Energy-Performance Assessment of Oscillatory Neural Networks based on VO2 Devices for Future Edge AI Computing

Corentin Delacour, Stefania Carapezzi, Madeleine Abernot, Aida Todri-Sanial

To cite this version:

HAL Id: lirmm-03591176
https://hal-lirmm.ccsd.cnrs.fr/lirmm-03591176
Preprint submitted on 28 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Energy-Performance Assessment of Oscillatory Neural Networks based on VO\textsubscript{2} Devices for Future Edge AI Computing

Corentin Delacour, Stefania Carapezzi, Madeleine Abernot, Aida Todri-Sanial

Abstract—Oscillatory Neural Network (ONN) is an emerging neuromorphic architecture composed of oscillators that implement neurons and coupled by synapses. ONNs exhibit rich dynamics and associative properties, which can be used to solve problems in the analog domain according to the paradigm let physics compute. For example, compact oscillators made of VO\textsubscript{2} material are good candidates for building low-power ONN architectures dedicated to AI applications at the edge like pattern recognition. However, little is known about the ONN scalability and its performances when implemented in hardware. Before deploying ONN, it is necessary to assess its computation time, energy consumption, performance and accuracy for a given application. Here, we consider a VO\textsubscript{2}-oscillator as an ONN building block and we perform circuit-level simulations to evaluate the ONN performances at the architecture level. Notably, we investigate how ONN computation time, energy and memory capacity scale with the number of oscillators. We show that ONN energy grows linearly when scaling up the network, making it suitable for large-scale integration at the edge. Furthermore, we investigate the design knobs for minimizing the ONN energy. Assisted by TCAD simulations, we report on scaling of VO\textsubscript{2} devices in crossbar geometry to decrease the oscillator voltage and energy. We benchmark ONN versus state-of-the-art architectures and observe that the ONN paradigm is a competitive energy efficient solution for scaled VO\textsubscript{2} devices. Finally, we present how ONN can efficiently detect edges in images captured on low-power edge devices.

Index Terms—Oscillatory Neural Network, Vanadium Dioxide (VO\textsubscript{2}), Edge AI, Hopfield Neural Network, Image Edge Detection

I. Introduction

The number of mobile devices connected to the internet has considerably increased the past few years and is estimated to reach 75 billion by 2025 [1]. The Internet of Things (IoT) paradigm is driven by continuous machine learning and AI progress, allowing mobile devices to predict and decide in interaction with their environment. IoT devices are connected and regularly exchange data on the internet, but depending on the workload, such connectivity may suffer from latency issues, bandwidth problems, and even confidentiality issues for some applications such as sending sensitive data as in healthcare devices. For these reasons, IoT devices require some local processing capability instead of transferring data over cloud or data centers [2]. However, with the sophistication of AI algorithms, computation at the edge becomes challenging for devices with limited resources [1]. Current algorithms depend on large neural networks with thousands of synapses and data propagate through several layers of neurons, via successive matrix multiplications between data and synaptic weights. Such algorithms implemented on a Von Neumann architecture (such as edge CPUs) suffer from large power consumption and data transfer bottleneck between memory and processing unit, also known as the Von Neumann bottleneck [3].

To overcome the limitations of the Von Neumann bottleneck, alternative brain-inspired computing paradigms are explored. Inspired by the biological neural networks, in-memory computing aims to merge memory and processing functions, where device physical properties can store the network’s weights while efficiently performing matrix products [4]. For instance, Ohm’s and Kirchhoff’s laws naturally describe multiplication and summation in the analog domain, allowing fast and efficient computations. Such as in crossbar architectures that have been shown to perform energy-efficient inference [4].

Based on the in-memory computing paradigm and inspired by the olfactory system in biological neural networks, Oscillatory Neural Networks (ONNs) compute by harnessing the rich dynamics of coupled oscillators for parallel processing. In ONNs, neurons are oscillators that are physically connected by electrical components (synapses). By exploiting nonlinear oscillator, dynamics allows to compute in phase [5] or frequency [6]. In ONNs, the memory is locally stored in synaptic elements, and it is distributed among oscillators that act as processing units interacting in parallel, in contrast with Von Neumann’s architecture. Mathematicians have studied ONN for decades and have proved the collective computational capability in ONNs [5].

