electric signal, amplifies the electric signal, and outputs an analogue voltage signal. The signal processing circuit converts the analogue voltage signal into a stochastic digital signal.

US2008313250 (A1) discloses a random signal generator including a differential noise generation circuit, an amplification circuit and a single-ended amplifier. The differential noise generation circuit includes a pair of input nodes and a pair of output nodes, and is configured to receive noise signals at the pair of input nodes and to generate differential noise signals at the pair of output nodes. The differential noise generation circuit is self-biased such that the pair of input nodes is coupled to the pair of output nodes. The amplification circuit is configured to amplify the differential noise signals output from the differential noise generation circuit to generate amplified differential signals. The single-ended amplifier is configured to generate a random signal based on the amplified differential signals, the random signal having irregular transition time points.

"A neuromorphic systems approach to in-memory computing with non-ideal memristive devices: from mitigation to exploitation" by Melika Payvand et al., XP081248190, presents a spiking neural network architecture that supports the use of memristive devices as synaptic elements and propose mixed-signal analogue-digital interfacing circuits that mitigate the effect of variability in their conductance values and exploit their variability in the switching threshold for implementing stochastic learning.

SUMMARY OF THE INVENTION

The objective of the present invention is thus to overcome at least some of the above limitations relating to circuits capable of generating random weights. More specifically, the present invention addresses the size and power consumption limitations of existing random weight generating circuits.

According to a first aspect of the invention, there is provided a random weight generating circuit as recited in claim 1.

The present invention has the advantage that the proposed circuit consists of only very few components and thus is very compact. The energy consumption of the circuit can also be made very small. Thus, the proposed circuit can be used in ultra-low power applications, such as biomedical implants, optionally implanted under the skin, and energy-harvesting smart sensors. The proposed circuit can enable a new family of smart sensory applications that have not been possible so far. Examples of these applications are a solar-energy harvesting sticker with smart monitoring capabilities, or a glucose (or other property) harvesting implant which detects disruptions in natural bio-rhythms.

According to a second aspect of the invention, there is provided a neural network accelerator comprising the circuit according to the first aspect of the present invention.

According to a third aspect of the invention, there is provided a neuromorphic processor comprising the circuit according to the first aspect of the present invention.

According to a fourth aspect of the invention, there is provided a method of generating one or more random weights as recited in claim 22.

Other aspects of the invention are recited in the dependent claims attached hereto.

BRIEF DESCRIPTION OF THE DRAWINGS

Other features and advantages of the invention will become apparent from the following description of a non-limiting example embodiment, with reference to the appended drawings, in which:

- Figure 1 is a neuromorphic processor block diagram illustrating a circuit, where the teachings of the present invention may be applied;
- Figure 2 is a circuit diagram showing a reservoir synapse circuit using a differential pair according to an example of the present invention;
- Figure 3 is a circuit diagram showing a reservoir synapse circuit using a differential pair and further comprising a differential pair integrator circuit according to an example of the present invention;
- Figure 4a is circuit diagram showing a portion of the circuit of Figure 2;
- Figure 4b is a circuit diagram showing another portion of the circuit of Figure 2;
- Figure 5a is a circuit diagram showing a portion of the circuit of Figure 3;
- Figure 5b is a circuit diagram showing the circuit of Figure 5a, but where some of the circuit elements are memristive devices instead of transistors;
- Figure 5c is a circuit diagram showing another portion of the circuit of Figure 3;
- Figure 6 is a circuit diagram illustrating a variant of the proposed random weight generating circuit; and
- Figure 7 is a circuit diagram illustrating another variant of the proposed random weight generating circuit.

DETAILED DESCRIPTION OF AN EMBODIMENT OF THE INVENTION

An embodiment of the present invention will now be described in detail with reference to the attached figures. This embodiment is described in the context of a neuromorphic processor implementing reservoir computing, but the teachings of the invention are not limited to this environment. For example, the teachings of the present invention are not limited to reservoir computing or to neuromorphic processors. The teachings of the present invention may equally be used in cryptography, for instance. The

example non-limiting embodiment of the present invention relates to a differential current-mode circuit arrangement. In the present description, by a differential circuit is understood a circuit in which critical signals are represented by the difference or by another mathematical relationship of two signals (or their signal amplitude values) rather than a single signal. It is to be noted that the word “signal” is used in the present description in its broad sense, and this word does not imply that any information would be coded in the signal. Furthermore, because the circuit of the present embodiment is a current-mode circuit, these signals are currents. Identical or corresponding functional and structural elements which appear in different drawings are assigned the same reference numerals.

It will be shown below that memory cells of a neuromorphic processor can be made much smaller by restricting the chip functionality to reservoir networks and its masked variant. To this end, the present invention proposes a compact, such as a three-transistor, analogue memory cell that implements a real-valued and tuneable memory by exploiting complementary metal-oxide-semiconductor (CMOS) mismatch or more broadly silicon mismatch or other type of mismatch between two circuit devices or elements. The analogue memory cell forms a synaptic cell or synapse for short. This approach enables implementing for instance compact and ultra-low power accelerators that consume less than 1 mW when running an RNN with a few hundred units. A circuit variant suitable for use with memristive or capacitive circuit elements will also be proposed.

The block diagram of Figure 1 illustrates an example neuromorphic processor 1, where the teachings of the present invention may be applied. The processor comprises an input interface 3 for receiving input signals and for feeding the input signals into first synapses 5, referred to also as input or non-recurrently connected synapses, illustrated in Figure 1 by blocks with forward hashing. The processor 1 comprises a given number of input synapses forming a set or array of input synapses 5, which are non-recurrently connected synapses. The processor 1 also comprises second synapses 7 forming a set or array of recurrently connected synapses. The recurrently connected synapses 7 are illustrated in Figure 1 by blocks with backward hashing. As is further shown in Figure 1, a given row of synapses (consisting of recurrently and non-recurrently connected synapses) is connected to its respective amplifier and input stage unit 9. More specifically, the outputs of all the synapses are summed and the sum is fed into the amplifier and input stage unit 9. The respective amplifier and input stage unit 9 receives a first output signal and a second, different output signal from a given row of synapses as will be explained later. An output of a given amplifier and input stage unit 9 is connected to an input of a respective neuron 11 (of a set of neurons), operating in this example as a spiking neuron, arranged to generate output signals, also referred to as spikes, which are binary control

signals (voltages), and which in this example are fed back to the synapses 7 to thus form recurrently connected synapses 7. This feed-back mechanism also enables the neuromorphic processor to operate as a neuromorphic reservoir. The outputs of the neurons 11 are connected to a readout unit or block 13, which thus forms the output of the entire neural network.

