Regular paper

An ultra-low power, ±0.3 V supply, fully-tunable Gaussian function circuit architecture for radial-basis functions analog hardware implementation

Marios Gourdouparis *, Vassilis Alimisis, Christos Dimas, Paul P. Sotiriadis

Department of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece

ARTICLE INFO

Keywords:
Gaussian function circuit
Bulk-controlled circuit
Fully tunable implementation
Ultra-low power design
Radial Basis Function

2010 MSC:
00-01
99-00

ABSTRACT

An ultra-low power (3.9 nW), low supply-voltage (0.6 V) and fully-tunable Gaussian-Bump circuit architecture for hardware-friendly implementation of Kernel functions is presented. The proposed architecture can be used to form hidden neuron cells for analog implementation of Radial Basis Functions in Neural Networks. It consists of only eleven transistors, all operating in sub-threshold. The Gaussian’s center, height and width are independently and electronically controlled. The proposed architecture is used as a building block to construct a 2-D Gaussian cascaded Bump structure, demonstrating its dimensional scalability. Proper operation, sensitivity and accuracy are confirmed via theoretical analysis and post-layout simulation results. The presented architectures were realized in TSMC 90 nm CMOS process and were simulated using the Cadence IC Suite.

1. Introduction

Radial Basis Functions (RBFs) are an especially useful tool in mathematics, with a variety of applications ranging from mathematical approximation and interpolation [1] to engineering applications [2]. RBFs are also extensively used in Machine Learning (ML). In RBF Neural Networks (NN), which perform efficiently in classification and function approximation problems, the activation function of the hidden layers’ neurons is an RBF. Moreover, they are also used as kernels in non-linearly separable cases of Support Vector Machine (SVM) algorithm, in which the kernel trick allows operation in a high-dimensional feature space [3].

RBF NN are commonly implemented in software [4–6]. However, in recent years traditional software realizations of ML algorithms are becoming ever more demanding in terms of memory and computational resources (power-hungry) [7,8]. A key focus then becomes the realization of such algorithms, when power constraints apply. In recent years there is a trend towards developing specialised hardware architectures for implementations of ML algorithms that meet those constraints. Apart from digital architectures [9], alternative approaches are being researched involving the use of analog integrated circuits (IC) [10–16]. Analog IC architectures have the advantage of low power consumption, low area and precise parallel computation due to the physical properties of analog circuits [14].

RBF kernels are suitable for analog hardware implementation. This is achieved via specific analog circuits which operate in sub-threshold. They are known in the literature as Gaussian function circuits or Bump circuits. The original Bump circuit was proposed by Delbruck [17] and it is a compact structure with only 8 transistors. It consists of two subcircuits, a differential pair and a current correlator. The current correlator provides a measure of similarity between the two currents of the differential pair and its output is a Gaussian function with respect to the input voltages of the differential pair [17].

The design of Gaussian function circuits generally aims to independent tunability of the Gaussian curve’s three characteristics, which are height, center (mean value) and width [17,18]. In addition, such circuits should be ultra-low power and low-area due to the fact that in system level implementations of multivariate kernels or RBF NN with many neurons in each hidden layer, many Bump cells have to operate in parallel fashion [19–21]. In RBF NN in particular, the tunability in width is a highly desirable feature as it allows the update in learning parameters.

This article is motivated by Delbruck’s Bump [17] and proposes a new architecture of an ultra-low power, fully tunable, bulk-controlled Gaussian function circuit, with supply voltage at only 0.6 V and power consumption at 3.9 nW, which are both lower in comparison to previous works. The width, height and mean value of the Gaussian function are independently programmable while only 11 transistors are used.

* Fully documented templates are available in the elsarticle package on CTAN.
* Corresponding author.
E-mail address: mariosgourd97@gmail.com (M. Gourdouparis).

https://doi.org/10.1016/j.aeue.2021.153755
Received 8 March 2021; Accepted 11 April 2021
Available online 17 April 2021
1434-8411/© 2021 Elsevier GmbH. All rights reserved.
The remainder of this article is organized as follows. Section 2 refers to related implementations and applications, while Section 3 introduces the proposed circuitry and explains the operation of the circuit based on the parameters. Moreover, the theoretical analysis of the proposed Gaussian function circuit is explained in Section 4. The behavior of the presented architecture is evaluated in Section 5, in TSMC 90 nm CMOS process and simulated using Cadence IC Suite. Section 6 discusses a comparison study between this work and previous implementations, and Section 7 concludes the article.

2. Related work

Delbruck’s Bump circuit [17], shown in Fig. 1, achieves tunability in height and center, while width is determined by the effective W/L ratio of the transistors, which also affects the height. There have been several implementations of Gaussian function circuits, with transistors operating in sub-threshold, many of them inspired by Delbruck’s Bump circuit architecture. There are realizations in which the Gaussian curve’s width is not electronically tunable. Instead, the width is tuned by setting a ratio of transistor W/L which necessitates the use of different Bump cells to generate different widths of the Bump [19,22]. An example of such an architecture with ultra-low power consumption (only 13.5 nW) is analyzed in [19].