In hardware, ONNs have been implemented with various technologies such as CMOS ASICs [7], [8], field-programmable gate arrays [9], spintronic oscillators [10], micro-electromechanical systems [11] for solving tasks varying from image processing [12], [13], [14], [15] to combinatorial optimization problems [16], [17], [18], [19], [20] and to implement reservoir computers [10], [21], [22], [23]. Insulator-to-metal phase transition (IMT) devices such as vanadium dioxide (VO\textsubscript{2}) are promising candidates to design compact nano-oscillators as they only require an additional load to produce oscillations at room temperature and are CMOS-compatible [14], [24]. It is believed that scaled VO\textsubscript{2} devices would provide fast and energy-efficient oscillations and has...
The contributions of this work are as follows. We resistances to study ONN architecture and performance. We model coupling elements not considered. We represent phases by black and white pixels where a pixel corresponds to a single oscillator.

Prior experimental work using VO₂ oscillators have reported on ONN performances for less than ten oscillators, but information on 1) VO₂ device scaling and 2) ONN architecture scaling are yet to be explored. For example, for image processing application, Shukla et al. reported the power consumption for six-coupled VO₂-oscillators [25], [24] but do not mention the energy and delay for larger networks. Though, at the device level, a power projection motivates the scaling down of the VO₂ channel length in planar geometry. For spoken vowel detection, Dutta et al. [6] propose to use four coupled planar VO₂ oscillators that consume 6 µW each, but scalability and computation time are not discussed. Corti et al. [14] describe how four and nine coupled VO₂ oscillators can be used as input filters in convolutional neural networks and make a projection of the ONN energy-delay for scaled VO₂ devices in crossbar geometry. However, the estimation remains empirical as VO₂ device physics and coupling elements parameters are not considered.

In this work, we investigate VO₂-ONN scaling at device, circuit, and architecture levels. We model coupling elements by resistances to study ONN architecture and performance. The contributions of this work are as:

- we show that the memory capacity of a fully-coupled ONN scales linearly with the number of oscillators, similarly to Hopfield Neural Networks (HNN).
- we analytically derive and express the trade-off between the ONN size, the oscillating frequency and the Signal to Noise Ratio (SNR).
- we determine the ONN linear energy scaling and constant computation time with the number of oscillators.
- assisted by Technology Computer-Aided Design (TCAD) simulations, we demonstrate how to minimize the oscillating energy for crossbar VO₂ devices.
- we benchmark the VO₂-based ONN energy and delay with respect to state-of-the-art neural accelerators and neuromorphic chips. We highlight that ONN can be a competitive computing paradigm for high oscillating frequencies.
- finally, we showcase a VO₂-based ONN benchmark for image edge detection and compare it with the state-of-the-art CMOS ASICs.

II. ONN Description

A. ONN as an Associative Memory

In ONNs, the information is encoded in phase differences among oscillators and a reference oscillator (the first one) [5]. [13]. ONN inputs are the phase initialization \( \Delta \Phi^\text{in} \in [0°, 180°]\), and outputs are the phase differences measured once ONN stabilizes. ONN output phases lock to binary values \( \Phi^\text{out} = 0° \) and \( \Phi^\text{out} = 180° \), which in the case of image processing can be represented by white and black pixels, respectively. Hence, we represent the ONN phase state by black and white images, where every oscillator corresponds to a single pixel (Fig.1).

In this work, we investigate ONN for associative memory applications like a Hopfield Neural Network (HNN) [26]. HNNs have been used for various applications such as solving optimization problems [27], [28] and image processing and encryption [29]. To study the ONN memory capacity, we perform pattern recognition like HNN where the network is fully connected, meaning every oscillator is connected to all the others. We train the network using the Hebbian learning rule [26] and we map the synaptic weights to coupling resistances [30] to store training images in the ONN. When ONN settles to a stable phase state (corresponding to a training image), its dynamics can be interpreted as the minimization of an energy function defined in [26]. An example of ONN computation with 60 VO₂-oscillators is shown in Fig.2 where ONN retrieves the noiseless digit ‘1’ after a few oscillation cycles. Next, we present how the dynamics of coupled VO₂-oscillators lead to associative properties, similarly to HNN.

B. VO₂ Oscillator

VO₂ is an IMT material that can switch in two different resistive states depending on its voltage \( V \) [13]. It transitions from an insulating state to a metallic state when \( V \) is above a threshold \( V_{tr} \), and reciprocally when \( V \) reaches the threshold \( V_{t} \). Hence, VO₂ presents a hysteresis in its \( I-V \) plan (Fig.3).
We use this property to design a relaxation oscillator. We bias the VO₂ device with a load resistance \( R_S \) in series and we connect a capacitor \( C_P \) in parallel with the output node \( V_{out} \) to adjust the oscillation frequency (Fig. 3a). The oscillator’s dynamics are described as \( C_P \frac{dV}{dt} = i - i_L \), where \( i \) is the VO₂ current capturing its hysteresis behavior. To produce oscillations, the line \( i_L \) must intercept \( i \) in the VO₂ negative differential resistance region (NDR) (Fig. 3b). To emulate VO₂ behavior in circuit simulations, we use the compact model from Maffezzoni et al. with circuit parameters listed in Table I.