The inputs to the synaptic array are encoded in spike or pulse trains. Each spike event turns on a synapse for a brief duration during which it generates two different output signals or currents, which are used to obtain its weight or importance coefficient or parameter defining the strength of a connection between two circuit nodes. The currents from all the synapses 5, 7 in a row are integrated by a single neuron 11. The neuron applies a non-linearity to the integrated current and transmits the output by encoding it in spike or pulse trains. The input stream to the reservoir consisting of the recurrently connected synapses 7 is fed in through at least some of the input synapses 5, which are similar to the synapses in the reservoir in all aspects other than they are not recurrently connected to the neurons 11. The synaptic weights are in this example read in the form of currents that are collected by the neurons and transformed to spike trains again.

Implementing a neuromorphic reservoir requires three components: 1) a sufficiently fast communication protocol to carry the spikes, 2) a neuron capable of encoding the temporal signals in spike trains with sufficient precision, and 3) randomly generated synaptic weights with a tunable eigen-spread. The address-event representation (AER) communication protocol has traditionally been the workhorse for neuromorphic event-based computation and performs well. Both the input interface 3 and the readout unit 13 may use this protocol. The third component is a memory cell (i.e. a synapse in this example) with random synaptic weights and a tunable eigen-spread. This can be built by incorporating several bits of memory in each synaptic cell and programming them with suitable values. However, for implementing neuromorphic reservoirs, it will be shown that the synaptic weights can be implemented by a compact 3-transistor circuit, for instance. The circuit eliminates the need for digital memory cells in synapses entirely. This dramatically reduces the synaptic cell area enabling neuromorphic processors that can advantageously implement recurrent networks with an order of magnitude more neurons and synapses.

The key idea in the proposed synapse circuit is to use the mismatch between a given internal property of two separate components or devices in the synapse circuit. In other words, current carrier concentrations of these two devices are mismatched, which makes it possible to generate the random weights according to the teachings of the

present invention. In the embodiment explained below, the key idea in the proposed synapse circuit is to use the CMOS mismatch.

Figure 2 shows a reservoir synapse circuit using a differential pair according to the present example. The synapse circuit comprises a first device Ma1, which in this example is a first transistor, and more specifically a first metal-oxide-semiconductor field-effect transistor (MOSFET) optionally operated in the sub-threshold region. The first device is part of a first circuit output branch and outputs a first output signal I_a, which in this example is a first output current. During operation of the circuit, the first output current I_a also flows through the first device Ma1. The operation of the first device is arranged to be controlled by a first control signal, referred to as a first control voltage V_{in}, at its gate terminal. The synapse circuit also comprises a second device Mb1, which in this example is a second transistor, and more specifically a second MOSFET optionally operated in the sub-threshold region. The second device is part of a second circuit output branch and outputs a second output signal I_b, which in this example is a second output current. During operation of the circuit, the second output current I_b also flows through the second device Mb1. The operation of the second device is arranged to be controlled by the first control signal V_{in} at its gate terminal. This means that the same control signal is used to control the first and second devices Ma1, Mb1. The first and second devices Ma1, Mb1 are configured in a differential mode and form a differential pair because their source terminals are connected to a common circuit node 15.

A first synapse readout circuit is connected between a first voltage reference node, referred to as a supply voltage source VDD, and the first device Ma1 for reading or copying the first output signal I_a. The supply voltage source is configured to produce a constant analogue voltage. The first synapse readout circuit is in this example a first current mirror. As is shown in Figure 2, the first current mirror comprises a third transistor Ma2 and a fourth transistor Ma3, whose source nodes or terminals are connected to the first reference voltage node VDD. The drain terminal of the third transistor Ma2 is connected to the drain terminal of the first transistor Ma1.

A second synapse readout circuit is connected between the first voltage reference node VDD and the second device Mb1 for reading or copying the second output signal I_b. The second synapse readout circuit is in this example a second current mirror. As is further shown in Figure 2, the second current mirror comprises a fifth transistor Mb2 and a sixth transistor Mb3, whose source terminals are connected to the first reference voltage node VDD. The drain terminal of the fifth transistor Mb2 is connected to the drain terminal of the second transistor Mb1.

The source terminals of the first and second transistors Ma1, Mb1 are connected through the common node 15 to a seventh transistor Mbias, and more specifically to its drain terminal. The gate terminal of the seventh terminal is arranged to be controlled by a second control signal, referred to as a bias voltage V_{bias}. The seventh transistor Mbias, and more specifically its source terminal, is connected to an eight transistor Men, and more specifically to its drain terminal. The source terminal of the eight transistor is connected to a second reference voltage node VG, which in this example is at zero electric potential, i.e. the source terminal of the eight transistor is grounded. The gate terminal of the eight transistor is arranged to receive a third control signal, referred to as a spike signal or spike voltage V_{spike}, to selectively turn on or off the switch formed by the eight transistor. In the present example, all the transistors of the circuit of Figure 2 are MOSFETs operated in the sub-threshold region. More specifically, the transistors of the current mirrors are in this example p-channel MOSFETs while the other transistors of the circuit are n-channel MOSFETs.

The weight of a respective synaptic cell 5, 7 can be implemented as the difference at a given time instant between the current (signal amplitude) flowing in the two output branches of a differential pair as shown in Figure 2, i.e. $I_w = I_a - I_b$. Using a differential arrangement eliminates systematic effects caused by location-dependent properties during fabrication such as doping and temperature on the wafer. The weights can be tuned up or down by scaling the tail current I_{bias}, where $I_{bias} = I_a + I_b$, of the circuit, i.e. the current flowing through the transistor Mbias. This can be carried out by adjusting the control voltage V_{bias}. In this manner, the output currents I_a and I_b can be adjusted. Clearly, the synaptic weight, I_w will scale linearly with the tail current. It is to be noted that in the present example, a synaptic weight is calculated individually for each synapse in the neuromorphic processor. Furthermore, the biasing state or condition of the circuit defined e.g. by V_{in} and V_{bias} may be kept constant once initially set. This would mean that the weights of the synapses would not change over a plurality of spiking cycles. Alternatively, it would be possible to change the biasing condition for each spiking cycle or at least for some of the spiking cycles. The same biasing condition is typically applied for all of the synapses of the neuromorphic processor. However, instead, it would be possible to apply different biasing conditions for at least some of the synapses of the neuromorphic processor.

The weights implemented by the differential pair circuit of Figure 2 were sampled by running Monte-Carlo mismatch simulations. The results show a linear relationship between the eigen-spread and the tail current. The weights in the network for different values of the tail current exhibit a Gaussian-like distribution with zero mean.

