A Spin Wave-Based Approximate 4:2 Compressor

Seeking the most energy-efficient digital computing paradigm

Mahmoud, Abdulqader Nael; Vanderveken, Frederic; Ciubotaru, Florin; Adelmann, Christoph; Hamdioui, Said; Cotofana, Sorin

DOI
10.1109/MNANO.2021.3126095

Publication date
2022

Document Version
Final published version

Published in
IEEE Nanotechnology Magazine

A Spin Wave-Based Approximate 4:2 Compressor

Seeking the most energy-efficient digital computing paradigm.

IN THIS ARTICLE, WE PROPOSE an energy-efficient spin wave (SW)-based approximate 4:2 compressor including three- and five-input majority gates. We validate our proposal by means of micromagnetic simulations and assess and compare its performance with state-of-the-art SW 45-nm CMOS and spin-CMOS counterparts. The evaluation results indicate that the proposed compressor consumes 31.5% less energy than its accurate SW-design version. Furthermore, it has the same energy consumption and error rate as a directional coupler (DC)-based approximate compressor, but it exhibits a

©SHUTTERSTOCK.COM/STILESTA

ABDULQADER NAEL MAHMOUD, FREDERIC VANDERVEKEN, FLORIN CIUBOTARU, CHRISTOPH ADELMANN, SAID HAMDIOUI, AND SORIN COTOFANA

Digital Object Identifier 10.1109/MNANO.2021.3126095
Date of current version: 30 November 2021

3× shorter delay. In addition, it consumes 14% less energy while having a 17% lower average error rate than its approximate 45-nm CMOS counterpart. When compared with other emerging technologies, the proposed compressor outperforms the approximate spin-CMOS-based compressor by three orders of magnitude in terms of energy consumption while providing the same error rate. Finally, the proposed compressor requires the smallest chip real estate measured in terms of devices.

BEYOND CMOS

The information technology revolution has led to a rapid raw data increase, requiring high-performance computing platforms for data processing [1]. Up-to-date downscaling CMOSs have been effective for satisfying these requirements. However, Moore's law has nearly reached its economical end, as CMOS feature size reduction is becoming increasingly difficult due to leakage, reliability, and cost issues [2]. As a result, different technologies have been investigated to replace CMOSs, such as graphene devices [3], memristors [4], and spintronics [5]. In this article, we study one type of spintronics—SW technology—that appears to open the way toward the most energy-efficient digital computing paradigm [6]–[9]. SW-based computing is promising for three main reasons [6]–[9]: 1) it has ultralow energy consumption potential because it does not rely on electron movements but on spinning around the magnetic field orientation [6]–[9], 2) it is highly scalable because SWs' wavelength can reach the nanometer scale [6]–[9], and 3) it has an acceptable delay [6]–[9]. As a consequence of these promising features, different research groups have made use of SW interaction to build logic gates and circuits.

The first experimental SW logic gate was an inverter designed by utilizing a Mach–Zehnder interferometer [10]. Moreover, the Mach–Zehnder interferometer has been used to build single-output majority, (N)AND, (N)OR, and X(N)OR gates [10], while multioutput SW logic gates have been introduced in [9] and [11]–[13]. Furthermore, multifrequency logic gates that enhance SW computing and storage capabilities have been proposed in [8] and [14], and wave pipelining has been achieved with pulse mode operation in the SW domain by employing four cascaded majority gates [15]. In addition, different SW circuits have been demonstrated at a conceptual level [16], in simulation [7], [17], and in practical millimeter-scale prototypes [18]. Different network-alike approaches have been suggested, e.g., reconfigurable meshes [19], cellular nonlinear networks [20], [21], and systolic arrays [22], which enable the implementation of SW-based neuromorphic computing approaches suggested in, e.g., [23]–[29]. 3D structures have been experimentally realized due to the development of focused electron beam-induced deposition [30], which opens the road for the implementation of SW-based human brain-inspired neuromorphic networks [30], [31]. All the aforementioned logic gates and circuits were designed to provide accurate results; however, many applications, such as multimedia processing and social media, are error tolerant, and within certain limits, they function correctly [32]. Hence, those applications can benefit from approximate computing circuits, which significantly save energy, delay, and area.

Based on the previous discussion of SW technology's potential and approximate computing benefits, one can conclude that SW approximate circuits are of great interest. With this view, and given that multiplication is heavily utilized in error-tolerant applications and that fast, state-of-the-art multipliers are built with 4:2 compressors [33], we introduce a novel approximate SW 4:2 compressor. The article's main contributions can be summarized as follows:

- We develop and design an approximate SW 4:2 compressor consisting of two majority gates that provides an average error rate of 31%.
- We enable DC-free approximate circuit design by demonstrating that majority gates can be directly cascaded, i.e., without amplitude normalization and domain conversion, to form a 4:2 compressor with no additional average error rate penalty.
- We validate the proposed 4:2 compressor by means of MuMax3 micromagnetic simulations.
- We demonstrate superiority: the proposed approximate SW 4:2 compressor's performance is assessed and compared with state-of-the-art SW 45-nm CMOS and spin-CMOS counterparts. The results indicate that the proposed compressor saves 31.5% more energy than the accurate SW design, whereas it has the same energy consumption and error rate as the approximate compressor with a DC while being 3× faster. In addition, the proposed compressor consumes 14% less energy while providing 17% less error when compared with the approximate 45-nm CMOS counterpart. Moreover, it outperforms the approximate spin-CMOS equivalent design by three orders of magnitude in terms of energy while having the same error rate. Finally, it requires the least chip real estate.

SW-BASED TECHNOLOGY FUNDAMENTALS AND COMPUTING PARADIGM

Magnetization dynamics when the magnetic material magnetization is out of equilibrium are described by the Landau–Lifshitz–Gilbert equation [6]:

$$\frac{d\vec{M}}{dt} = -|\gamma|\mu_0(\vec{M}\times\vec{H}_{\text{eff}}) + \frac{\alpha}{M_s}\left(\vec{M}\times\frac{d\vec{M}}{dt}\right), \quad (1)$$

where γ, is the gyromagnetic ratio, μ_0 is the vacuum permeability, M is the magnetization, M_s is the saturation magnetization, α is the damping factor, and H_{eff} is the effective field, which consists of external, exchange, demagnetizing, and magnetocrystalline fields.

Equation (1) has wave-like solutions under small magnetic disturbances, which are called *SWs*. These can be seen as collective excitations of the magnetization within the magnetic material [6]. An SW, as with any other wave, is described by its amplitude A, phase ϕ, wavelength λ, wavenumber $k = 2\pi/\lambda$, and frequency f, as illustrated in Figure 1 [6]. The SW frequency and wavenumber are linked by

the so-called dispersion relation, which plays a fundamental role during the SW circuit design process [6]. The dispersion relation slope determines the SW group velocity $\mathbf{v}_g = \nabla_{\mathbf{k}}\omega$, where ω is the angular frequency and describes the wave energy flow direction and speed, and the wave phase front propagation direction and speed are calculated by the phase velocity $\mathbf{v}_p = \mathbf{k}\omega/\|\mathbf{k}\|^2$ [6].