### Table I

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_{DD} )</td>
<td>2.5 V</td>
</tr>
<tr>
<td>( R_S )</td>
<td>20 kΩ</td>
</tr>
<tr>
<td>( C_P )</td>
<td>500 pF</td>
</tr>
<tr>
<td>( R_{ins} )</td>
<td>100.2 kΩ</td>
</tr>
<tr>
<td>( R_{set} )</td>
<td>0.99 kΩ</td>
</tr>
<tr>
<td>( V_L )</td>
<td>1 V</td>
</tr>
<tr>
<td>( V_H )</td>
<td>1.99 V</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>200</td>
</tr>
<tr>
<td>( V^* = V_{DD} - V_L )</td>
<td>1.5 V</td>
</tr>
<tr>
<td>( V^c = V_{DD} - V_H )</td>
<td>0.501 V</td>
</tr>
<tr>
<td>( \tau_0 )</td>
<td>10 ns</td>
</tr>
<tr>
<td>( T_{osc} )</td>
<td>21.6 μs</td>
</tr>
<tr>
<td>Simulation time step</td>
<td>1 ns</td>
</tr>
</tbody>
</table>

#### C. Two-Coupled Oscillators

ONN initialization assigns oscillators initial phases that correspond to the input image. Here, we initialize input phases by delaying oscillators’ starting time with respect to the reference oscillator, as \( \Delta \Phi_{in}^n = \Delta \Phi_{in} - 2\pi \), with \( T_{osc} \) the natural oscillation period [30]. Oscillators’ dynamics evolve by exchanging current through the coupling resistor \( R_C \). Fig. 4 shows the simplest configuration with two VO₂-oscillators. Coupling switches allow precise ONN initialization and are closed once all oscillators are turned on. Analogous to a synaptic weight in HNN, \( R_C \) determines the coupling strength between two oscillators, and hence the final phase state. Fig. 5 shows the relationship between 1) the input phase, 2) the coupling resistance \( R_C \), and 3) the output phase once the oscillators settle. When \( R_C \ll R_C^0 \), oscillators are in-phase and \( R_C \) implements a positive synaptic weight. Whereas for \( R_C \gg R_C^0 \), oscillators are out-of-phase and \( R_C \) emulates a...
negative weight \( \Delta \Phi \). \( \Delta \Phi \) allows the retrieval of one of the two states. Fig. 5b and c show examples with \( R_{C_2} = R_{C_2} \) where the two oscillators settle to in-phase and out-of-phase states, respectively. This two-coupled oscillator case represents the smallest ONN used as an associative memory. In the next section, we investigate large size ONNs with \( N \) oscillators coupled by resistances \( R_{C_N} \), and we report on their performances.

### III. ONN Scaling

#### A. Simulation set-up

We have developed an ONN circuit simulation platform in Matlab that includes VO\(_2\) device parameters (compact model [31]) and coupling parameters to allow transient simulation of different size ONNs. We consider each oscillator as a pixel of an image. To avoid biased results due to specific training sets, we generate random training sets composed of \( M \) random black and white patterns \( \xi^{\mu} \), \( \mu \in \{1, 2, ..., M\} \) with \( \xi^{\mu}_i = -1 \) if pixel \( i \) is black, or \( \xi^{\mu}_i = +1 \) if white. We impose the same proportion of black and white pixels in the training patterns to avoid any effect emerging from unbalanced patterns. Next, we apply the Hebbian learning rule [5] to the training set and we obtain a matrix of synaptic coefficients \( H \). We compute the corresponding coupling resistances \( R_{C_N} \) using the mapping function described in [30].

To generate test sets, we use the training patterns in which we apply a random uniform noise taking values between -1 and +1 (noisy pixels are gray), as in the example of Fig. 1. We vary the number of noisy pixels up to 50%. For inference, we initialize the ONN with a test image by delaying the starting time of oscillators. Then, we solve ONN dynamics using circuit equations along with the VO\(_2\) model [31]. As VO\(_2\)'s state equation is nonlinear, we solve it numerically at each time step using Newton-Raphson’s algorithm.