Using a differential arrangement addresses the problem of systematic mismatch between the first and second devices and tunable eigen-value spread. However, a few aspects may still be optimised. The output current can get saturated to the tail bias current if the two branches of the differential pair are highly mismatched. This will cause clipping in the output weight distribution. Furthermore, the power consumption of the differential circuit in the absence of input spikes can be non-zero because of parasitic charges in the gates of transistors Ma3 and Mb3. This can be addressed by pulling the gates to VDD in the absence of a spike, but that would need more devices increasing the cell footprint. These problems can be overcome by using a differential pair integrator (DPI) to implement the amplifier stage as for example explained by E. Chicca et al. “Neuromorphic electronic circuits for building autonomous cognitive systems”, as published in Proceedings of the IEEE 102.9 (Volume: 102, Issue: 9, Sept. 2014), pp. 1367–1388, ISSN: 0018-9219.

The DPI synapse circuit configuration as shown in Figure 3 consists of two symmetric blocks, namely a first block and a second block, each comprising six transistors according to this example. Transistors Ma2-5 and Mb2-5 form the DPI circuit where transistors Ma5 and Mb5 are the output transistors. The transistors Ma2 and Ma5 form a first DPI circuit current mirror while the transistors Mb2 and Mb5 form a second DPI circuit current mirror. The transistor Ma3 is in parallel with the transistors Ma2 and Ma4 (which are in series) while the transistor Mb3 is in parallel with the transistors Mb2 and Mb4 (which are in series). It is to be noted that the configuration of Figure 3 comprises two switches, namely Mena and Menb, i.e. one switch for each one of the DPI circuit input branches. However, it would be possible to build the circuit by using only one switch (Men) instead, which would be shared between both the DPI circuit input branches (similar to the configuration of Figure 2). The equation describing the behaviour of the circuit valid in the sub-threshold mode is given by

$$V_\tau \frac{d}{dt} I_{out} + I_{out} = \frac{I_{th}}{I_\tau} I_{in}, \quad (1)$$

where, V_τ is control voltage, and more specifically a fourth control voltage, i.e. the voltage controlling the operation of the transistors Ma4 and Mb4, and more specifically their gate voltages, and is close to zero because capacitors are taken out of the DPI circuit in this implementation. I_{out} denotes the output current, which is either I_a or I_b. I_{in} is the input current of the DPI circuit (flowing either in or out of the DPI circuit depending on its implementation) and is also the current through the transistors Ma1 or Mb1, I_{th} is a current generated by a fifth control voltage V_{th}, which is the control voltage controlling the gates of the transistors Ma3 and Mb3, while I_τ is the current through the transistors Ma4

or Mb4 and generated by the control voltage V_τ. The gain of this circuit is given by $\gamma = \frac{I_{th}}{I_\tau}$. As in the differential version of Figure 2, the reservoir synapse circuit stores the weight in the difference between the two outputs as $I_w = I_a - I_b$.

The circuit addresses the problems present in the differential pair of Figure 2 because it does not saturate and has no floating gates in the absence of a spike. One potential problem with the circuit is that the output currents could deviate from Equation 1 for above-threshold currents.

In the absence of a mismatch,

$$I_a = I_b = I_{in} \times \frac{I_{th}}{I_\tau} = I_{in} \times \gamma, \quad (2)$$

where, I_{in} is again the current flowing through the transistors Ma1 and Mb1 when there is a spike event. The mismatch between I_a and I_b can be modelled to a first order approximation as

$$I_x = (I_{mean} + \Delta I_{in}/2) \times (\gamma_{mean} + \Delta \gamma/2), \quad (3)$$

where, $x = a$ or b. Using this convention, after some basic algebra, the stored synaptic weight, I_w can be expressed as

$$I_w = \Delta\gamma \times I_{mean} + \Delta I_{in} \times \gamma_{mean} \quad (4)$$

Equation 4 shows that the effective value of the weight stored in the arrangement of Figure 3 is the sum of the mismatch between the gain and I_{in} of the “a” and “b” branches of the synapse scaled by the mean values of the bias and gain values, respectively. The mean and standard deviation in the distribution of synaptic weights, I_w can be scaled by either increasing I_{in} (by adjusting V_{in}) or the gain in the circuit. This also changes the eigen-spread of the reservoir weights.

Monte Carlo simulations show that the resultant distribution of the weights resembles a Gaussian distribution (when considering the weights of the entire synapse array). A scaling factor is applied to the output currents representing the synaptic weight, I_w, to account for the input range of the neuron integrating the output currents. By tuning the gain γ and I_{in}, the spread of the weights and the largest eigen-value can be controlled in an almost linear fashion.

The synaptic circuits presented so far comprise circuits that generate a mismatched differential output (the two output signal values being different), amplify this difference, and send it out to the integrating neuron 11. To make the memory cell compact

(i.e. the synapse), only those devices that are essential for generating mismatch need to remain within the synaptic cell and the rest can be shared. This is similar to how a readout sense amplifier stage is shared between multiple memory cells in a static random-access memory (SRAM) memory array. The shared components should then be integrated into the input stage to the neuron, i.e. into the amplifier and input stage unit 9, which may also comprise a subtraction unit to carry out the subtraction $I_a - I_b$. Concretely, a compact implementation of the differential circuit of Figure 2 only retains the differential pair and tail inside the synaptic cell as shown in Figure 4a. The current mirrors at the top (each comprising the transistors Mx2 and Mx3), which are arranged to sum the synaptic outputs, are shared between all the memory cells in a row and incorporated in the input stage of the neuron as shown in Figure 4b. It is to be noted that $\sum I_a = I_p$, where I_p is a first differential output signal, and $\sum I_b = I_n$, where I_n is a second differential output signal of the amplifier and input stage unit 9.

Similarly, the DPI version of the reservoir synapse cell (shown in Figure 3) is made compact by only retaining the mismatch-generating devices, Ma1 and Mb1, and optionally the switch Men inside the memory cell, and sharing the DPI amplifier stages between all the synapses in a row. Figure 5a shows the devices that are inside the compact synapse cell 5, 7. It is easy to modify the synapse circuit by replacing the transistors Ma1 and Mb1 transistors by for example memristive devices Da and Db as shown in Figure 5b. The shared DPI circuit to be integrated into the input stage of the neuron is shown in Figure 5c. The cost of doing this is that the mismatch contributing to the synaptic weights becomes smaller. However, scaling down the device sizes could compensate for this effect as that increases mismatch.

Several variants of compact synapses (e.g. the three-transistor cells) for use in neuromorphic reservoirs in a two-device differential arrangement were presented. As mismatch increases for smaller devices, the proposed memory cell scales well with technology unlike traditional analogue circuits. In other words, the proposed circuit takes advantage of the increased mismatch as the devices are made physically smaller. These synaptic cells have been used to build a neuromorphic reservoir system and may be used for temporal processing tasks in ultra-low power edge applications (i.e. in applications having only very limited amount of energy available).