SW exchange interaction is dominant for large k values (short wavelengths), whereas dipolar interaction is dominant for small k values (long wavelengths) [6]. In the exchange regime, the dispersion relation exhibits a quadratic behavior regardless of the magnetization direction $\omega_{n,ex} = \omega_M \lambda_{ex} k_{tot}^2$, where $\omega_M = \gamma\mu_0 M_0$, M_0 is the static magnetization component, $k_{tot}^2 = k^2 + k_{nn}^2$, $k_{nn} = nn\pi/w$ is the quantized wavenumber, and nn is the mode number. In contrast, in the dipolar regime, the SW dispersion relation in a thin film becomes $\omega_{nn,dip} = \sqrt{\omega_0(\omega_0 + \omega_M F)}$, where $\omega_0 = \gamma\mu_0 H_0$, H_0 represents the static component of the effective magnetic field, and F denotes a factor that takes into account the demagnetization field. Here, F also depends on the magnetization orientation, which results in anisotropic SW properties [6].

Three common types of dipolar SWs, whose classification depends strongly on the static magnetization orientation and SW propagation direction, exist, i.e., surface SWs (SSWs), forward volume waves (FVSWs), and backward volume waves (BVSWs). For SSWs, the static magnetization and propagation directions are perpendicular to each other and lie in the waveguide plane. In contrast, for FVSWs, the static magnetization is perpendicular to the propagation direction and waveguide plane. On the other hand, for BVSWs, the static magnetization and propagation directions are parallel and lie in the waveguide plane [34]. Note that because the dispersion relation slope is positive for SSWs and FVSWs and negative for BVSWs, the group and phase velocities are in the same direction for SSWs and FVSWs and opposite directions for BVSWs [6], [34].

Generally speaking, information can be encoded in the SW amplitude and phase at different frequencies [6], [8], while the interaction between SWs coexisting in the same waveguide is governed by the interference principle. Figure 2(a) shows a simulation where two SWs are interacting: if they have the same phase, i.e., $\Delta\phi = 0$, they interfere constructively, resulting in a larger-amplitude SW. On the other hand, if they have different phases, i.e., $\Delta\phi = \pi$, they interfere destructively, producing a diminished-amplitude SW. Due to their nature, SWs provide natural support for majority function evaluation, as the interference of an odd number of SWs emulates a majority decision. For instance, if three SWs with the same amplitude, frequency, and wavelength interfere, the result is a zero-phase SW (logic 0) if no more than one of the waves has a π phase. In the other case, they result in a π-phase SW (logic 1), which is equivalent to the behavior of a three-input majority gate. Note that a CMOS three-input majority gate implementation requires 18 transistors, while SW technology requires only one waveguide. We observe that if the SWs have a varying A, λ, and f, their interaction results in more sophisticated interference patterns, which might open different SW-based computation paradigms. However, in this article, we consider only the interaction of SWs with phase-encoded information; i.e., logic 0 and logic 1 are represented by zero and π phases, respectively.

Figure 2(b) describes an SW device consisting of an excitation region I, waveguide B, functional region FR, and detection region O [6], [35]. The SW can be excited by means of, e.g., microstrip antennas and magnetoelectric (ME) cells in region I [6], [35]. Traditionally, an SW is excited by sending a current through a microstrip antenna to create an alternating Oersted magnetic field, which results in a magnetic torque that creates the SW [6], [35]. However, a more power-efficient way is voltage-controlled SW excitation by means of a so-called ME cell introduced in [16]. The waveguide is made out of a ferromagnetic material, e.g., permalloy, yttrium iron garnet, and cobalt iron boron, which must be properly chosen since it has a direct impact on SW properties [6], [35]. SWs can be amplified and normalized, and they can interfere with other SWs within the same waveguide, and the resulting SW can be

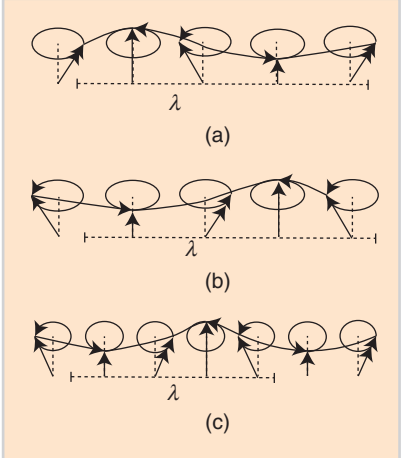

FIGURE 1 The SW parameters. (a) $\varphi = 0$, $k = 1$. (b) $\varphi = \pi$, $k = 1$. (c) $\varphi = \pi$, $k = 3$.

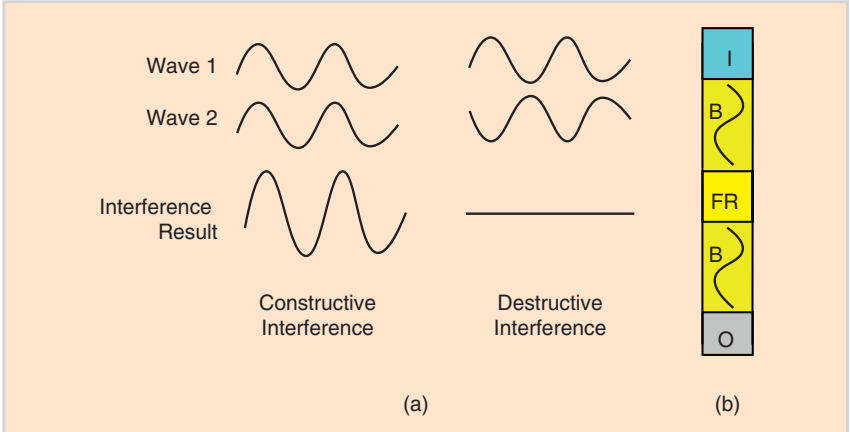

FIGURE 2 (a) Constructive and destructive interference. (b) An SW device.

identified by using components similar to the ones employed for SW excitation [6], [35] by means of phase and threshold detection techniques [6].

Conventional SW detection makes use of a strip antenna within which the SW generates an ac current, which is subsequently rectified with a diode. A more power-efficient way is ME cell SW detection, which follows the inverse conversion trajectory; i.e., the output SW generates a dynamic strain, which is sensed by the ME piezoelectric layers and subsequently converted into a voltage swing. Output value detection can be done by comparing the resulting SW phase with a predefined phase such that, e.g., if the phase difference $\Delta\phi = 0$, the output is logic 0, whereas the output is logic 1 if the phase difference is $\Delta\phi = \pi$. Alternatively, threshold detection can be utilized by comparing the output SW amplitude with a predefined threshold value T; i.e., if the SW amplitude is larger than T, the output is logic 1 and 0 otherwise [6], [35].

4:2 COMPRESSOR

For many state-of-the-art applications, e.g., artificial neural networks and machine vision, which heavily rely on multiplication, the availability of fast multipliers is essential. Tree multipliers are the fastest and can perform a multiplication within two processor clock cycles [36]. They embed three stages, i.e., partial product generation, a reduction tree, and a carry propagation adder. In an n-bit multiplier, the first stage requires n^2 gates to produce the partial products matrix, the second stage provides a logarithmic depth reduction of n n-bit partial products to two numbers (n:2 reduction) without carry propagation, and the final stage is a carry propagate adder that sums up the reduction tree outputs [36]. n:2 reduction is a carry-propagation-free addition and has been traditionally implemented by means of full and half adders, resulting in a reduction tree depth of O(log n). More recently, the theoretical concept of n:2 compressors was introduced, and its practical implementation, e.g., a 4:2 compressor, was reported in [33] and [37]–[39]. When built with 4:2 compressors, each element in the reduction tree processes four instead of three bits, which results in shallower reduction trees with a more regular layout [36]. To get some insight on this, let us assume 8-bit multiplication as a discussion vehicle. In this case, eight 8-bit partial products have to be reduced to two numbers. Figure 3 presents the processing of an 8-bit column in dot notation, with full and half adders in Figure 3(a) and 4:2 compressors in Figure 3(b). As one can observe in the figure, the full and half adder-based reduction requires two stages, while the 4:2 compressor reduction needs only one stage.