The most compact Gaussian function circuit consisting of only four transistors, but without width tunability is presented in [23]. Moreover, there are compact Bump circuits using back-gate control voltage to achieve width tunability [24,25]. In contrast, there is a variety of complicated architectures for the realization of fully-tunable Gaussian functions. Examples of such circuits are synthesized using the translinear principle and the exponential characteristics of subthreshold MOS, with increased number of transistors [26–28]. Also, an architecture based on the translinear principle and BJT transistors is presented in [29].

Other complicated architectures use extra stages to achieve tunability of the Gaussian curve’s characteristics. These extra stages may be implemented with pseudo-differential transconductors [18], operational transconductance amplifiers [30], Digital to Analog Converters (DAC) [31], different values of series-connected resistances [32] or prescaling circuits with floating gate transistors [33,34]. E.g. a complicated Gaussian function architecture is presented in [35], consisting of DAC, operational amplifier, floating resistor, multiplier and exponentiant. Also, binary switches are incorporated in the Bump circuit of [36], which exhibits compensation on temperature and $V_{DS}$ variations and has peak output currents exceeding 1 $\mu$A. It is a modification of the architecture presented in [37].

Apart from the previous implementations in which transistors operate in sub-threshold, there are architectures in which transistors operate in strong inversion. There is a trade-off between low power consumption in sub-threshold and higher speed operation in strong inversion. Two Gaussian generation circuits consisting of transistors operating in both regions are presented in [34,38]. An implementation which consists of a symmetric current correlator and a differential pair with extra current sources is presented in [39], with the ability to operate both in strong and weak inversion regions.

There are also other architectures that consist of transistors operating exclusively in strong inversion region. There are also Gaussian circuits without width tunability [40,41]. More specifically, a CMOS and a BiCMOS RBF circuit [40] and a Bump circuit which consists of a voltage correlator [41] are presented. The appropriate tunability in Gaussian curve’s characteristics is achieved through complicated implementations. A complicated and accurate implementation of a Gaussian function circuit using a $\alpha$-order approximation (based on Taylor mathematical series) with current mode squaring circuits operating in saturation is presented in [42]. A Gaussian membership function architecture is implemented in [43] based on current mode squaring and exponential approximation circuits and in [44] based on a differential coupled amplifier (using two differential pairs). Differential pairs and a minimum value circuit are used for the realization of a Gaussian function output [45]. A modified current rectifier is proposed in [46] for the realization of the Gaussian function. Moreover, a different architecture based on current conveyors is proposed in [47].

Gaussian function circuits and Bump circuits are useful in a variety of applications. In modern neuromorphic systems there is a need for tunable Gaussian function circuits to implement weight update mechanisms [48] and adaptive stop learning procedures [25,49]. Tunable Gaussian function circuits have also been used in architectures implementing SVM [31,38,50], support vector domain description algorithm [51], k-means clustering algorithm [52], RBF NN [19,21], RBF classifiers [33] and Gaussian Kernels used in support vector regression implementations for approximate computing [53]. Memristive RBF NN architectures are also being investigated using hybrid CMOS-memristor Bump circuits [20]. Bump circuits have also been used in various sensor applications such as image edge detection [54] and unsupervised anomaly detection [32].

3. Proposed circuit architecture

The architecture of the original Bump circuit [17] in Fig. 1, can be modified to achieve electronically and adjustable width, independent of the other Bump parameters. This is done by the proposed architecture in Fig. 2 consisting of a differential difference pair ($M_{b1}$–$M_{b4}$) and a modified current correlator ($M_{1}$–$M_{4}$).

3.1. Differential difference pair

The differential difference pair consists of two differential pairs which produce currents with sigmodal shaped curves of adjustable slopes. The differential difference pair’s transistors $M_{b1}$ and $M_{b3}$ have their bulks connected to a control voltage $V_c$. Input voltage $V_{in}$ is connected to the gates of $M_{b1}$ and $M_{b3}$ while a parameter voltage $V_r$ is applied to the gates of $M_{b2}$ and $M_{b4}$. To increase the linearity of the differential difference pairs block, the ratio of the transistors’ $M_{b2}$/$M_{b1}$ and $M_{b4}$/$M_{b3}$ sizes is set to 2 instead of 1 [55–57].

Transistors’ dimensions are summarized in Table 1. The transistor sizes are fixed and selected so that they result in a good balance of the circuit’s performance. It is the topology and control nodes that offer the desirable tunability. The Bump circuit structure is biased with current $I_{bias}$. In our implementation, we set the power supply rails at $V_{DD} = -V_{SS} = 0.3 \, V$, and all transistors operate in the sub-threshold region.