The proposed random weight generator may be used for instance in bio-medical smart implants that constantly monitor patients, and the circuit will never need to be recharged. Enabling these products requires ultra-low power computation and a small footprint. The proposed invention addresses this problem. Another interesting application is in “smart” wearables that monitor various modalities, such as temperature or pressure

patterns. These devices have a small battery or even an energy-harvesting mechanism, such as a solar panel patch. All of this needs a processing unit which is compact and has a very low power consumption. Neuromorphic computation addresses this space. The present invention helps make the neuromorphic processor extremely compact saving cost and space.

A neural network accelerator or processor that is described in earlier paragraphs uses the proposed invention to generate random weights. These weights can be used for reservoir computation. This idea can be extended by incorporating a masking matrix as explained in the following paragraphs so that this kind of circuit would no longer be a reservoir computation circuit. The idea behind the masking matrix is to restructure the weight matrix of the neural network W as a product of a random weight matrix R and a masking matrix with a single-bit value M by using element by element multiplication (also known as Hadamard product), as follows:

$$W = R \cdot M. \quad (5)$$

The elements (M_x) in the matrix thus have a value of either 0 or 1. It is to be noted that originally the matrix W was made up of random weight components. The masking matrix is obtained either online or offline as the outcome of the training.

It is well known that a network with fully random weights has a lesser representational power than a fully trained network. The motivation for the use of the masking matrix is to bridge this performance gap while retaining the area savings obtained by the compact random weight generation circuit. As shown in Figure 6, this can be implemented for example by adding a single bit memory 17 to the memory cell formed by the entire circuit shown in Figure 6.

The masking matrix can be used in at least two ways:

1) In the first mode of operation, the optimisation algorithm finds the best matrix M suitable for the network while also training the read-out layer.

2) In the second mode of operation, the connectivity or weight matrix is masked by different random matrices for different use cases. For example, M_1 for training on task 1, M_2 for task 2 and so on. Depending on the use-case, the appropriate masking matrix is loaded.

The masking bit implementation is achieved by including a masking bit or masking signal (which is understood to be a control signal) in the circuit. The bit can be implemented using a range of memory technologies. The bottom half of Figure 6

represents the random weight generation circuit. The rest of Figure 6 describes the arrangement by which the masking bit can be incorporated into a memory cell or unit 17, which in this example is an SRAM cell, connected to a new switch Mm (which in this example is in a series configuration with it), and in this example implemented by a transistor, and more specifically by a MOSFET. Instead of an SRAM memory unit, any other memory types can be used, as long as they are able to store at least one data bit. In essence, a single bit is required to turn on or off the random weight value coming out of the random weight generation circuit. When implemented across the entire neural network or at least across a portion of it, the arrangement of Figure 6 is thus configured to individually enable each random weight in the neural network. It is to be noted that one spike signal would collectively control many random weight generation circuits by controlling the operation of the switches Men. The random weight generation part of the circuit of Figure 6 can incorporate any of the random weight generation circuits described above. The frequency of the masking signal or its activity pattern more specifically would typically be lower than the frequency or activity pattern of the spike signal, which in this example is not a periodic signal. It is also to be noted that each memory cell could comprise its own dedicated memory unit, or one memory unit could be shared between more than one random weight generating circuit.

The arrangement of Figure 7 shows an alternative solution for the arrangement of Figure 6 to be used in a neural network accelerator. The function of the circuit of Figure 7 is the same as the circuit of Figure 6. Regarding the circuit structure, the difference is that no memory cell is required in the circuit of Figure 7, but a digital logic AND gate 19 is used instead, which is connected to the switch Men. One of the inputs of the AND gate is the spike signal, while the other input is the matrix element M_x. In this manner, the switch is turned on only when both the inputs have a logic one level. The logic AND gate may by implemented by a transistor circuit, such as a CMOS circuit.

While the invention has been illustrated and described in detail in the drawings and foregoing description, such illustration and description are to be considered illustrative or exemplary and not restrictive, the invention being not limited to the disclosed embodiment. Other embodiments and variants are understood, and can be achieved by those skilled in the art when carrying out the claimed invention, based on a study of the drawings, the disclosure and the appended claims. For example, instead of operating the circuit in current mode, it could be operated in voltage mode. More specifically, by adding a load, such as a resistor or a transistor, to the electric circuit output nodes, i.e. to the nodes where the output currents I_a and I_b are measured, the output from the circuit would be voltages instead of currents. Therefore, the circuit can be used either in voltage mode

or current mode. It is to be noted that the exact mismatch values between the second circuit element output signal and the first circuit element output signal depend on the fabrication process.

In the claims, the word "comprising" does not exclude other elements or steps, and the indefinite article "a" or "an" does not exclude a plurality. The mere fact that different features are recited in mutually different dependent claims does not indicate that a combination of these features cannot be advantageously used. Any reference signs in the claims should not be construed as limiting the scope of the invention.

CLAIMS

1. A random weight generating circuit (7, 9) comprising:
   - a first voltage node (VDD) for providing a supply voltage for the random weight generating circuit;
   - a second voltage node (VG) at a given electric potential;
   - a first circuit element (Ma1) characterised by a first electric current carrier concentration for outputting a first circuit element output signal; and
   - a second circuit element (Mb1) characterised by a second electric current carrier concentration for outputting a second circuit element output signal, the first and second circuit elements being located between the first and second voltage nodes having a given voltage difference between them, the first circuit element output signal and the second circuit element output signal being different due to the first and second electric carrier concentrations being mismatched,

wherein the random weight generating circuit (7, 9) comprises a subtraction unit (9) configured to generate one or more random weights such that a respective random weight is represented by a difference between the first circuit element output signal and the second circuit element output signal, wherein the random weight generating circuit (7, 9) further comprises a switch (Men), the operation of which is configured to be controlled by a control signal for selectively turning on or off the random weight generating circuit (7, 9).

2. The circuit (7, 9) according to claim 1, wherein the first circuit element (Ma1) and/or the second circuit element (Mb1) are one of the following elements: a transistor, a memristive element, and a capacitor.

3. The circuit (7, 9) according to claim 1 or 2, wherein the circuit further comprises a scaling circuit (Mbias) for scaling the one or more weights up or down.

4. The circuit (7, 9) according to claim 3, wherein the scaling circuit (Mbias) is a transistor circuit.

5. The circuit (7, 9) according to claim 3 or 4, wherein the scaling circuit (Mbias) is between the first circuit element circuit (Ma1) or the second circuit element (Mb1) and the second voltage node (VG).