#### B. ONN Recognition Accuracy

We study how the ONN recognition accuracy varies with \( M \) and the number of noisy pixels in the test images. We set \( N=36 \) and we consider that the ONN recognition fails if at least one pixel differs from the corresponding training image. Fig. 6 shows the simulation results. When \( M = 2 \), the ONN recognition accuracy is larger than 70\% for test images having up to 30\% of noise. However, the accuracy dramatically drops for more stored patterns. Such as, for \( M = 4 \), the ONN recognition accuracy is around 30\% for the same level of input noise. To predict the ONN memory capacity for \( N \) oscillators, we perform multiple simulations in the following subsection. To the best of our knowledge, this is the first systematic study to derive the ONN memory capacity versus ONN size and accuracy.

#### C. ONN Memory Capacity

We report on ONN memory capacity when trained using the Hebbian learning rule. For different ONN sizes from \( N=8 \) up to \( N=100 \), we vary the size \( M \) of the training set. Fig. 7 shows the ONN recognition accuracy for test images having 10\% of noise. As expected, larger networks can store more patterns. ONN with \( N=100 \) stores \( M=16 \) patterns with a recognition accuracy larger than 50\%. Whereas 16 oscillators are limited to \( M=6 \) patterns for a similar accuracy. This trend is in accordance with Hopfield’s results [26]. Based on Fig. 7, we extract the ONN memory capacity when recognition accuracy reaches 50\%. Results are shown in Fig. 8. We derive that the ONN memory capacity grows linearly with a fitted slope of 0.146, in accordance with the scaling factor of 0.15 derived by Hopfield [26].

To increase the ONN memory capacity, one could think of having large networks but the number of synapses scales quadratically as \( N(N - 1)/2 \) and would make the physical implementation of large designs very challenging. Moreover, it is worthwhile to mention that noise would also limit the ONN scaling as the thermal noise increases with the number of synaptic resistors \( R_{C_N} \). Next, we derive a first-order estimation of the maximum fully-connected ONN size when the synaptic thermal noise is the predominant noise source.
thermal noise limits the number of oscillators to \( N \) oscillators. With that has been reported \[32\] in the case of two coupled-ring tudes. We consider SNR=3.5 as minimum achievable SNR with respect to the oscillation frequency for various ampli-

f

f for high frequency ONN operation.

In a previous work \[32\], Csaba and Porod highlighted the ONN robustness to electronic noise and have shown that ONN can tolerate a smaller SNR compared to amplitude-

based computing systems achieving the same functionality. ONN can tolerate a smaller SNR compared to amplitude-

Here, we study the ONN energy scaling using a bottom-up approach, i.e., starting from device and circuit level before scaling up to the architecture level. We show analytically and by circuit simulations that the ONN energy scales linearly with the number of oscillators.

A. Single Oscillator Energy Footprint

From circuit equations and Fig 3h, we derive the instantaneous power consumption of a single oscillator as:

\[
P(t) = V_{DD} \left( \frac{V_{out}}{R_S} + C_P \frac{dV_{out}}{dt} \right)\]

As \( V_{out} \) is a \( T_{osc} \)-periodic signal, the oscillator energy loss for one oscillation is given by

\[
E_{osc} = \frac{V_{DD}}{R_S} \int_0^{T_{osc}} V_{out} dt
\]

Then, we introduce the output mean voltage \( \overline{V_{out}} = 1/T_{osc} \int_0^{T_{osc}} V_{out} dt \) to reformulate the last expression as:

\[
E_{osc} = \frac{V_{DD}}{R_S} \overline{V_{out}} T_{osc}
\]

IV. ONN Energy Scaling

D. ONN Size and Noise Limitation

In a fully-connected ONN of size \( N \), each oscillator sees \( N-1 \) noisy synaptic resistors \( R_{C_n} \) (no self coupling) and its equivalent noise source expressed in \([V^2/Hz]\) is:

\[
\overline{V^2} = (N-1) \frac{4k_B T R_{C_n} f_{osc}}{2} \tag{1}
\]

where \( k_B \) is the Boltzmann's constant, \( T \) the temperature, and \( f_{osc} = 1/T_{osc} \) is the oscillation frequency. We assume that the intrinsic oscillator noise is negligible with respect to the synaptic thermal noise when \( N \) is large. As a first-order approximation, we only consider thermal noise \( \[1\] \) because we are interested in scaling up \( f_{osc} \) for high frequency ONN operation.