The proposed architecture provides a Gaussian function output as shown in Fig. 3 (for $V_r = 0, V_c = -300 \, mV$ and $I_{bias} = 1 \, nA$).
In each differential pair, connecting the bulk of one of the two transistors to a control voltage $V_c$ results in a shift of the differential currents along the $V_{in}$ axis by altering $V_c$. The currents of the first differential pair ($M_{n1}$ and $M_{n2}$) are shifted in a symmetric way about the origin ($V_{in} = 0$) relative to the currents of the second differential pair ($M_{n3}$ and $M_{n4}$). As shown in Fig. 4 an increase in $V_c$ results in a shift of the current $I_{Mn1}$ transfer curve to the left. The same change is applied to current $I_{Mn2}$. For the second differential pair, an increase in $V_c$ results in a shift of the currents' transfer curve to the right, as shown in Fig. 5 for the case of $I_{Mn3}$.

Bulk-controlled implementations of Bump circuits are also presented in [24,25]. In contrast to those designs, in the proposed architecture there is summation of currents from the two differential pairs. Adding the currents flowing through $M_{n1}$ and $M_{n3}$ achieves the desired variability in the slope of current $I_1$, as shown in Fig. 6. Similarly for $M_{p1}$ and $M_{p3}$, currents $I_{Mn2}$ and $I_{Mn4}$ are summed, thus tunability in the slope of current $I_2$ is achieved, as shown in Fig. 7. The fact that the current correlator’s input currents $I_1$ and $I_2$ have adjustable slopes leads to the desired tunability in the Gaussian output curve’s width.

### 3.2. Modified current correlator

In the proposed implementation a modified current correlator has been used, in order to tackle the inherent asymmetries in the standard current correlator topology. In this work, the transistors’ dimensions are

![Fig. 2. Proposed Gaussian function circuit.](image)

![Fig. 3. Output current of the Gaussian function circuit for $V_c = 0$, $V_r = -300$ mV and $I_{bias} = 1$ nA (post-layout simulation).](image)

![Fig. 4. Displacement of current $I_{Mn1}$ via parameter voltage $V_c$, for $I_{bias} = 1$ nA and $V_r = 0$ mV (post-layout simulation).](image)

![Fig. 5. Displacement of current $I_{Mn3}$ via parameter voltage $V_c$, for $I_{bias} = 1$ nA and $V_r = 0$ mV (post-layout simulation).](image)

![Fig. 6. Tuning of current $I_1$ via parameter voltage $V_c$, for $I_{bias} = 1$ nA and $V_r = 0$ mV (post-layout simulation).](image)

### Table 1
MOS Transistors Dimensions.

<table>
<thead>
<tr>
<th>Block</th>
<th>W/L ($\mu$m/$\mu$m)</th>
<th>Current Correlator</th>
<th>W/L ($\mu$m/$\mu$m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{n1}, M_{n2}$</td>
<td>1.6/0.4</td>
<td>$M_{p1}, M_{p2}$</td>
<td>0.4/1.6</td>
</tr>
<tr>
<td>$M_{n3}, M_{n4}$</td>
<td>0.8/0.4</td>
<td>$M_{p3}, M_{p4}$</td>
<td>0.4/1.6</td>
</tr>
<tr>
<td>$M_{n5}, M_{n6}$</td>
<td>0.8/1.6</td>
<td>$M_{n5}, M_{n6}$</td>
<td>0.6/1.6</td>
</tr>
<tr>
<td>$M_{n7}, M_{n8}$</td>
<td>1.2/1.6</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

![Fig. 2. Proposed Gaussian function circuit.](image)

![Table 1](image)
respectively). A single current source
values. Thus, we choose
not all equal (non-symmetric implementation). While transistors \( M_{d1} \), \( M_{d2} \) and \( M_{d3} \) have \( \frac{W}{L} = 0.4 \ \mu m \) the transistor’s \( M_{d4} \) value of \( W/L \) has been changed. This modification enables the elimination of small dc offsets of the Bump’s transfer curve along the \( V_{in} \) axis, with \( \frac{W}{L} = 0.4 \ \mu m \) proving to be the optimal value for this purpose, as shown in Fig. 8.

In our implementation, we want to achieve more symmetrical results in the Gaussian function curve for the minimum bias current of 1 nA. The available values of width are multiples of a single transistor with \( W = 0.2 \ \mu m \). Through parametric simulations for different values of transistor’s width (\( M_{d4} \)) demonstrated in Fig. 8, it can be observed that for values of \( W \) greater than \( W = 0.6 \ \mu m \) there is a considerable offset of the Gaussian curve’s center along the \( V_{in} \) axis. Furthermore, comparing the two Gaussian curves for \( W = 0.6 \ \mu m \) and \( W = 0.4 \ \mu m \), the one with \( W = 0.6 \ \mu m \) is slightly more symmetrical across the whole range of \( V_{in} \) values. Thus, we choose \( W = 0.6 \ \mu m \) as the optimal value for our design (layout).