6. The circuit (7, 9) according to claim 3 or 5, wherein the scaling circuit comprises a differential pair integrator.

7. The circuit (7, 9) according to any one of claims 3 to 6, wherein the first and second circuit elements (Ma1, Mb1) are comprised in a memory cell, and the circuit comprises a plurality of memory cells, and wherein the scaling circuit is shared between a plurality of the first circuit elements (Ma1) and a plurality of the second circuit elements (Mb1).

8. The circuit (7, 9) according to any one of the preceding claims, wherein the first and second circuit elements (Ma1, Mb1) are arranged as a differential pair.

9. The circuit (7, 9) according to any one of the preceding claims, wherein the circuit further comprises a readout circuit for reading the first element output signal and the second element output signal.

10. The circuit (7, 9) according to claim 9, wherein the readout circuit comprises a current mirror and/or a differential pair integrator.

11. The circuit (7, 9) according to claim 9 or 10, wherein the first and second circuit elements (Ma1, Mb1) are comprised in a memory cell, and the circuit comprises a plurality of memory cells, and wherein the readout circuit is shared between a plurality of the first circuit elements (Ma1) and a multiplicity of the second circuit elements (Mb1).

12. The circuit (7, 9) according to any one of the preceding claims, wherein the random weight generating circuit (7, 9) comprises a control circuit (17, 19) for selectively turning on or off the random weight generating circuit (7, 9) by using a masking signal.

13. The circuit (7, 9) according to claim 12, wherein the control circuit (17, 19) comprises a control switch (Mm) and a memory unit (17) for storing a masking bit so that the masking signal at an output terminal of the memory unit is configured to selectively turn or off the control switch (Mm) to thereby turn on or off the random weight generating circuit (7, 9).

14. The circuit (7, 9) according to claim 12, wherein the control circuit (17, 19) comprises a logic AND gate (19) connected to the switch (Men) for selectively turning on or

off the switch (Men), the control signal and the masking signal are configured to be fed into the AND gate (19).

15. The circuit (7, 9) according to any one of claims 12 to 14 wherein the control signal is characterised by a first signal activity pattern, and the masking signal is characterised by a second, different signal activity pattern.

16. A neural network accelerator comprising the circuit (7, 9) according to any one of the preceding claims.

17. A neuromorphic processor (1) comprising the circuit (7, 9) according to any one of the preceding claims.

18. The neuromorphic processor (1) according to claim 17, wherein the neuromorphic processor comprises an artificial neuron (11), and the control signal is configured to be received from the artificial neuron (11), and wherein the first and second circuit element output signals are currents.

19. The neuromorphic processor (1) according to claim 17 or 18, wherein the neuromorphic processor (1) comprises a synapse cell, and wherein the first and second circuit elements (Ma1, Mb1) are comprised in the synapse cell.

20. The neuromorphic processor (1) according to any one of claims 17 to 19, wherein the neuromorphic processor (1) comprises a set of artificial neurons (11), and a neuromorphic reservoir of recurrently connected synaptic cells (7) formed by connecting a respective artificial neuron (11) by a feed-back mechanism to a set of synaptic cells (7) comprised in the neuromorphic reservoir.

21. The neuromorphic processor (1) according to any one of claims 17 to 20, wherein the neuromorphic processor (1) is configured to implement reservoir computing.

22. A method of generating one or more random weights in a circuit (7, 9) comprising:

- a first voltage node (VDD) for providing a supply voltage for the circuit;
- a second voltage node (VG) at a given electric potential;

- a first circuit element (Ma1) characterised by a first electric current carrier concentration for outputting a first circuit element output signal; and
- a second circuit element (Mb1) characterised by a second electric current carrier concentration for outputting a second circuit element output signal, the first and second circuit elements being located between the first and second voltage nodes having a given voltage difference between them, the first circuit element output signal and the second circuit element output signal being different due to the first and second electric carrier concentrations being mismatched,

the method comprising:

- generating a respective random weight by subtracting the second circuit element output signal from the first circuit element output signal to represent the respective random weight as a difference between the first circuit element output signal and the second circuit element output signal; and
- controlling the operation of a switch (Men) of the circuit (7, 9) by means of a control signal for selectively turning on or off the circuit (7, 9).