We express the oscillator SNR as the ratio between the peak-to-peak voltage amplitude over the thermal noise standard deviation:

\[
SNR = \frac{\Delta V_{max}}{\sqrt{\overline{V^2}}} \tag{2}
\]

When increasing \( N \), we scale \( R_{C_n} \) as \( R_{C_n} = (N-1)R_{C_2} \), where \( R_{C_2} \) is the coupling resistance for an ONN composed of two oscillators only \[30\]. We then express the maximum fully-connected ONN size \( N_{max} \) combining \( [1] \) and \( [2] \):

\[
N_{max} = 1 + \frac{\Delta V_{max}}{4k_B T R_{C_2} f_{osc} SNR} \sqrt{2} \tag{3}
\]

Fig[9] shows the maximum fully-connected ONN size \( N_{max} \) with respect to the oscillation frequency for various amplitudes. We consider SNR=3.5 as minimum achievable SNR that has been reported \[32\] in the case of two coupled-ring oscillators. With \( \Delta V_{max}=21 \text{ mV} \), we observe that the synaptic thermal noise limits the number of oscillators to \( N_{max}=300 \) for \( f_{osc}=10 \text{ MHz} \). For applications that need large ONNs, the

oscillation frequency has to be reduced and the amplitude should increase. It is worthwhile to note that the minimum SNR might also depend on the ONN size, as Csaba and Porod \[32\] reported correct functionality for 100 coupled oscillators even for SNR<1. In literature, coupling oscillators has been shown efficient to reduce the phase noise \[33\], but yet little is known on the impact of noise on the oscillator synchronization and scaling of phase-based computing systems.
capacitors in digital circuits $E_{\text{dyn}} = C_p V_D^2$. However, in our case, the oscillator energy loss (6) is modulated by the DC output voltage operating point $V_{\text{out}}$. Note that closed-form expressions for $V_{\text{out}}$ and $T_{\text{osc}}$ are established in [30] but are not listed here for clarity. We observe that the two key knobs to obtain low energy ONN are low operating voltages and low parasitics. Next, we derive the oscillator energy when coupled to $N-1$ other oscillators.

**B. ONN Synaptic Operations**

We first define the intrinsic synaptic operation between coupled oscillators. For oscillator $i$, we conceptually express its synaptic input weighted sum $h_i(t)$ as:

$$h_i(t) = \sum_{j=1}^{N-1} W_{ij} \Delta \phi_j(t)$$

where $W_{ij}$ are the synaptic weights and $\Delta \phi_j(t)$ are the phases of other oscillators (Fig.10c). Then, the role of the oscillating neuron is to produce an output phase by applying a non-linear activation function $a$ to its input:

$$\Delta \phi_i(t) = a(h_i(t))$$

We define a synaptic operation (SOP) in ONN as the evaluation of the quantity $W_{ij} \Delta \phi_j(t)$. Note that up to now, we have not considered any hardware, and SOP could be implemented in various manners such as with digital circuits [9] or using the analog Ohm’s law. Using these definitions, we express the neuron energy as the sum of two contributions:

$$E_{\text{neuron}} = E_{\text{input}} + E_{\text{activation}}$$

$E_{\text{input}}$ is the loss related to the evaluation of the input weighted sum, whereas $E_{\text{activation}}$ is the energy needed to produce an output, i.e., determine the phase difference. Again, (6) is general enough so it can capture any type of implementation and computing (sequential or parallel). In the interesting case where neurons process information in parallel, we can then express the neuron energy as:

$$E_{\text{neuron}} = ((N-1) E_{\text{SOP}} + E_{\text{osc}}) \frac{T_{\text{osc}}}{N_{\text{cycles}}}$$

where $N_{\text{cycles}}$ is the number of oscillating cycles before settling to a stable output phase state, and $E_{\text{osc}}$ is the energy of a single oscillation. One interesting aspect of analog ONN is that sometimes SOP can be energy-free. For instance, when two coupled oscillators are in-phase the synaptic current is null and $E_{\text{SOP}}=0$ (see Fig.10c). The worst-case SOP energy occurs when two oscillators are out-of-phase: the maximum amplitude across the synaptic resistor reaches $\Delta V_{\text{max}} = V_H - V_L$ and induces Joule’s loss (see Fig.10b). As SOP analytical expression depends on the oscillating waveform, we evaluate here the worst-case for simplicity and we consider that a DC voltage $\Delta V_{\text{max}}$ is applied to every coupling resistor $R_{\text{C}}$ during the entire oscillating period:

$$E_{\text{SOP}} = \frac{\Delta V_{\text{max}}^2}{R_{\text{C}}} T_{\text{osc}}$$

To assess how the ONN energy scales with $N$, we must first evaluate the ONN computation time, $N_{\text{cycles}}$. Next, we perform circuit simulations of various ONN sizes dedicated to pattern recognition to estimate $N_{\text{cycles}}$.