The height, width and center of the produced Gaussian function are electronically tuned via three circuit’s parameters (\( I_{bias}, V_c \) and \( V_r \) respectively). A single current source \( I_{bias} \) provides the bias current for both differential pairs and controls the height of the Bump while as explained above a control voltage \( V_c \) at the bulks of \( M_{d1} \) and \( M_{d2} \) alters the width. The center of the Gaussian function is set by the parameter voltage \( V_r \). The output current of the Bump reaches its maximum value when the input voltage matches the parameter voltage (\( V_{in} = V_r \)).

3.3. 2-D Implementation

A typical characteristic of Bump circuit architectures is the dimensional scalability. It is realized with two or more cascaded Bump circuits, in which the output current of one Bump circuit is used as bias current for the next identical Bump cell. Each Bump cell has each own input voltage \( V_{in} \) and parameter voltages \( V_c \) and \( V_r \), while only the first Bump cell is biased with a current \( I_{bias} \). A cascaded 2-D implementation using the proposed Gaussian function circuit as a basic component, is shown in Fig. 9. The scalability of the proposed Bump circuit makes it an interesting candidate for implementations of multivariate RBFs. Functions of this type are used as activation functions of the hidden layers’ neurons in an RBF NN architecture. An example of such RBF NN architectures is shown in Fig. 10. Thus, the proposed Gaussian function circuit could be used as a building block for analog hardware implementations of RBF NN architectures.

In practice, in system level implementations using such Gaussian function circuits as building blocks, the input and parameter voltages are generated by previous blocks of the system. Such blocks could be D/A Converters\[14,21,31,32,38\] or sensor systems such as analog circuit extraction blocks\[19\] or active pixel sensors\[54\].

4. Circuit theoretical analysis

In this Section, a mathematical analysis of the proposed Gaussian circuit is presented. All transistors operate in the sub-threshold region and we use the MOS model in\[22\], i.e. for the PMOS and NMOS are respectively:

\[
I_{max} = I_{o} e^{(V_{S}-V_{Th})/\kappa} \left( e^{(V_{Dr}-V_{Th})/\kappa} - e^{(V_{Dd}-V_{Th})/\kappa} \right) \quad (1)
\]

\[
I_{max} = I_{o} e^{(V_{S}-V_{Th})/\kappa} \left( e^{(V_{Dr}-V_{Th})/\kappa} - e^{(V_{Dd}-V_{Th})/\kappa} \right) \quad (2)
\]

where: \( \kappa_{p} \) and \( \kappa_{n} \) are the slope factors for PMOS and NMOS transistors respectively, \( V_{G}, V_{S}, V_{Dr} \) and \( V_{Dd} \) are the gate voltage, source voltage, drain voltage and bulk voltage respectively, \( V_{Th} \) is the thermal voltage and \( I_{o} \) and \( I_{o} \) are the characteristic currents (pre-exponential currents) for PMOS and NMOS transistors, respectively\[22\].

In this work, \( I_{o} \) and \( I_{o} \) are the pre-exponential currents of transistors \( M_{d3} \) and \( M_{d2} \) respectively. For every transistor we consider, we use a scaling factor (m), i.e. \( mI_{o} \) or \( mI_{o} \), to capture relative W/L value according to Table 1.

4.1. Differential difference pair analysis

In this work two transistors, \( M_{d3} \) and \( M_{d4} \), of the differential difference pair are bulk-controlled and their bulks are connected to the parameter voltage \( V_c \), as shown in Fig. 2. All four transistors \( M_{d3} \) to \( M_{d4} \) operate in saturation region with voltages \( V_{in} \). Thus, we are led to a simplified version of (2). More specifically for transistors \( M_{d3} \) and \( M_{d4} \) their saturation currents are given by the following expressions (using the appropriate aspect ratio W/L):

\[
I_{d3} = 2I_{o} e^{(V_{S}-V_{Th})/\kappa} \left( e^{(V_{Dr}-V_{Th})/\kappa} - e^{(V_{Dd}-V_{Th})/\kappa} \right) \quad (3)
\]

\[
I_{d4} = 2I_{o} e^{(V_{S}-V_{Th})/\kappa} \left( e^{(V_{Dr}-V_{Th})/\kappa} - e^{(V_{Dd}-V_{Th})/\kappa} \right) \quad (4)
\]