Essentially, a 4:2 compressor can be implemented by two cascaded full adders (FAs) processing 4 bits in the same column and generating one sum bit in the current column and a carry to the next column, as depicted in Figure 4. Since the 4:2 compressor output can assume only a value between zero and three, while the input value can be between zero and four, a transport for the 4:2 compressor in the adjacent column is required to conserve the input value. Thus, the 4:2 compressor applied in column i of the partial product matrix processes four dots in that column (I_1, I_2, I_3, and I_4) and a carry-in C_{in} provided by a 4:2 compressor in column $i-1$ [$C_{o1}(i-1)$], and it generates three outputs: one intermediate transport $C_{o1}(i)$ that serves as C_{in} for a counter in column $i+1$, the sum $S(i)$, and carry-out $C_{o2}(i)$. Note that as $C_{o1}(i-1)$ participates in the second stage of the calculation, there is no extra delay induced, and the reduction is still performed in a carry-propagation-free manner. We remark that if the 4:2 compressor is implemented by two cascaded FAs, as in Figure 4, the reduction schemes in Figure 3 have the same delay. Fortunately, most state-of-the-art CMOS 4:2 compressor implementations rely on different circuitry and are faster than two cascaded FAs [33], [37]–[39].

Given that multiplication-dominated, error-tolerant applications exist, e.g., multimedia processing and social media [32], approximate CMOS 4:2 compressors have been proposed [33] to enable

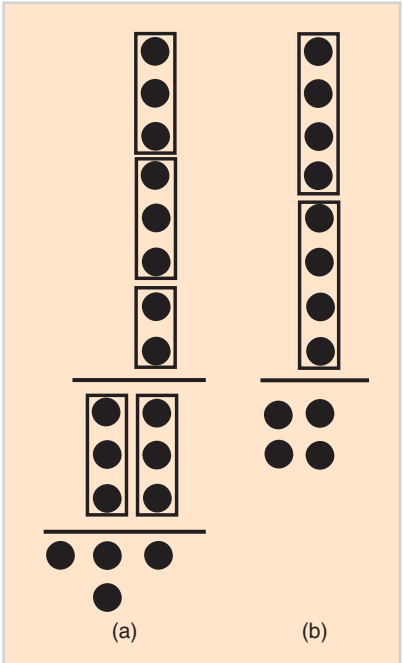

FIGURE 3 (a) Eight dots in an 8-bit-multiplier partial product stage processed by full and half adders. (b) Eight dots in an 8-bit-multiplier partial product stage processed by a 4:2 compressor.

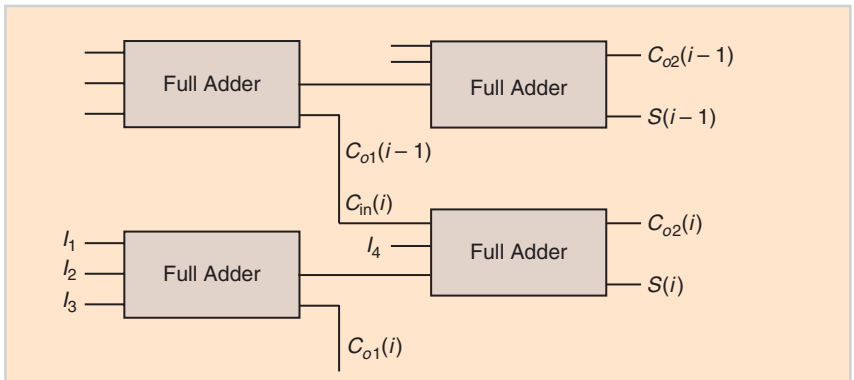

FIGURE 4 A 4:2 compressor.

significant energy consumption and area savings. A straightforward SW 4:2 compressor implementation can be built using the SW FA proposed in [17], which provides accurate results with acceptable delay and energy efficiency, as discussed in the "Performance Evaluation and Discussion" section. However, as mentioned, many applications are error tolerant and work properly within certain limits [32]; therefore, by enabling approximate computing, a more energy-efficient SW 4:2 compressor can be made.

SW APPROXIMATE 4:2 COMPRESSOR

The straightforward implementation of an SW approximate 4:2 compressor can be done by means of the two approximate SW FAs proposed in [40]. This requires the cascading of two FAs, which cannot be straightforwardly performed because different FA input combinations generate varying output SW strengths [7]. To solve this issue and make the compressor function correctly, a DC is required [7] to normalize the output of the first FA before passing it to the second. Figure 5 details the approximate compressor obtained by cascading two approximate FAs by means of a normalizer (DC). However, the DC induces substantial delay and area overheads, which makes working without it desirable. Therefore, we propose the novel DC-free approximate compressor presented in Figure 6. The behavior of the two directly cascaded FAs is now obtained with three- and five-input majority gates computing $C_{o1} = MAJ(X, Y, C_i)$ and $S = \bar{C}_{o2} = \overline{MAJ}(I_1, I_2, I_3, \bar{I}_4, \bar{C}_{in})$, respectively, where I_1, I_2, I_3, and I_4, and C_{in} are the excitation cells and C_{o1}, S, and C_{o2} are the detection cells.

Note that each input must be excited with a separate transducer, and each output must be sensed by a separate transducer. As mentioned in the "SW-Based Technology Fundamentals and Computing Paradigm" section, these transducers can be antennas and ME cells [6]. The proposed 4:2 approximate compressor generates C_{o1} without any error and S and C_{o2} with an average error rate of 31.25% and 18.75%, respectively.

Table 1 presents the truth table of the accurate 4:2 compressor C_{o1}, S_{ac}, and C_{o2ac}; the approximate 4:2 compressor without a DC C_{o1}, C_{o2ap1}, and S_{ap1}; and the approximate 4:2 compressor with a DC C_{o1}, C_{o2ap2}, and S_{ap2}. As can be