**C. ONN Settling Time and Energy Scaling**

We define the ONN settling time as the time $t_{\text{settle}}$ required for ONN signals to be periodically stable:

$$t_{\text{settle}} = N_{\text{cycles}} T_{\text{osc}}$$

For $t \geq t_{\text{settle}}$, ONN phases can be measured as they are stable. For example in Fig.11, ONN stabilizes to a stable pattern after $N_{\text{cycles}} = 1.75$ cycles. To derive the ONN settling time, we perform simulations for different ONN sizes by varying 1) the number $M$ of stored patterns and 2) the number of noisy pixels.
in test images from 10% to 50%. Fig.11 shows the simulation results. Interestingly, the ONN settling time is approximately constant and is smaller than 5 cycles in most cases. Hence, ONN parallel computation can allow to compute in constant time even for large networks. This result corroborates what has been observed with oscillator-based Ising machines [16]. i.e., coupled oscillators converge to a solution (not necessarily the optimal one) in constant time.

Moreover, we derive that the ONN energy scales linearly (see Fig.11) when ONN satisfies the two following properties:

1) **Parallelism:** the computation time \( t_{\text{settle}} \) remains quasi-constant.

2) **Downscaling of synaptic energy:** we scale the coupling resistors \( R_{C_b} \) as:

\[
R_{C_b} = (N - 1)R_{C_2} \tag{13}
\]

where \( R_{C_2} \) is the coupling resistance between two coupled oscillators [30]. The synaptic loss \( E_{\text{SOP}} \) becomes:

\[
E_{\text{SOP}} = \frac{\Delta V_{\text{max}}^2}{(N-1)R_{C_2}} T_{\text{osc}} \tag{14}
\]

Therefore, even though the number of synapses grow quadratically, the ONN energy grows only linearly with the number of oscillators. This can be verified using our previous definitions [6] [10] [14]:

\[
E_{\text{analog}} = N E_{\text{neuron}} = N((N-1)E_{\text{SOP}} + E_{\text{osc}})N_{\text{cycles}} \tag{15}
\]

\[
= N(\frac{\Delta V_{\text{max}}^2}{R_{C_2}} + \frac{V_{DD}}{R_S} V_{\text{out}})N_{\text{cycles}}T_{\text{osc}}
\]

\[
E_{\text{digital}} = N((N-1)E_{\text{MAC}} + E_{\text{acc}})N_{\text{cycles}} \tag{16}
\]

In our simulations, we considered ONNs with a large supply voltage \( V_{DD} = 2.5V \) leading to an important energy consumption of 2 nJ/oscillator/cycle. Whereas, Jackson et al. in [8], have designed an ONN consuming 1.21 pJ/oscillator using a hybrid design (analog synapses and digital neurons) in 28 nm CMOS technology. Next, we study how to scale VO\(_2\) devices to achieve competitive performances with respect to state-of-the-art solutions.

**D. Oscillator Energy Minimization using Scaled VO\(_2\) Devices**

Here, we study how to minimize the energy for a VO\(_2\)-based oscillator using the formulation [6] and assisted by TCAD simulations. The TCAD modeling and simulation flow are further described in recent work [34], [35]. We consider VO\(_2\) devices in crossbar (CB) geometry [14] as a potentially scalable geometry to lower the oscillator energy consumption (Fig.12a). By reducing the VO\(_2\) CB size, the overall VO\(_2\) thermal dissipation decreases and the VO\(_2\) device can transition to a metallic state with less power [35]. The applied voltage can then be reduced for given insulator and metallic states that are set by material properties and contact area (Fig.12b).

As our model predicts that the oscillator energy scales quadratically with voltages [6], it is of interest to scale down VO\(_2\) CB dimensions. Fig.13 shows results of TCAD simulations for various CB (500nm, 1\(\mu\)m, 1.5\(\mu\)m, 2\(\mu\)m, 3\(\mu\)m and 4\(\mu\)m) and biasing parameters. We see from Fig.13b that VO\(_2\) threshold voltages \( V_H \) and \( V_L \) are approximately proportional to CB and allow a linear \( V_{DD} \) scaling. With reduced CB, the oscillating voltage amplitude can be decreased (Fig.13b) for low power operation (Fig.13b). As we kept the same material, contact area, and load capacitor for all CB sizes, the oscillating period does not vary significantly and the minimum energy is obtained for CB=500nm (Fig.13b).