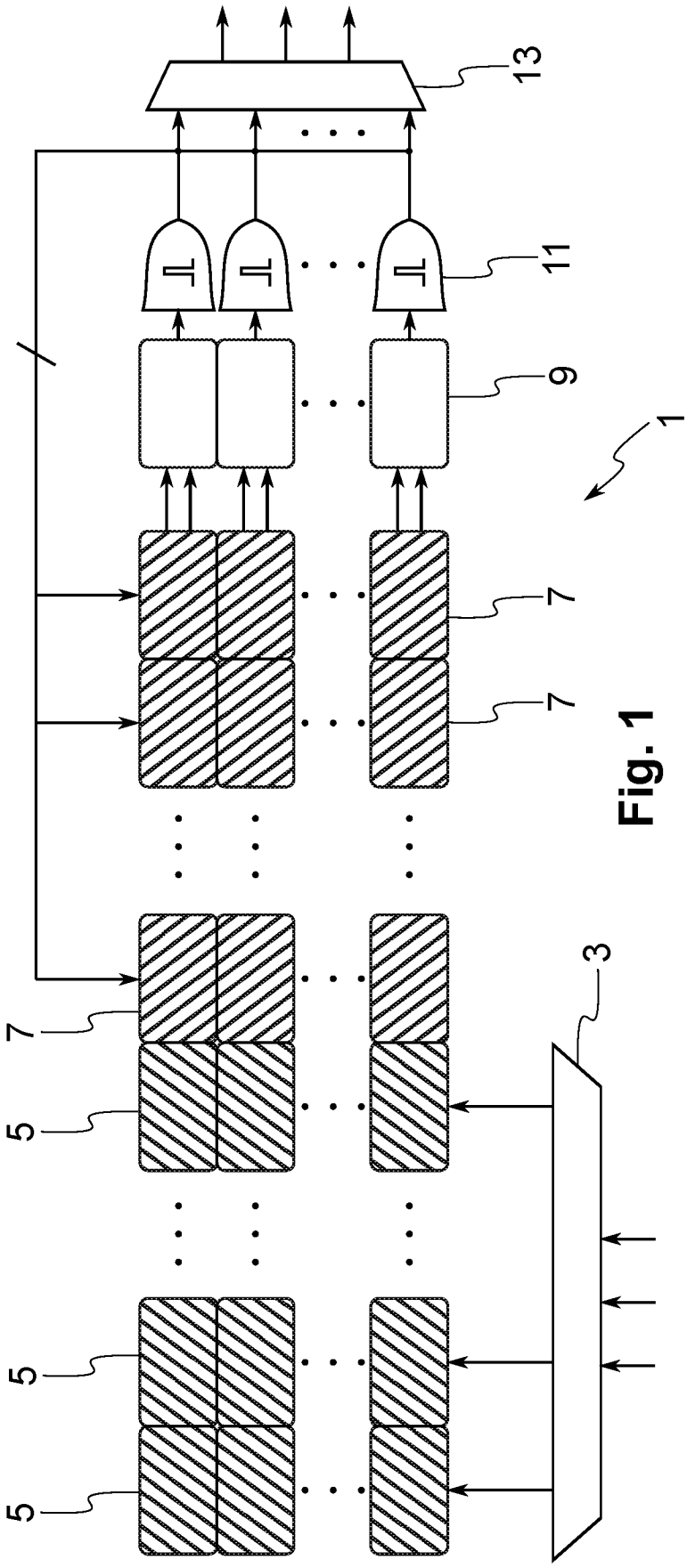


Fig. 1

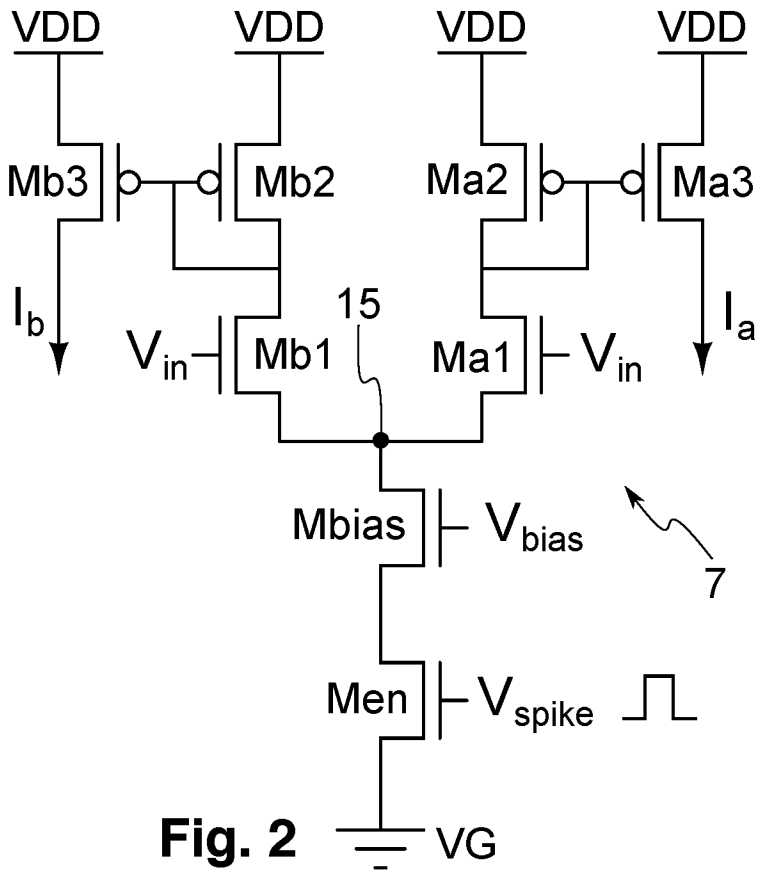

VDD
VDD
VDD
VDD
Mb3
Mb2
Ma2
Ma3
I_b
V_{in}
Mb1
15
Ma1
V_{in}
I_a
Mbias
V_{bias}
7
Men
V_{spike}
VG

Fig. 2

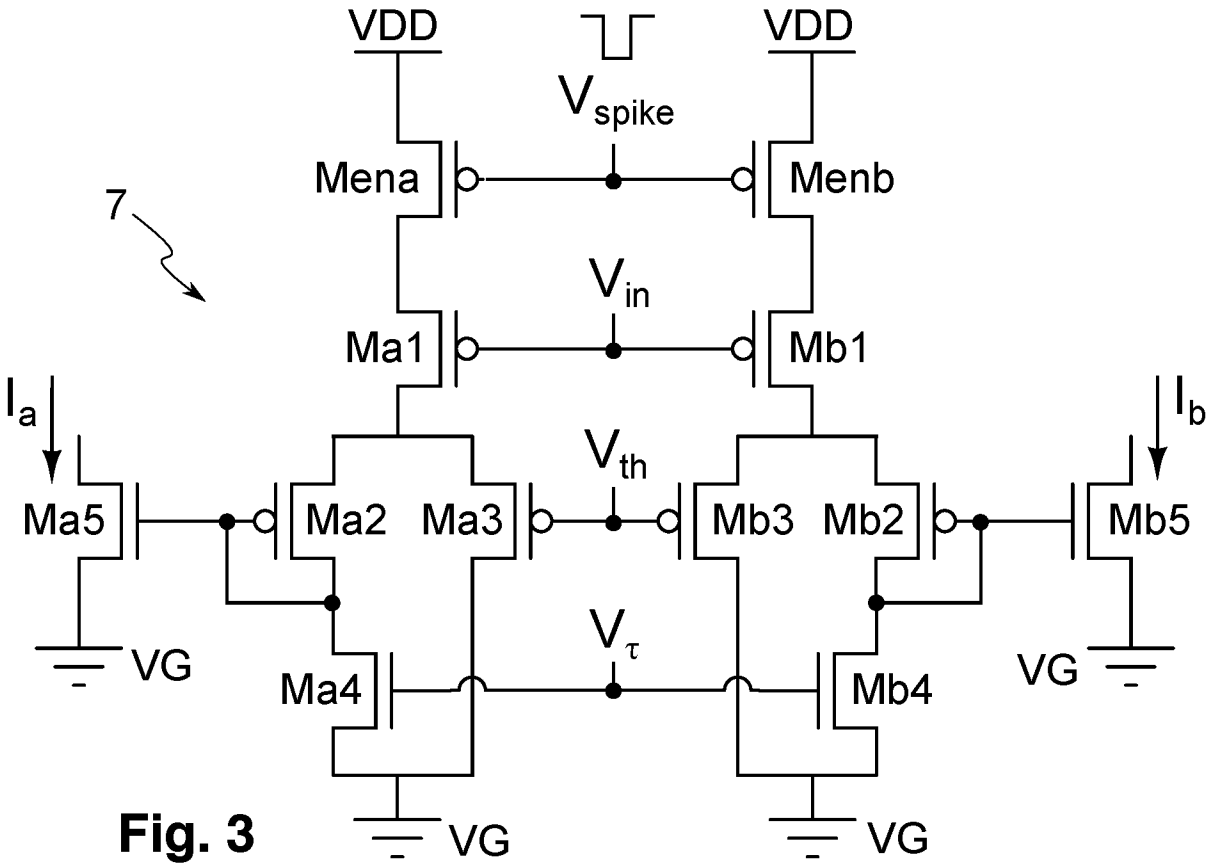

VDD
VDD
V_{spike}
Mena
Menb
7
V_{in}
Ma1
Mb1
I_a
I_b
V_{th}
Ma5
Ma2
Ma3
Mb3
Mb2
Mb5
V_τ
VG
Ma4
Mb4
VG
VG
VG