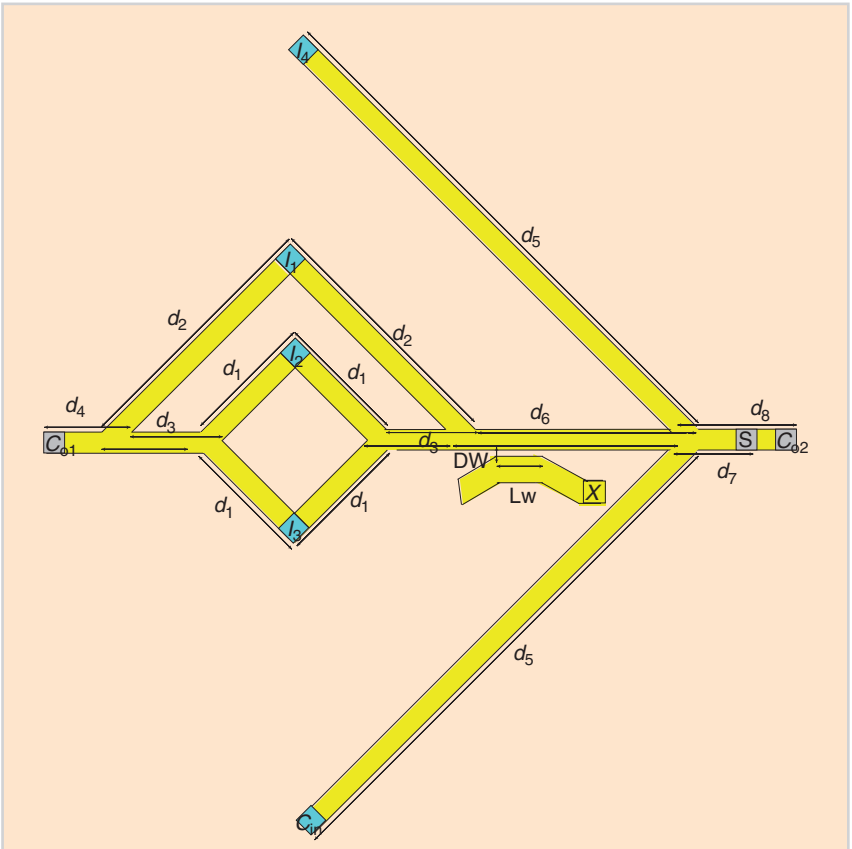

FIGURE 5 An approximate SW-based 4:2 compressor with a normalizer. DW: distance between the two waveguides; Lw: length of the coupled waveguide (based on the design specifications clarified in [7]).

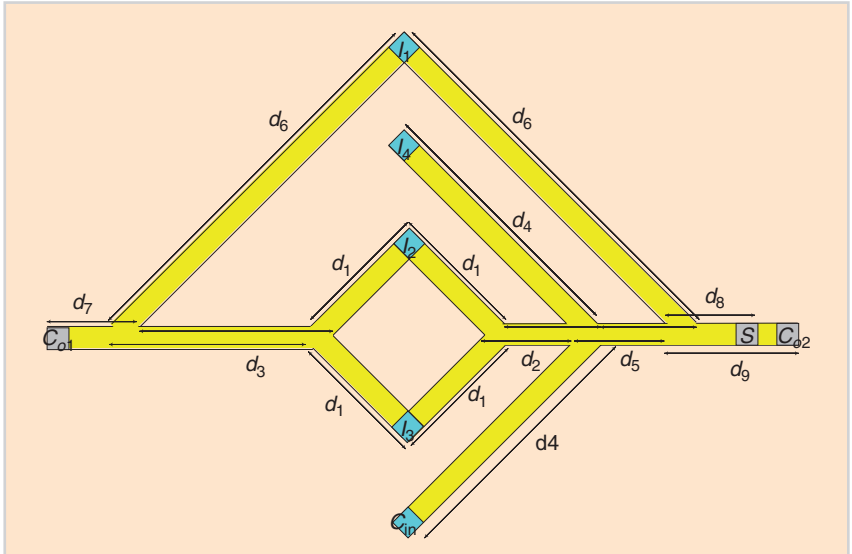

FIGURE 6 An approximate SW-based 4:2 compressor without a normalizer.

observed from the table, approximate 4:2 compressors with and without a DC provide the same average error rate of 25% because S_{ap1} and C_{o2ap1} have an error rate of 37.5%, and 12.5%, respectively, whereas S_{ap1} and C_{o2ap1} have an error rate of 31.25% and 18.75%, respectively. Note that the erroneous output values in the table are underlined and boldfaced to highlight them.

To achieve proper functionality for the structure in Figure 6, the waveguide width must be smaller than or equal to the SW wavelength to simplify the interference patterns. Furthermore, all SWs must be excited with the same amplitude, wavelength, and frequency, and the waveguide lengths must be accurately computed, as they determine the SW interaction modes. For example, if SW constructive (destructive) interference is envisaged for in-phase (out-of-phase) SWs, the distances must be equal to $n \times \lambda$, where $n = 0, 1, 2, \ldots$; this is the case for d_1, d_3, d_4, and d_6 in Figure 6. In contrast, if SW constructive (destructive) interference is envisaged for out-of-phase (in-phase) SWs, the distances must be equal to $(n + 1/2) \times \lambda$; this is the case for d_2 and d_5 in Figure 6. On the output side, it is important to detect the output at specific positions; i.e., if the desired output is the output itself, which is the case for C_{o1} in Figure 6, d_7 must be equal to $n \times \lambda$. On the other hand, if the inverted output is desired, the distance must be equal to $(n + 1)/2 \times \lambda$. Moreover, the outputs must be detected as near as possible to the last interference point to capture a large SW amplitude.

TABLE 1 The accurate and approximate SW-based 4:2 compressor.

$C_{IN}I_4I_3I_2I_1$	C_{O1}	C_{O2ac}	C_{O2ap1}	C_{O2ap2}	S_{ac}	S_{ap1}	S_{ap2}
00000	0	0	0	**1**	0	**1**	0
00001	0	0	0	0	1	1	1
00010	0	0	0	0	1	1	1
00011	1	0	0	0	0	**1**	**1**
00100	0	0	0	0	1	1	1
00101	1	0	0	0	0	**1**	**1**
00110	1	0	0	0	0	**1**	**1**
00111	1	0	0	0	1	1	1
01000	0	0	**1**	**1**	1	**0**	**0**
01001	0	1	1	1	0	0	0
01010	0	1	1	1	0	0	0
01011	1	0	0	0	1	1	1
01100	0	1	1	1	0	0	0
01101	1	0	0	0	1	1	1
01110	1	0	0	0	1	1	1
01111	1	1	**0**	**0**	0	**1**	**1**
10000	0	0	**1**	**1**	1	**0**	**0**
10001	0	1	1	1	0	0	0
10010	0	1	1	1	0	0	0
10011	1	0	0	0	1	1	1
10100	0	1	1	1	0	0	0
10101	1	0	0	0	1	1	1
10110	1	0	0	0	1	1	1
10111	1	1	**0**	**0**	0	**1**	**1**
11000	0	1	1	1	0	0	0
11001	0	1	1	1	1	**0**	**0**
11010	0	1	1	1	1	**0**	**0**
11011	1	1	1	1	0	0	0
11100	0	1	1	1	1	**0**	**0**
11101	1	1	1	1	0	0	0
11110	1	1	1	1	0	0	0
11111	1	1	1	**0**	1	**0**	1

The proposed SW 4:2 compressor operation principle is as follows:

- C_{o1}: SWs are excited at I_1, I_2, and I_3 with the same amplitude, wavelength, and frequency at the same time. The I_2 SW interferes constructively or destructively with the I_3 SW, depending on their phase difference. The resulting SW propagates through the waveguide and subsequently interferes with the I_1 SW. The resulting SW is captured at the output C_{o1}, based on phase detection.
- S and C_{o2}: The I_2 SW interferes constructively or destructively with the I_3 SW, depending on their phase difference, and the resulting SW propagates through the waveguide to interfere with the SWs excited at I_4 and C_{in}. The resulting SW propagates and subsequently interferes with the I_1 SW. Finally, the produced SW is captured at the outputs S and C_{o2}, based on threshold detection.