Transistors \( M_{d2} \) and \( M_{d3} \) have their bulks connected to \( V_{SS} \). Their saturation currents are given by the following expressions (using the appropriate aspect ratio W/L):

\[
I_{d2} = I_{o} e^{(V_{S}-V_{Th})/\kappa} \left( e^{(V_{Dr}-V_{Th})/\kappa} - e^{(V_{Dd}-V_{Th})/\kappa} \right) \quad (5)
\]

\[
I_{d3} = I_{o} e^{(V_{S}-V_{Th})/\kappa} \left( e^{(V_{Dr}-V_{Th})/\kappa} - e^{(V_{Dd}-V_{Th})/\kappa} \right) \quad (6)
\]

Transistors \( M_{d5} \) to \( M_{d7} \) operate in saturation as current mirrors. According to their relative W/L values, the bias current \( I_{bias} \) is equal to:

\[
I_{bias} = \frac{2I_{d3} + I_{d4}}{3} \quad (7)
\]
which is given by:

\[ I = \frac{3I_{\text{lat}}}{2 + e^{(\kappa_3 V_i - \kappa_3 V_{i0})/V_t}} \]

for the two differential pairs. Combining (3), (5) and (7) we conclude:

\[ I_{\text{lat}} = \frac{3I_{\text{lat}}}{2 + e^{\kappa_3 V_i - \kappa_3 V_{i0}}/V_t} \]

In order to simplify the expression, we set \( \Delta V = V_i - V_{i0} \) and \( V_{i1} = V_i - V_{SS} \) and the previous equation is transformed in the following way:

\[ I_{\text{lat}} = \frac{3I_{\text{lat}}}{2 + e^{\kappa_3(V_i - V_{i0})/V_t}} \]

In this step we combine (4), (6) and (8). The drain current of \( M_{p3} \) is given by:

\[ I_{\text{d}} = \frac{3I_{\text{lat}}}{2 + e^{\kappa_3(V_i - V_{i0})/V_t}} \]

The current \( I_1 \) is equal to the sum of \( I_{\text{lat1}} \) and \( I_{\text{lat2}} \), as shown in Fig. 2. The total expression of \( I_1 \) is:

\[ I_1 = \frac{3I_{\text{lat}}}{2 + e^{\kappa_3(V_i - V_{i0})/V_t}} + \frac{3I_{\text{lat}}}{2 + e^{\kappa_3(V_i - V_{i0})/V_t}} \]

The current \( I_2 \) is equal to the sum of \( I_{\text{d}1} \) and \( I_{\text{d}2} \), as shown in Fig. 2. By using the same methodology we can calculate the total expression of \( I_2 \) which is given by:

\[ I_2 = \frac{3I_{\text{lat}}}{2 + e^{\kappa_3(V_i - V_{i0})/V_t}} + \frac{3I_{\text{lat}}}{2 + e^{\kappa_3(V_i - V_{i0})/V_t}} \]

4.2. Modified current correlator analysis

The current correlator’s transistors’ dimensions are shown in Table 1. Supposing that the output node’s voltage (drain of transistor \( M_{p3} \)) is sufficiently low to ensure operation in saturation region, its drain current, using (1), is given by the following expression:

\[ I_{\text{out}} = I_{\text{d}3} = I_{\text{o}3} e^{V_{D2} - V_{S2} + (V_i - V_{i0})/V_t} \]

where \( V_{D2} \) and \( V_{S2} \) are drain voltage of transistors \( M_{p3} \) and \( M_{p4} \). Transistor’s \( M_{p3} \) drain current is given by:

\[ I = I_{\text{d}3} = I_{\text{o}3} e^{V_{D2} - V_{S2} + (V_i - V_{i0})/V_t} \]

Transistor’s \( M_{p4} \) current is given by:

\[ I = I_{\text{d}4} = I_{\text{o}4} e^{V_{D4} - V_{S4} + (V_i - V_{i0})/V_t} \]

Transistor \( M_{p4} \) operates in triode region. The drain current of transistor \( M_{p4} \) is equal to the drain current of transistor \( M_{p3} \) and given by:

\[ I_{\text{d}4} = \frac{3I_{\text{lat}}}{2 + e^{\kappa_3(V_i - V_{i0})/V_t}} \left( e^{(V_i - V_{i0})/V_t} - e^{(V_{D2} - V_{S2})/V_t} \right) \]

\[ I_{\text{out}} = I_{\text{d}4} = I_{\text{o}4} e^{V_{D4} - V_{S4} + (V_i - V_{i0})/V_t} \]

where \( V_{D2} = V_{D4} \) and \( I_{\text{out}} = I_{\text{d}4} = I_{\text{o}4} \), as shown in Fig. 2. Solving (14)–(16) for \( V_{D2} \), \( V_{D4} \), and \( V_{S2} \), and replacing them into (18) we get:

\[ I_{\text{out}} = \frac{3}{2} I_1 \left( 1 - \frac{I_{\text{out}}}{I_1} \right) \]