Fig.14 shows the comparison between our analytical model [6] and mean power and energy computed with TCAD for different CB sizes. We observe a good match for the mean power but some deviation when evaluating the energy. We believe this is mainly due to non-linearities induced by thermal
effects, which result in a larger oscillating period thus a higher energy consumption [32]. This aspect is not captured by our analytical formalism as it only considers electrical variables (Fig. 13b). Nevertheless, the scaling trend of our model is in agreement with TCAD simulations and we use it for benchmarking ONN with state-of-the-art chips.

V. ONN BENCHMARKING

A. Neuron Energy-Delay Benchmark

Benchmarking ONN with other architectures is not trivial as ONN is a phase-based system and does not perform conventional MAC operations. However, the concept of synaptic operation is shared among all sorts of neural inference chips and can serve as common ground for benchmarking. In Artificial Neural Networks (ANN), a SOP is defined by the multiplication between the input and the synaptic weight. Then, it can be naturally implemented in digital hardware by a MAC operator and in this case, there is the equivalence described in [41] that consumes 20.5 \( \mu \)W in 28nm CMOS technology suitable for low power edge applications. For the phase measurement, we take the example of the circuit model is in agreement with TCAD simulations and we use it for benchmarking ONN with state-of-the-art chips.

Fig. 13. TCAD simulation results for the same crossbar (CB) geometry and \( C_p=50 \text{fF} \). (a) Oscillator parameters with respect to the \( V_{O_2} \) CB. By scaling down the \( V_{O_2} \) CB, the thermal dissipation decreases and the device needs less power to transition from one state to the other. Therefore, the \( V_{O_2} \) thresholds \( V_H \) and \( V_L \) decrease with CB. We scale down \( V_{DD} \) approximately linearly with \( V_H \) and \( V_L \). The load resistor \( R_S \) is adapted in each case to place the load line in the NDR region and obtain oscillation. (b) Transient voltage across \( V_{O_2} \) devices for different CB. (c) Instantaneous power for different CB. Scaling down CB leads to low oscillation amplitude and low power. (d) Oscillator energy vs period for various CB.

In Digital Neural Accelerators performing MACs, the neuron delay is estimated as \( \text{delay}_\text{neuron} = N/\text{T}_{\text{ch}} \). However, in literature, faster \( V_{O_2} \) oscillations where circuits employ nano-farad load capacitors could reach a similar speed with lower load capacitors. Thus, we project the oscillator energy and delay for lower capacitances down to 500 fF using our analytical model (6) and (12).

To obtain a more precise energy assessment, we include the power consumed by peripheral circuits, i.e., the phase initialization and measurement circuits. To set the oscillator input phase, we would use in the worst case one digital-to-time converter (DTC) per oscillator. As an example, we consider a 9-bits DTC consuming 31 \( \mu \)W at 40 MHz in 28nm CMOS technology suitable for low power edge applications. For the phase measurement, we take the example of the circuit described in [41] that consumes 20.5 \( \mu \)W in 28nm CMOS technology. Overall, we consider \( P_{\text{periph}}=60 \mu \)W of peripheral circuits per oscillator clocked at 30 MHz, which gives 2 pJ per cycle. As a first-order estimation, we consider that \( P_{\text{periph}} \) is proportional to the neuron oscillating frequency and we obtain a constant peripheral energy loss \( E_{\text{periph}} = P_{\text{periph}} T_{\text{osc}} N_{\text{cycles}} \). We use \( N_{\text{cycles}} \approx 5 \) derived in section IV.C and we obtain \( E_{\text{periph}}=10 \) pJ. Note that our estimation remains optimistic as we use a bottom-up type of energy-delay assessment, whereas state-of-the-art data correspond to real chip measurements.

Fig. 14 shows the neuron energy-delay for various SNN neuromorphic chips (blue circled dots), digital neural accelerators (red squared points) considered in previous work [36], [37] and \( V_{O_2} \) oscillators with different load capacitances.
When the oscillator load capacitance increases, the oscillator slows down and its energy to produce a stable output phase increases. Similarly, neuromorphic SNN chips lie on the right-hand side of the plot as they generally produce spikes at lower frequency than digital neural accelerators [45]. From the neuron energy point of view, it appears that VO\textsubscript{2}\textsuperscript{-}based ONN can compete with state-of-the-art SNN neuromorphic chips for a similar neuron delay. With real chip measurements which would include all peripheral energies, we expect the ONN region to shift up and to lie in the SNN neuromorphic region in the worst case.