SIMULATION SETUP AND RESULTS

To validate the proposed structure by using MuMax3 [41], we made use of 1-nm thick and 50-nm wide $Fe_{60}Co_{20}B_{20}$ waveguides in addition to the following parameters [42]: saturation magnetization $M_s = 1.1$ MA/m, damping constant $\alpha = 0.0025$, and exchange stiffness $A_{exch} = 18.5$ pJ/m. Note that no external field is applied, as the shape anisotropy is strong enough to push the magnetization in the plane along the waveguide length. This configuration enables the propagation of backward volume SWs. The SWs are excited with a carrier frequency of 10 GHz modulated

by Gaussian pulses with a 500-ps sigma to save energy and guarantee the excitation of single-frequency SWs. From the backward volume SW dispersion relation presented in Figure 7, at 10 GHz, we determine k as being 26 rad/μm, which results in $\lambda = 2\pi/k = 240$ nm. As discussed in the "SW Approximate 4:2 Compressor" section, the distances d_1, d_3, d_4, d_6, and d_8 should be equal to integer multiples of λ, whereas d_2 and d_5 should be equal to integer multiples of $1/2\lambda$, and they are d_1 = 240 nm (n = 1), d_2 = 600 nm (n = 2.5), d_3 = 1,440 nm (n = 6), d_4 = 720 nm (n = 3), d_5 = 840 nm (n = 3.5), d_6 = 1,680 nm (n = 7), d_7 = 240 nm (n = 1), d_8 = 120 nm (n = 0), and d_9 = 240 nm (n = 1).

Figure 8 presents C_{o1} MuMax3 simulation results for $\{I_1,I_2,I_3\}$ = {0,0,0}, {0,0,0}, {0,0,1}, {0,1,0}, {0,1,1}, {1,0,0}, {1,0,1}, {1,1,0}, and {1,1,1}. One can observe in the figure that C_{o1} is correctly detected; $C_{o1} = 0$ for $\{I_1,I_2,I_3\}$ = {0,0,0}, {0,0,1}, {0,1,0}, and {0,1,1}, whereas $C_{o1} = 1$ for $\{I_1,I_2,I_3\}$ = {1,0,0}, {1,0,1}, {1,1,0}, and {1,1,1}, as it should, for a 0.4-ns reading window starting 1.8 ns after the input application.

Table 2 lists the normalized magnetization of the 4:2 approximate compressor outputs C_{o2} and S for all possible input combinations; i.e., $\{C_{in},I_4,I_3,I_2,I_1\}$ = {0,0,0,0,0}, {0,0,0,0,1}, ..., {1,1,1,1,0}, and {1,1,1,1,1}. The threshold technique is used to detect C_{o2} and S; i.e., if the normalized magnetization of the output SW is larger than the threshold value T, the output is logic 1 and zero otherwise. The output magnetization in the table is normalized with respect to the highest achieved magnetization, which is obtained when $\{C_{in},I_4,I_3,I_2,I_1\}$ = {0,0,1,1,1}. For C_{o2}, detection T = 0 is appropriate, which results in $C_{o2} = 1$ for input combinations $\{C_{in},I_4,I_3,I_2,I_1\}$ = {0,0,0,0,0}, {0,1,0,0,0}, {0,1,0,0,1}, {0,1,0,1,0}, {0,1,1,0,0}, {1,0,0,0,0}, {1,0,0,0,1}, {1,0,0,1,0}, {1,0,1,0,0}, {1,1,0,0,0}, {1,1,0,0,1}, {1,1,0,1,0}, {1,1,0,1,1}, {1,1,1,0,0}, {1,1,1,0,1}, and {1,1,1,1,0}, and $C_{o2} = 0$ for the remaining cases, as it should.

The same threshold value is suitable for S, but the threshold condition is flipped; i.e., if the resulting SW normalized magnetization is larger than zero, S is logic 0 and logic 1 otherwise. This results in S = 0 for $\{C_{in},I_4,I_3,I_2,I_1\}$ = {0,0,0,0,0}, {0,1,0,0,0}, {0,1,0,0,1}, {0,1,0,1,0}, {0,1,1,0,0}, {1,0,0,0,0}, {1,0,0,0,1}, {1,0,0,1,0}, {1,0,1,0,0}, {1,1,0,0,0}, {1,1,0,0,1}, {1,1,0,1,0}, {1,1,0,1,1}, {1,1,1,0,0}, {1,1,1,0,1}, and {1,1,1,1,0}, and S = 1 for the remaining cases, as it should. Therefore, the MuMax3 simulations prove that the proposed 4:2 approximate compressor provides the expected functionality.

PERFORMANCE EVALUATION

We evaluate the proposed SW approximate 4:2 compressor and compare it in terms of its error rate, energy consumption, delay, and area (the number of utilized devices) with its state-of-the-art SW 45-nm CMOS [43] and spin-CMOS [44] counterparts. To assess the performance of our proposal, we make the following assumptions: 1) ME cells are utilized for SW excitation and detection and have a power consumption of 34 nW and a delay of 0.42 ns [45], and 2) SWs consume negligible energy during interference and propagation through waveguides. Note that these assumptions might need to be revisited to better capture SW technology in future developments.

The proposed compressor with and without DC delays can be calculated by adding the SW propagation determined

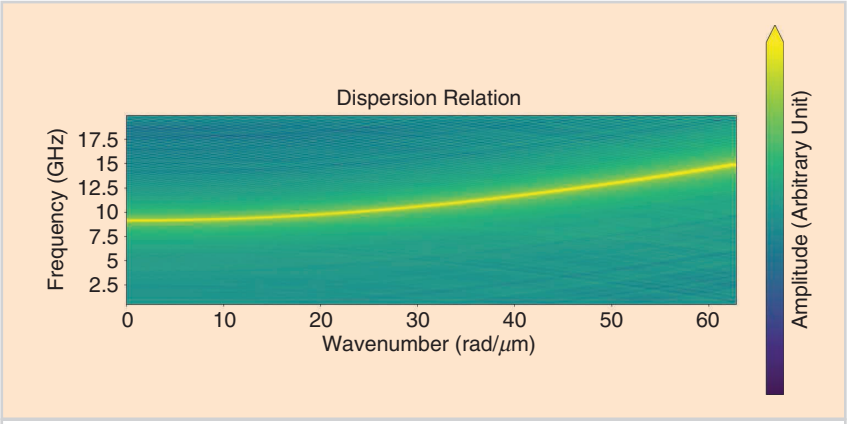

FIGURE 7 The SW dispersion relation.

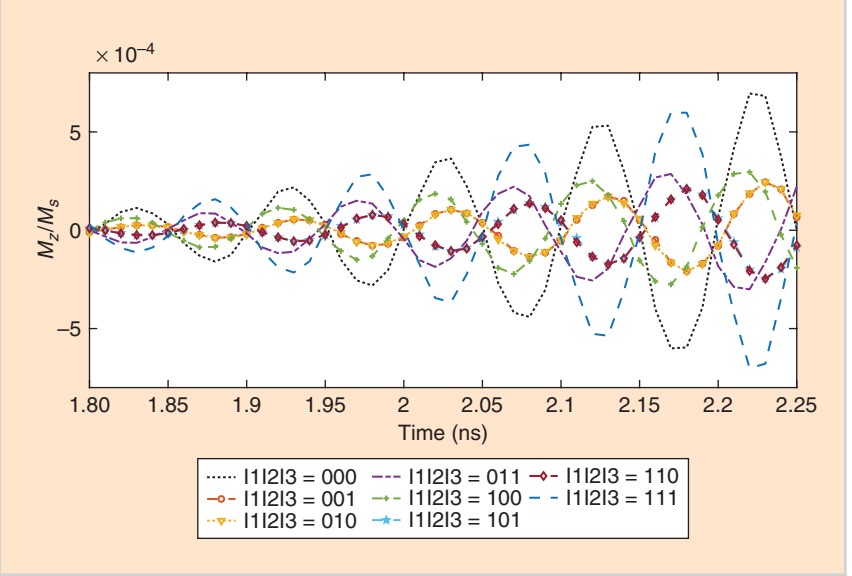

FIGURE 8 The normalized 4:2 compressor output C_{o1}.