The expression of current correlator for the proposed topology is expressed in the following way:

\[ I_{\text{out}} = I_1 + \frac{3I_1}{2} \]
4.3. Bump Circuit Analysis

In order to simplify the expression (12) and (13) we set \( x = \kappa_n \Delta V/V_T \) and 
\[ y = (\kappa_n - 1)\frac{V_{in}}{V_T}; \]
\[ I_1 = \frac{3M_{bias}}{2} \left( \frac{4 + 4e^{y} + 4e^{-y} + 2y^2}{2 + 4e^{y} + 4e^{-y} + 2y^2} \right) \]  
(21)

By using the same methodology the simplified expression of \( I_2 \) which is given by:
\[ I_2 = \frac{3M_{bias}}{2} \left( \frac{4e^{y} + 4e^{-y} + 2y^2}{2 + 4e^{y} + 4e^{-y} + 2y^2} \right) \]  
(22)

Combining (20) with (21) and (22):
\[ I_{out} = \frac{3M_{bias}}{2} \left( \frac{(6e^{-y} + 3M)(2e^{y} + M)}{2(2e^{y} + 6e^{y} + 4e^{-y} + 5M)} \right) \]  
(23)

Finally, the output current of the fully tunable Bump circuit is expressed using hyperbolic cosine equation as:
\[ I_{out} = \frac{3M_{bias}}{2} \left( \frac{12 + 3M^2 + 12MCoshx}{(2Coshx + M)(6e^{y} + 4e^{-y} + 5M)} \right) \]  
(24)

where intermediate variable \( M \) is given by:
\[ M = 2e^{-y} + \frac{e^y}{2} = 2e^{-(\kappa_n - 1)(V_{in} - V_{o}))/y_T} + \frac{\kappa_n - 1)(V_{in} - V_{o})/y_T}{2} \]  
(25)

and \( x \) has been defined as:
\[ x = \frac{\kappa_n(V_{in} - V_{o})}{V_T} \]  
(26)

The Gaussian circuit’s current \( I_{out} \) depends on the input voltage \( V_{in} \), and parameters \( M_{bias}, V_{T} \) and \( V_{C} \). The theoretical output current of the Gaussian circuit, according to (24), is presented in Figs. 11-13. Each parameter \( V_{C}, V_{T} \) and \( M_{bias} \) independently tunes a characteristic of the Gaussian curve. We alter the value of one parameter while the others are kept constant. Parameter \( V_{C} \) adjusts the Gaussian curve’s width as shown in Fig. 11. Parameter \( V_{T} \) sets the mean value (center) of the Gaussian function’s output, as shown in Fig. 12. Parameter \( M_{bias} \) scales the height of the Gaussian curve, as shown in Fig. 13. The results of the theoretical analysis illustrate the correct operation of the proposed Gaussian function circuit.

5. Simulation results

The proposed ultra-low power, low-voltage, fully tunable, bulk-controlled Gaussian function circuit has been designed in TSMC 90 nm CMOS process, using the Cadence IC design suite. The power supply rails are \( V_{DD} = -V_{SS} = 0.3 \) V, and all transistors operate in the sub-threshold region. The Gaussian circuit’s simulation results are from post-layout simulations. The layout of the proposed 1-D Gaussian function architecture (Fig. 2) is shown in Fig. 14, where the area is 50.45 \( \mu \)m \( \times \) 25.85 \( \mu \)m. In order to deal with manufacturing considerations and mismatches, dummy transistors were used and transistor matching is applied only for transistors with the same length (L) based on the common-centroid technique.

There are cases where, the Gaussian function circuit could be part of a system with higher supply voltage or in a sensitive place of a system (need for elimination of PVT variations). In such cases it would be preferable to use extra circuits in order to bias the Gaussian function circuit. These circuits are called Bandgap References [58–66]. They have the ability to deal with PVT variations [67] and provide a stable and appropriate ultra-low bias voltage. Additionally, there are Bandgap References which can produce a bias current of 1 nA [59], less than 1nA [60] and up to 40 nA [66], which can be mirrored and bias the Gaussian function circuit.