Fig. 3

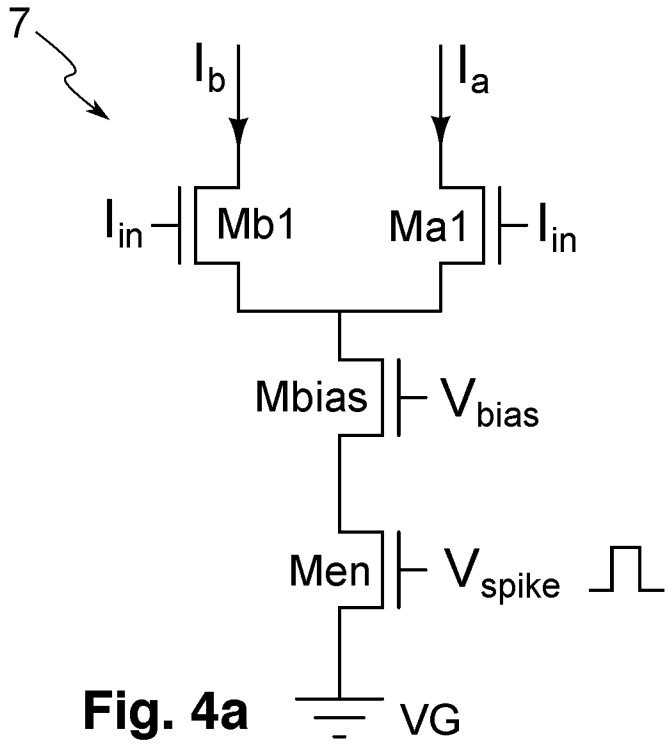

7
I_b
I_a
I_{in}
Mb1
Ma1
I_{in}
Mbias
V_{bias}
Men
V_{spike}
VG

Fig. 4a

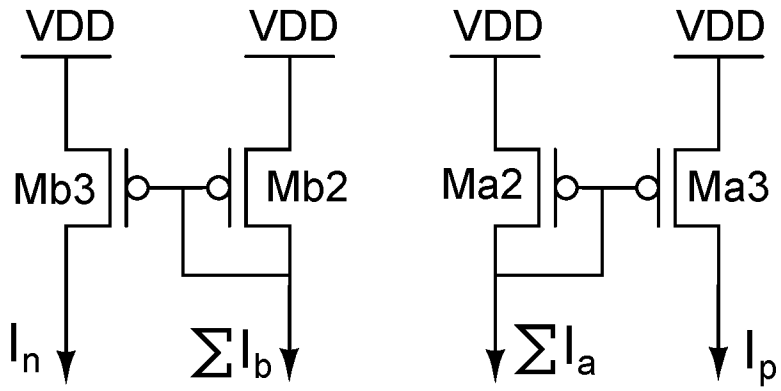

VDD
VDD
VDD
VDD
Mb3
Mb2
Ma2
Ma3
I_n
$\sum I_b$
$\sum I_a$
I_p

Fig. 4b

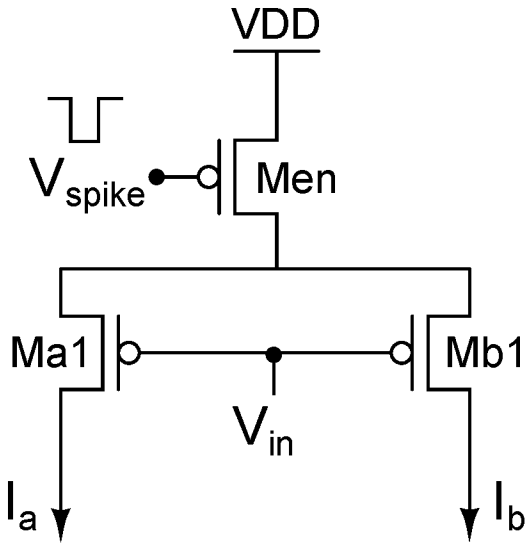


Fig. 5a

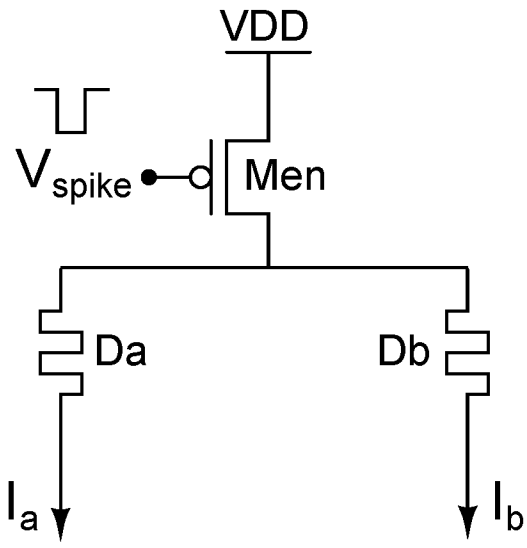


Fig. 5b

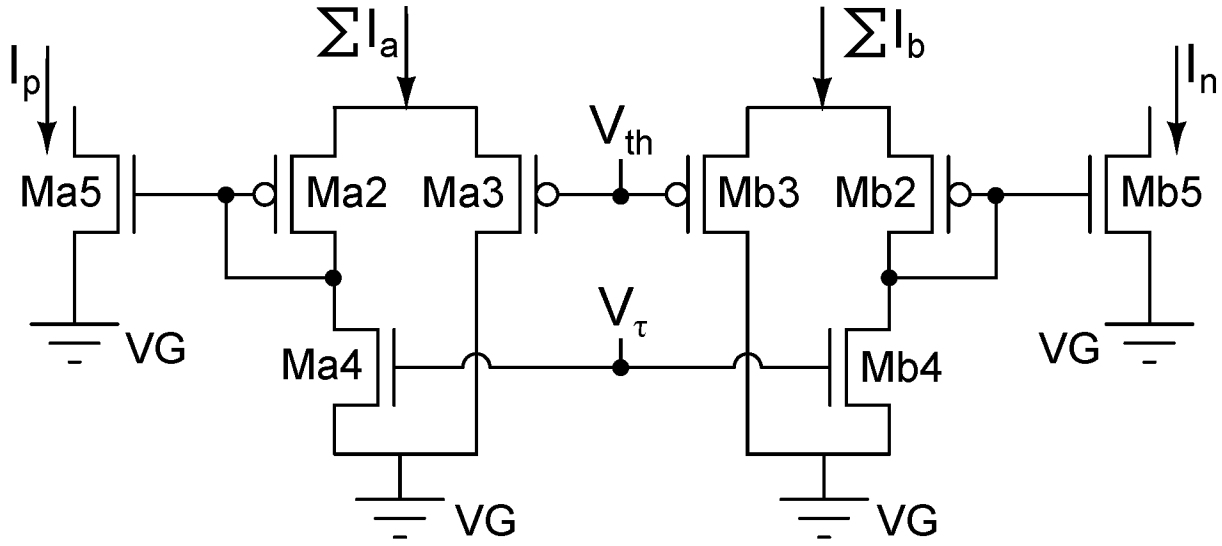


Fig. 5c

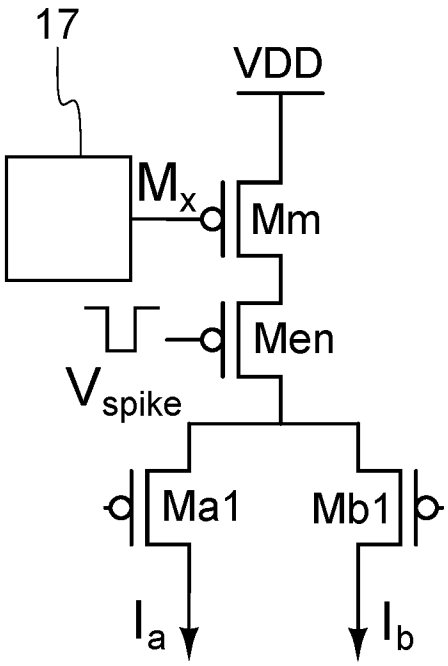

17
VDD
Mx
Mm
Men
Vspike
Ma1
Mb1
Ia
Ib

Fig. 6

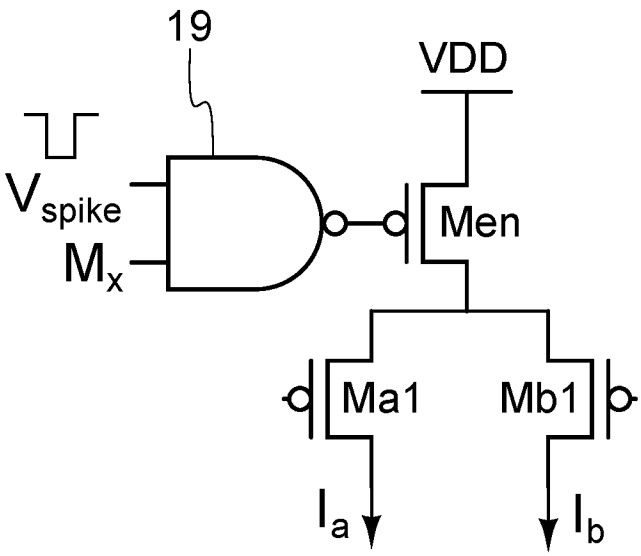

19
VDD
Vspike
Mx
Men
Ma1
Mb1
Ia
Ib

Fig. 7

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2020/050058

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06N3/063 G06N3/04
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CHIARA BARTOLOZZI ET AL: "Synaptic Dynamics in Analog VLSI", NEURAL COMPUTATION, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, US, vol. 19, no. 19, 1 October 2007 (2007-10-01), pages 2581-2603, XP002557203, ISSN: 0899-7667 page 2581 - page 2595 ----- -/--	1-22

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
7 April 2020	24/04/2020

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Cilia, Elisa

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	RICHTER OLE ET AL: "Device mismatch in a neuromorphic system implements random features for regression", 2015 IEEE BIOMEDICAL CIRCUITS AND SYSTEMS CONFERENCE (BIOCAS), IEEE, 22 October 2015 (2015-10-22), pages 1-4, XP032824893, DOI: 10.1109/BIOCAS.2015.7348416 [retrieved on 2015-12-04] abstract; figure 1 paragraph [000I] - paragraph [0III] paragraph [000V]	1-22
X	US 8 384 569 B2 (IPGOAL MICROELECT SICHUAN CO [CN]; ZHU GUOJUN [CN]) 26 February 2013 (2013-02-26)	1-5,7-9, 11,22
A	abstract; claims 1-20; figures 1, 3 column 1, line 37 - column 4, line 54	6,10, 12-15
X	US 2008/313250 A1 (LEE CHOONG-HOON [KR] ET AL) 18 December 2008 (2008-12-18)	1-8,10, 11,16,22