TABLE 2 The normalized approximate SW-based 4:2 compressor outputs C_{O2} and S.

$C_{IN}I_4I_3I_2I_1$	RESULTING SW	C_{O2} AFTER THRESHOLDING	S AFTER THRESHOLDING
00000	0.45	1	0
00001	−0.08	0	1
00010	−0.07	0	1
00011	−0.59	0	1
00100	−0.01	0	1
00101	−0.46	0	1
00110	−0.49	0	1
00111	−1	0	1
01000	0.66	1	0
01001	0.23	1	0
01010	0.22	1	0
01011	−0.3	0	1
01100	0.3	1	0
01101	−0.21	0	1
01110	−0.2	0	1
01111	−0.69	0	1
10000	0.68	1	0
10001	0.18	1	0
10010	0.21	1	0
10011	−0.28	0	1
10100	0.28	1	0
10101	−0.22	0	1
10110	−0.18	0	1
10111	−0.73	0	1
11000	1	1	0
11001	0.51	1	0
11010	0.47	1	0
11011	0.012	1	0
11100	0.59	1	0
11101	0.07	1	0
11110	0.09	1	0
11111	−0.4	0	1

TABLE 3 The approximate 4:2 compressor performance comparison.

TECHNOLOGY	TYPE	ERROR RATE	ENERGY (fJ)	DELAY (ns)	DEVICE NUMBER
SW	Accurate	0	0.2	6.56	14
SW (with DC)	Approximate	0.31	0.137	11.4	8
SW (without DC)	Approximate	0.31	0.137	3.4	8
CMOS1 [43]	Approximate	0.125	0.172	0.049	40
CMOS2 [43]	Approximate	0.375	0.16	0.048	28
Spin-CMOS1 [44]	Approximate	0.31	173	3	28
Spin-CMOS2 [44]	Approximate	0.25	338	4	42

by means of micromagnetic simulations and the delay of the excitation and detection cells, which sums to 11.4 and 3.4 ns, respectively. We remark that, to perform amplitude normalization, the DC has to be rather long [7], which results in a large delay overhead. Note that two approximate CMOS 4:2 compressor designs were reported in [43]; the first (CMOS1) consists of an approximate FA and two multiplexers (nine AND, two XOR, and six OR two-input gates); the second (CMOS2) consists of two approximate FAs (six two-input XOR gates). In addition, two approximate spin-CMOS 4:2 compressor designs were suggested in [44]; the first (Spin-CMOS1) consists of two approximate FAs (two three-input majority gates); the second (Spin-CMOS2) consists of an accurate and approximate FA (two three-input and one five-input majority gates).

Table 3 provides the evaluation results. When compared with the accurate SW compressor, which is a direct implementation consisting of two accurate SW adders in [17], the proposed 4:2 compressor saves 31.5% more energy and is 1.93× faster. Moreover, it has the same energy consumption and error rate as the approximate compressor with a DC, but it has 3× less delay. In addition, it consumes 20% and 14% less energy, has higher delay by approximately two orders of magnitude, and exhibits 61% more and 17% less average error when compared with CMOS1 and CMOS2 in Table 3, respectively. When compared with same-error-rate spin-CMOS (Spin-CMOS1 design in Table 3), it consumes three orders of magnitude less energy and provides a 17% delay reduction. Although Spin-CMOS2 provides a 19% better average error rate, it is three orders of magnitude less effective in terms of energy consumption and speed. Note that the proposed compressor requires the smallest number of devices, which indicates that it potentially demands the least chip real estate.

To get some insight into the implications of our proposal at the application level, we consider the well-known JPEG encoding, which relies on the discrete cosine transform (DCT) [46], as a discussion vehicle. Given that JPEG

encoding is error tolerant and that DCT is a multiplication-dominated algorithm, 4:2 approximate compressors based on tree multipliers are quite attractive for practical JPEG code implementations. Such an approach has been presented in [44], and briefly, it is as follows: DCT and the inverse DCT are implemented by means of accurate adders and approximate compressor-based multipliers; thus, additions provide accurate results, and multiplications yield approximative results. The 16 × 16 approximate signed multiplier implementations are based on the Baugh–Wooley algorithm and Dadda [36] partial product reduction implemented with 4:2 counters. Given that the approximate 4:2 compressor in [44] has the same average error rate as the one we propose, we can infer that replacing that compressor with ours does not change the image quality, while resulting in three orders of magnitude less energy consumption.

We note that the main goal of this article is to propose and validate an SW 4:2 approximate compressor, and hence we do not take into consideration thermal and variability effects. However, in [47], it was suggested that thermal noise, edge roughness, and the waveguide trapezoidal cross section do not have a noticeable impact on gate functionality. Thus, we expect that the 4:2 approximate compressor functions correctly in their presence. However, further investigation of such phenomena is of great interest, but it cannot be performed before technology data and suitable simulation tools become available.

DISCUSSION

Our evaluation proved that SW technology can improve the state of the art in terms of energy and area [6]. However, several challenges have to be solved, such as [6] the following:

- *Immature technology*: As mentioned in the "SW-Based Technology Fundamentals and Computing Paradigm" section, SW excitation and detection can be done by means of antennas and ME cells. Note that inductive antenna excitation is neither energy efficient nor highly scalable [6], while ME cells are energy efficient and highly scalable but not experimentally demonstrated. Thus, ME cells are expected to be utilized for SW excitation and detection. In addition, ME phase detection transducers are general purpose, whereas ME threshold detection transducers require specific threshold value tuning by means of dimensions and internal stack structures.
- *Scalability*: SW devices are highly scalable because, conceptually speaking, the essential requirement for achieving functionality is that the spin wave device should be larger than the SW wavelength, which can be in the nanometer range. In addition, the figures reported in [45] indicate that SW circuit footprints favorably compare with CMOS equivalents; e.g., the area of a hybrid SW–CMOS 32-bit divider is 3.5× smaller than that of its 10-nm CMOS counterpart. Although benchmarking might build on assumptions that are not fully realistic, as SW technology is still immature, it gives some insight into the SW circuit area and performance potential.

Although benchmarking might build on assumptions that are not fully realistic, as SW technology is still immature, it gives some insight into the SW circuit area and performance potential.

CONCLUSION

This article proposed an SW-based 4:2 approximate compressor that consists of three- and five-input majority gates. We reported the design of approximate circuits without DCs, which are essential to normalize gate outputs when cascading them in accurate circuit designs. We validated the proposed compressor by means of micromagnetic simulations and compared the design to the state-of-the-art SW 22-nm CMOS, 45-nm CMOS, and spin-CMOS counterparts.