The width tunability of the proposed Gaussian function circuit, via parameter voltage \( V_{C} \), is shown in Fig. 15, for constant values of \( I_{bias} = 1 \) nA and \( V_{C} = 0 \) mV. An increase in parameter voltage \( V_{C} \) leads to an increase in the Gaussian curve’s width. The mean value of the derived Gaussian function is determined by voltage \( V_{C} \), as illustrated in Fig. 16, for constant values of \( I_{bias} = 1 \) nA and \( V_{C} = -300 \) mV. Proper operation is achieved for a wide parameter voltage \( V_{C} \) range. Its values are between \( V_{Ran} = -250 \) mV and \( V_{Ran} = 250 \) mV (83.3% of the power supply range). The height of the Bump output current is set by the bias current \( I_{bias} \) as shown in Fig. 17, for constant \( V_{C} = 0 \) and \( V_{C} = -300 \) mV. The tunability of the Gaussian function’s characteristics derived from post-layout simulations.
simulations and shown in Figs. 15–17 matches the expected behavior depicted in Figs. 11–13. The output current of two cascaded Bump cells is represented in 3-D space and is depicted in Figs. 18–21. The independent tunability of the Gaussian curve’s characteristics (width, height, center) is also achieved for 2-D RBFs. The first of the two cascaded Bump circuits of the architecture in Fig. 9, is biased with $I_{\text{bias}} = 2 \text{nA}$. Setting control parameters $V_{c1} = V_{c2} = 300 \text{ mV}$ and $V_{r1} = V_{r2} = 0 \text{ V}$, maximum width is achieved as shown in Fig. 18. By keeping $I_{\text{bias}}$, $V_{r1}$ and $V_{r2}$ values constant while altering $V_{c1} = V_{c2} = -300 \text{ mV}$, the Gaussian curve’s width is independently adjusted, as shown in Fig. 19. Keeping the parameter voltages constant at $V_{c1} = V_{c2} = 300 \text{ mV}$ and $V_{r1} = V_{r2} = 0 \text{ mV}$ and increasing $I_{\text{bias}} = 4 \text{nA}$, height scaling of the 2-D output is achieved, as shown in Fig. 20. The tunability of the RBF’s center is presented in Fig. 21, by setting parameter voltages $V_{r1} = V_{r2} = 100 \text{ mV}$ while the other parameters are the same as in Fig. 20. The 2-D implementation’s results confirm the desirable scalability of the proposed circuit.
The sensitivity behavior has been evaluated using the Monte-Carlo analysis tool for \( N = 100 \) points. The corresponding histogram for the Bump circuit’s center of voltage is shown in Fig. 22. The center of the voltage is \( V_{\text{mean}} = 1.7 \text{ mV} \), and the standard deviation is \( \sigma = 4.3 \text{ mV} \). It confirms the correct performance and accuracy of the proposed circuit.

The simulation results confirm the proper operation and performance of the proposed Gaussian function circuit, in accordance with the theoretical analysis. There are specific factors resulting in slight variations between the theoretical results as shown in Figs. 11–13 and simulation results as shown in Figs. 15–17. In theoretical analysis the sub-threshold slope factors \( \kappa_1 \) and \( \kappa_2 \) are considered to be constant, which is typical in approximate analysis of sub-threshold circuits. However, an increase in gate-to-bulk voltage results in a small increase in the slope factors’ values [22]. This approximation justifies small variations in the form of the Gaussian function output. For transistors \( M_{\text{DS}} - M_{\text{DS}} \) (current mirrors) the current mirroring ratio for small currents (1 nA–5 nA) diverges from the aspect ratio (2). This is illustrated in the height difference of the Gaussian function output between theoretical and simulation results. Thus, in simulation results the output current exceeds 1 nA for bias current \( I_{\text{bias}} = 1 \text{ nA} \). This is vital for the scalability of the proposed architecture without having degradation of the output current. In Figs. 11–13 and Figs. 15–17 there is a small dc offset of up to 10 nV for increased bias current (more than 5 nA). This is a result of the non-symmetric current correlator topology, which achieves the appropriate symmetry in small currents \( I_{\text{bias}} = 1 \text{ nA} \) and sufficient current scaling.

6. Comparison study and discussion

The proposed Gaussian function circuit is compared with recent literature works in terms of performance and design characteristics and the results are summarized in Table 2. All the implementations are in CMOS process except of [20] which is designed in LT-SPICE with PTM transistor models and MS memristor model and [24] which uses quad MOS transistor arrays ALD1106 and ALD1107, manufactured by Advanced Linear Devices (ALD).

Low power consumption is a vital characteristic of Bump circuits that are used for the implementation of RBF NNs. The previous works operating with ultra-low power consumption are [18] at 18.9 nW and [19] at only 13.5 nW. Our work, provides a significant improvement in power consumption compared to literature, operating at only 3.9 nW (reduction of power consumption by 71.1% compared to 13.5 nW).

Power supply voltages used in literature implementations range from 5 V down to 0.7 V. The minimum power supply among the previous works is achieved in [20,27] (0.7 V) and [34] (0.75 V). In the proposed architecture, power supply voltage is further reduced at only 0.6 V. Our implementation is biased with minimum \( I_{\text{bias}} = 1 \text{ nA} \), which is smaller than 2 nA [24] and equal to 1 nA[18].