A	abstract; claims 1-20; figures 1-7 paragraph [0003] paragraph [0008] - paragraph [0021] paragraph [0048] - paragraph [0079] paragraph [0029] - paragraph [0036] paragraph [0091] - paragraph [0094]	9,12-15
A	US 5 247 479 A (YOUNG IAN [US]) 21 September 1993 (1993-09-21) the whole document	1-22
A	MELIKA PAYVAND ET AL: "A neuromorphic systems approach to in-memory computing with non-ideal memristive devices: From mitigation to exploitation", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 13 July 2018 (2018-07-13), XP081248190, DOI: 10.1039/C8FD00114F the whole document	1-22
A	EP 1 130 771 A1 (FUJI PHOTO FILM CO LTD [JP]) 5 September 2001 (2001-09-05) paragraph [0001] - paragraph [0048] paragraph [0060] - paragraph [0069] paragraph [0080] - paragraph [0085]	1-22

-/--

2

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	ELISABETTA CHICCA ET AL: "Neuromorphic Electronic Circuits for Building Autonomous Cognitive Systems", PROCEEDINGS OF THE IEEE., vol. 102, no. 9, 1 May 2014 (2014-05-01), pages 1367-1388, XP055602172, US ISSN: 0018-9219, DOI: 10.1109/JPROC.2014.2313954 the whole document -----	1-22
A	US 9 646 684 B1 (CALVETTI EMANUELA [IT] ET AL) 9 May 2017 (2017-05-09) the whole document -----	1-22

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/IB2020/050058

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 8384569 B2	26-02-2013	CN 101834586 A	15-09-2010
		US 2011260897 A1	27-10-2011
US 2008313250 A1	18-12-2008	DE 102008015578 A1	16-10-2008
		FR 2914076 A1	26-09-2008
		KR 20080086567 A	26-09-2008
		US 2008313250 A1	18-12-2008
US 5247479 A	21-09-1993	JP H05198184 A	06-08-1993
		US 5247479 A	21-09-1993
EP 1130771 A1	05-09-2001	DE 60119010 T2	14-09-2006
		EP 1130771 A1	05-09-2001
		JP 4544683 B2	15-09-2010
		JP 2001244747 A	07-09-2001
		KR 20010085673 A	07-09-2001
		US 2001017569 A1	30-08-2001
US 9646684 B1	09-05-2017	NONE	