The evaluation results indicated that the proposed 4:2 compressor saves 31.5% more energy in comparison with the accurate SW compressor and that it has the same energy consumption and error rate as the approximate compressor with a DC but with 3× less delay. Moreover, it consumes 14% less energy while having a 17% lower error rate when compared with the approximate 45-nm CMOS counterpart. Furthermore, it outperformes the approximate spin-CMOS-based compressor by three orders of magnitude in terms of energy consumption while providing the same error rate. Last but not least, the proposed compressor requires the smallest number of devices; thus, it potentially needs the least chip real estate.

ACKNOWLEDGMENT

This work received funding from the European Union's Horizon 2020 research and innovation program within the Future and Emerging Technologies Open project Spin Wave Computing for Ultimately-Scaled Hybrid Low-Power Electronics, under grant 801055. It has also been partially supported by IMEC's industrial affiliate program on beyond-CMOS logic. Frederic Vanderveken acknowledges financial support from Flanders Research Foundation through grant 1S05719N.

ABOUT THE AUTHORS

Abdulqader Nael Mahmoud (a.n.n.mahmoud@tudelft.nl) is with the Computer Engineering Laboratory, Delft University of Technology, Delft, 2628 CD, The Netherlands.

Frederic Vanderveken (frederic.vanderveken@imec.be) is with IMEC, Leuven, 3001, Belgium.

Florin Ciubotaru (florin.ciubotaru@imec.be) is with IMEC, Leuven, 3001, Belgium.

Christoph Adelmann (christoph.adelmann@imec.be) is with IMEC, Leuven, 3001, Belgium.

Said Hamdioui (s.hamdioui@tudelft.nl) is with the Computer Engineering Laboratory, Delft University of Technology, Delft, 2628 CD, The Netherlands.

Sorin Cotofana (s.d.cotofana@tudelft.nl) is with the Computer Engineering Laboratory, Delft University of Technology, Delft, 2628 CD, The Netherlands.

REFERENCES

[1] N. D. Shah, E. W. Steyerberg, and D. M. Kent, "Big data and predictive analytics: Recalibrating expectations," *JAMA*, vol. 320, no. 1, pp. 27–28, 2018, doi: 10.1001/jama.2018.5602.

[2] N. Z. Haron and S. Hamdioui, "Why is CMOS scaling coming to an end?" in *Proc. IEEE Int. Des. Test Workshop*, 2008, pp. 98–103, doi: 10.1109/IDT.2008.4802475.

[3] Y. Jiang, N. C. Laurenciu, H. Wang, and S. D. Cotofana, "Graphene nanoribbon based complementary logic gates and circuits," *IEEE Trans. Nanotechnol.*, vol. 18, pp. 287–298, Mar. 2019, doi: 10.1109/TNANO.2019.2903480.

[4] H. A. D. Nguyen, J. Yu, M. A. Lebdeh, M. Taouil, S. Hamdioui, and F. Catthoor, "A classification of memory-centric computing," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 16, no. 2, pp. 1–26, Apr. 2020, doi: 10.1145/3365837.

[5] S. Agarwal *et al.*, "International roadmap of devices and systems 2017 edition: Beyond CMOS chapter," Sandia National Laboratories (SNL-NM), Livermore, CA, USA, Tech. Rep. SAND2018-3550R, 2018. [Online]. Available: https://irds.ieee.org/images/files/pdf/2017/2017IRDS_BC.pdf

[6] A. Mahmoud *et al.*, "Introduction to spin wave computing," *J. Appl. Phys.*, vol. 128, no. 16, p. 161,101, 2020, doi: 10.1063/5.0019328.

[7] A. N. Mahmoud, F. Vanderveken, C. Adelmann, F. Ciubotaru, S. Cotofana, and S. Hamdioui, "Spin wave normalization toward all magnonic circuits," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 1, pp. 536–549, Jan. 2021, doi: 10.1109/TCSI.2020.3028050.

[8] A. Mahmoud, F. Vanderveken, F. Ciubotaru, C. Adelmann, S. Cotofana, and S. Hamdioui, "N-bit data parallel spin wave logic gate," in *Proc. 2020 Design, Automat. Test Eur. Conf. Exhib. (DATE)*, pp. 642–645, doi: 10.23919/DATE48585.2020.9116368.

[9] A. Mahmoud, F. Vanderveken, C. Adelmann, F. Ciubotaru, S. Cotofana, and S. Hamdioui, "2-output spin wave programmable logic gate," in *Proc. 2020 IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, pp. 60–65, doi: 10.1109/ISVLSI49217.2020.00021.

[10] M. P. Kostylev, A. A. Serga, T. Schneider, B. Leven, and B. Hillebrands, "Spin-wave logical gates," *Appl. Phys. Lett.*, vol. 87, no. 15, p. 153,501, 2005, doi: 10.1063/1.2089147.

[11] A. Mahmoud, F. Vanderveken, C. Adelmann, F. Ciubotaru, S. Hamdioui, and S. Cotofana, "Fan-out enabled spin wave majority gate," *AIP Adv.*, vol. 10, no. 3, p. 035119, 2020, doi: 10.1063/1.5134690.

[12] A. Mahmoud, F. Vanderveken, C. Adelmann, F. Ciubotaru, S. Hamdioui, and S. Cotofana, "4-output programmable spin wave logic gate," in *Proc. 2020 IEEE 38th Int. Conf. Comput. Design (ICCD)*, pp. 332–335, doi: 10.1109/ICCD50377.2020.00062.

[13] A. Mahmoud, C. Adelmann, F. Vanderveken, S. Cotofana, F. Ciubotaru, and S. Hamdioui, "Fan-out of 2 triangle shape spin wave logic gates," in *Proc. 2021 Design, Automat. Test Eur. Conf. Exhib. (DATE)*, 2021, pp. 948–953, doi: 10.23919/DATE51398.2021.9474089.

[14] A. N. Mahmoud, F. Vanderveken, C. Adelmann, F. Ciubotaru, S. Hamdioui, and S. Cotofana, "Multi-frequency data parallel spin wave logic gates," *IEEE Trans. Magn.*, vol. 57, no. 5, pp. 1–12, 2021, doi: 10.1109/TMAG.2021.3062022.

[15] A. Mahmoud, F. Vanderveken, C. Adelmann, F. Ciubotaru, S. Hamdioui, and S. Cotofana, "Achieving wave pipelining in spin wave technology," in *Proc. 2021 22nd Int. Symp. Qual. Electron. Design (ISQED)*, pp. 54–59, doi: 10.1109/ISQED51717.2021.9424264.

[16] A. Khitun and K. L. Wang, "Non-volatile magnonic logic circuits engineering," *J. Appl. Phys.*, vol. 110, no. 3, p. 034306, 2011, doi: 10.1063/1.3609062.

[17] A. Mahmoud, F. Vanderveken, F. Ciubotaru, C. Adelmann, S. Cotofana, and S. Hamdioui, "Spin wave based full adder," in *Proc. 2021 IEEE Int. Symp. Circuits Syst. (ISCAS)*, pp. 1–5, doi: 10.1109/ISCAS51556.2021.9401524.

[18] F. Gertz, A. Kozhevnikov, Y. Filimonov, and A. Khitun, "Magnonic holographic memory," *IEEE Trans. Magn.*, vol. 51, no. 4, pp. 1–5, Apr. 2015, doi: 10.1109/TMAG.2014.2362723.