The implementation with the minimum number of transistors (only 4) is presented in [23], but it lacks tunability in width. The proposed Gaussian function circuit, is a compact and low area implementation, consisting of only 11 transistors. Architectures with similar number of transistors are [34] (8 transistors) [24] (10 transistors) and [18,38] (14 transistors). The compact Bump circuit in [19] (9 transistors) uses extra logic and multiplexer circuits to update the output curve’s width, while for the same purpose in [31] (11 transistors) an extra DAC is used. Also, in [21] (13 transistors) extra switches and multiplexer circuits are used while in [44] (14 transistors) an extra stage for the controlling current implementation is used in order to achieve the desirable behavior. The area of the proposed circuit (layout) is 1304 \( \mu \text{m}^2 \) and only [18] has a smaller layout area (988 \( \mu \text{m}^2 \)), while in other implementations [19,26,27,33,46] it ranges from 2475 \( \mu \text{m}^2 \) up to 13054 \( \mu \text{m}^2 \).
7. Conclusion

In this work, an ultra-low power, low voltage and compact Gaussian function circuit was presented. Three circuit’s parameters are used in order to independently tune the height, width and center of the Gaussian curve. The proposed circuit’s ultra-low power consumption allows for low power high dimensional RBF implementations, by using arrays of cascaded Bump circuits. Thus, it can be used as a building block for low power and low area analog hardware RBF NN accelerators. Theoretical analysis and both 1-D and 2-D post-layout simulation results confirm the proper operation of the proposed architecture.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme Human Resources Development, Education and Lifelong Learning in the context of the project Strengthening Human Resources Research Potential via Doctorate Research (MIS-500432), implemented by the State Scholarships Foundation (IKY).

Table 2

<table>
<thead>
<tr>
<th>Technology</th>
<th>Power Consumption (nW)</th>
<th>Power Supply (V)</th>
<th>Minimum ( I_{\text{min}} ) (μA)</th>
<th>No of Transistors</th>
<th>Ind. Param. Tunability</th>
</tr>
</thead>
<tbody>
<tr>
<td>This work</td>
<td>3.9</td>
<td>0.6</td>
<td>1</td>
<td>11</td>
<td>YES</td>
</tr>
<tr>
<td>[18] 130 nm</td>
<td>18.9</td>
<td>3</td>
<td>1 nA</td>
<td>14</td>
<td>YES</td>
</tr>
<tr>
<td>[19] 180 nm</td>
<td>13.5</td>
<td>0.9</td>
<td>40 nA</td>
<td>9</td>
<td>YES</td>
</tr>
<tr>
<td>[20] Discrete</td>
<td>4.1 μW</td>
<td>0.7</td>
<td>1 μA</td>
<td>22</td>
<td>YES</td>
</tr>
<tr>
<td>[21] 130 nm</td>
<td>10.5 μW</td>
<td>1.2</td>
<td>–</td>
<td>+13</td>
<td>YES</td>
</tr>
<tr>
<td>[23] 180 nm</td>
<td>–</td>
<td>1.3 – 2 V</td>
<td>–</td>
<td>4</td>
<td>NO</td>
</tr>
<tr>
<td>[26] 350 nm</td>
<td>650 nW</td>
<td>1.3</td>
<td>50 nA</td>
<td>17</td>
<td>YES</td>
</tr>
<tr>
<td>[27] 180 nm</td>
<td>350 nW</td>
<td>0.7</td>
<td>50 nA</td>
<td>31</td>
<td>YES</td>
</tr>
<tr>
<td>[31] 180 nm</td>
<td>–</td>
<td>1.8</td>
<td>100 nA</td>
<td>+11</td>
<td>YES</td>
</tr>
<tr>
<td>[33] 500 nm</td>
<td>90 μW</td>
<td>3.3</td>
<td>–</td>
<td>19</td>
<td>YES</td>
</tr>
<tr>
<td>[34] 180 nm</td>
<td>160 nW</td>
<td>0.75</td>
<td>35 nA</td>
<td>8</td>
<td>YES</td>
</tr>
<tr>
<td>[38] 180 nm</td>
<td>–</td>
<td>1.8</td>
<td>50 nA</td>
<td>14</td>
<td>YES</td>
</tr>
<tr>
<td>[42] 180 nm</td>
<td>100 μW</td>
<td>1</td>
<td>10 μA</td>
<td>30</td>
<td>YES</td>
</tr>
<tr>
<td>[43] 350 nm</td>
<td>–</td>
<td>3.3</td>
<td>10 μA</td>
<td>45</td>
<td>YES</td>
</tr>
<tr>
<td>[44] 350 nm</td>
<td>220 μW</td>
<td>3.3</td>
<td>9 μA</td>
<td>+14</td>
<td>YES</td>
</tr>
<tr>
<td>[45] 180 nm</td>
<td>27 μW</td>
<td>1.8</td>
<td>2 μA</td>
<td>15</td>
<td>YES</td>
</tr>
<tr>
<td>[46] 180 nm</td>
<td>23.7 μW</td>
<td>2</td>
<td>5 μA</td>
<td>32</td>
<td>YES</td>
</tr>
</tbody>
</table>

Fig. 22. Center value sensitivity via Monte-Carlo simulation.

References