[19] R. Miller, V. K. Prasanna-Kumar, D. I. Reisis, and Q. F. Stout, "Parallel computations on reconfigurable meshes," *IEEE Trans. Comput.*, vol. 42, no. 6, pp. 678–692, 1993, doi: 10.1109/12.277290.

[20] L. O. Chua and L. Yang, "Cellular neural networks: Theory," *IEEE Trans. Circuits Syst.*, vol. 35, no. 10, pp. 1257–1272, Oct. 1988, doi: 10.1109/31.7600.

[21] L. O. Chua and T. Roska, "The CNN paradigm," *IEEE Trans. Circuits Syst. I. Fundam. Theory Appl.*, vol. 40, no. 3, pp. 147–156, 1993, doi: 10.1109/81.222795.

[22] N. Petkov, *Systolic Parallel Processing*. New York, NY, USA: Elsevier, 1992.

[23] M. Rahman, S. Khasanvis, J. Shi, and C. A. Moritz, "Wave interference functions for neuromorphic computing," *IEEE Trans. Nanotechnol.*, vol. 14, no. 4, pp. 742–750, 2015, doi: 10.1109/TNANO.2015.2438231.

[24] C. Pan and A. Naeemi, "An expanded benchmarking of beyond-CMOS devices based on boolean and neuromorphic representative circuits," *IEEE J. Explor. Solid-State Computat.*, vol. 3, pp. 101–110, Dec. 2017, doi: 10.1109/JXCDC.2018.2793536.

[25] N. Locatelli, V. Cros, and J. Grollier, "Spin-torque building blocks," *Nature Mater.*, vol. 13, no. 1, pp. 11–20, 2014, doi: 10.1038/nmat3823.

[26] A. Khitun, M. Bao, and K. L. Wang, "Magnetic cellular nonlinear network with spin wave bus for image processing," *Superlattices Microstruct.*, vol. 47, no. 3, pp. 464–483, 2010, doi: 10.1016/j.spmi.2009.11.004.

[27] F. Macià, A. D. Kent, and F. C. Hoppensteadt, "Spin-wave interference patterns created by spin-torque nano-oscillators for memory and computation," *Nanotechnology*, vol. 22, no. 9, p. 095301, 2011, doi: 10.1088/0957-4484/22/9/095301.

[28] A. Kozhevnikov, F. Gertz, G. Dudko, Y. Filimonov, and A. Khitun, "Pattern recognition with magnonic holographic memory device," *Appl. Phys. Lett.*, vol. 106, no. 14, p. 142,409, 2015, doi: 10.1063/1.4917507.

[29] H. Arai and H. Imamura, "Neural-network computation using spin-wave-coupled spin-torque oscillators," *Phys. Rev. Appl.*, vol. 10, no. 2, p. 024040, 2018, doi: 10.1103/PhysRevApplied.10.024040.

[30] M. Huth, F. Porrati, and O. V. Dobrovolskiy, "Focused electron beam induced deposition meets materials science," *Microelectron. Eng.*, vols. 185–186, pp. 9–28, Jan. 2018, doi: 10.1016/j.mee.2017.10.012.

[31] P. Fischer, D. Sanz-Hernández, R. Streubel, and A. Fernández-Pacheco, "Launching a new dimension with 3D magnetic nanostructures," *APL Mater.*, vol. 8, no. 1, p. 010701, 2020, doi: 10.1063/1.5134474.

[32] S. Mittal, "A survey of techniques for approximate computing," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 1–33, Mar. 2016, doi: 10.1145/2893356.

[33] M. Kumar and J. Nath, "Design of an energy efficient 4-2 compressor," *IOP Conf. Series, Mater. Sci. Eng.*, vol. 225, p. 012136, Aug. 2017, doi: 10.1088/1757-899X/225/1/012136.

[34] J. R. Eshbach and R. W. Damon, "Surface magnetostatic modes and surface spin waves," *Phys. Rev. J. Arch.*, vol. 118, no. 5, pp. 1208–1210, Jun. 1960. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.118.1208, doi: 10.1103/PhysRev.118.1208.

[35] A. V. Chumak, A. A. Serga, and B. Hillebrands, "Magnonic crystals for data processing," *J. Phys. D, Appl. Phys.*, vol. 50, no. 24, p. 244,001, 2017. [Online]. Available: http://stacks.iop.org/0022-3727/50/i=24/a=244001, doi: 10.1088/1361-6463/aa6a65.

[36] B. Parhami, *Computer Arithmetic: Algorithms and Hardware Designs*, 2nd ed. London, U.K.: Oxford Univ. Press, 2009.

[37] A. Momeni, J. Han, P. Montuschi, and F. Lombardi, "Design and analysis of approximate compressors for multiplication," *IEEE Trans. Comput.*, vol. 64, no. 4, pp. 984–994, 2015, doi: 10.1109/TC.2014.2308214.

[38] J. Mori *et al.*, "A 10 ns 54*54 b parallel structured full array multiplier with 0.5 mu m cmos technology," *IEEE J. Solid-State Circuits*, vol. 26, no. 4, pp. 600–606, 1991, doi: 10.1109/4.75061.

[39] F. Ranjbar, Y. Forghani, and D. Bahrepour, "High performance 8-bit approximate multiplier using novel 4: 2 approximate compressors for fast image processing," *Int. J. Integr. Eng.*, vol. 10, no. 1, pp. 114–133, 2018, doi: 10.30880/ijie.2018.10.01.018.

[40] A. Mahmoud, F. Vanderveken, F. Ciubotaru, C. Adelmann, S. Hamdioui, and S. Cotofana, "Spin wave based approximate computing," 2021, arXiv:2103.12869.

[41] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. Van Waeyenberge, "The design and verification of mumax3," *AIP Adv.*, vol. 4, no. 10, p. 107,133, 2014, doi: 10.1063/1.4899186.

[42] T. Devolder et al., "Time-resolved spin-torque switching in MgO-based perpendicularly magnetized tunnel junctions," *Phys. Rev. B*, vol. 93, no. 2, p. 024420, Jan. 2016, doi: 10.1103/PhysRevB.93.024420.

[43] K. Manikantta Reddy, M. Vasantha, Y. Nithin Kumar, and D. Dwivedi, "Design and analysis of multiplier using approximate 4-2 compressor," *AEU – Int. J. Electron. Commun.*, vol. 107, pp. 89–97, Jul. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1434841118330085, doi: 10.1016/j.aeue.2019.05.021.

[44] S. Angizi, H. Jiang, R. F. DeMara, J. Han, and D. Fan, "Majority-based spin-CMOS primitives for approximate computing," *IEEE Trans. Nanotechnol.*, vol. 17, no. 4, pp. 795–806, 2018, doi: 10.1109/TNANO.2018.2836918.

[45] O. Zografos *et al.*, "Design and benchmarking of hybrid CMOS-spin wave device circuits compared to 10nm CMOS," in *Proc. IEEE 15th Int. Conf. Nanotechnol. (IEEE-NANO)*, Jul. 2015, pp. 686–689, doi: 10.1109/NANO.2015.7388699.

[46] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Trans. Consum. Electron.*, vol. 38, no. 1, pp. xviii–xxxiv, 1992, doi: 10.1109/30.125072.

[47] Q. Wang, P. Pirro, R. Verba, A. Slavin, B. Hillebrands, and A. V. Chumak, "Reconfigurable nanoscale spin-wave directional coupler," *Sci. Adv.*, vol. 4, no. 1, p. e1701517, 2018, doi: 10.1126/sciadv.1701517.