The VO\textsubscript{2} oscillator could compete with neural accelerators at energy level but would be orders of magnitude slower with load capacitances larger than 500fF. For instance, a neuron from PuDianNao [42] accelerator produces an output after 242 ps whereas it would take 16 ns to phase lock for a scaled VO\textsubscript{2} oscillator with \(C_P=500\text{fF}\). We notice that peripheral circuits set the minimum achievable neuron energy for load capacitances smaller than 50 pF (green diamond points), whereas the energy of the ONN neuron standalone can be below the picojoule range (orange star points). From our first-order estimation, we conclude that the energy-delay of a VO\textsubscript{2}-ONN can be very competitive under the two following conditions:

1) the oscillating frequency is in the GHz range, \textit{i.e.}, the load capacitance \(C_P<50\text{ fF}\) and assuming that the VO\textsubscript{2} thermal time constant remains negligible [45].

2) careful design of peripheral circuits to fully take advantage of ONN phase computing paradigm.

As an alternative of VO\textsubscript{2} oscillators, CMOS ONNs (purple triangular points) are currently very competitive as they use scaled transistors from a mature CMOS technology. For instance, the first phase-based ONN chip ever reported for pattern recognition is the digital ONN designed by Jackson et al. [8] with 100 neurons and 10,000 synapses using a 28nm CMOS technology. Their results are promising as they measured a 1.21 pJ neuron energy and 4 ns delay. For fast convolution inference, Nikonov et al. [44] recently reported on an ONN chip fabricated in 22nm FinFet CMOS process that computes in less than 10 ns and consumes 2 pJ/oscillator. In the field of oscillator-based Ising machines (OIM) [16], Ahmed et al. [17] revealed an OIM composed of 560 ring oscillators in 65nm CMOS technology that consume 1.74 pJ/oscillator for \(N_{cycles}=5\) and \(f_{osc}=118\text{ MHz}\). These recent examples further highlight the ONN potential to perform various tasks at high speed and low energy.

Finally, we would like to stress that benchmarking different architectures at the neuron level only gives a limited vision of chips' potential as they are ultimately used to solve practical problems. For example, Nikonov's ONN and the neural accelerator DianNao [43] have almost the same energy-delay when used to compute convolutions [7]. Next, we choose to benchmark a VO\textsubscript{2}-ONN in the case of image edge detection.
Fig. 16. a) 10 fully-connected oscillators trained to detect vertical, horizontal and diagonal edges in images. b) Mapping of Hebbian coefficients to coupling resistors. c) ONN state that detects the image background.

which is a widely used task in image processing.

B. Edge Detection Benchmark with ONN

Here, we aim to benchmark VO\textsubscript{2}-ONNs with other works on a specific image edge detection application. Similar to edge detection algorithms that employ 3x3 or 5x5 convolution kernels [44], we scan an input image with 3x3 ONN to extract edges. Analogous phase-based edge detection algorithms have already been proposed in literature [45], [46] but we rather focus on the analog hardware implementation to assess how a VO\textsubscript{2}-ONN benchmarks with state-of-the-art edge detection hardware.

We consider a fully-coupled ONN composed of 10 oscillators and 45 coupling resistors where 9 oscillators scan the input image with a padding of 1, and the 10th oscillator makes the final decision (Fig.16a). Using the Hebbian learning rule [5], we train the ONN to detect edges in the vertical, horizontal and diagonal direction and we map the Hebbian coefficients to coupling resistors using the mapping function defined in [30] (Fig.16b). To detect the background, we bias the VO\textsubscript{2} oscillators such that the 0\degree phase state is more likely to occur (we set $R_S=6k\Omega$ instead of $20k\Omega$), further explained in [30]. As shown in Fig.16c and Fig.17, oscillators converge in-phase when initialized with similar input phases and the ONN detects the background. Fig.17 shows an example where the ONN detects a vertical edge. Note that the 10th output oscillator is always initialized with an input phase of 90\degree to not favor any particular output state. As already highlighted in section IV.C, the ONN makes the decision after few oscillation cycles only (between 3 and 5).

We compare our ONN image detection with the state-of-the-art Sobel and Canny edge detection methods [44], [47] that we test in Matlab using built-in functions. The results from Fig.18 show that Sobel, Canny, and ONN edge detections are qualitatively similar for a binary input image. A more interesting case consists in detecting edges in a gray-scale image as shown in Fig.19 with the 8-bits 64x64 gray-scale example. We observe that ONN detects more edges than Sobel and therefore evaluates well the image gradient. However, our ONN edge detection seems more sensible to noise than Canny.

that initially smooths the input image with a 5x5 gaussian kernel. We believe that larger ONN kernels such as 5x5 or 7x7 could produce similar denoising property but is beyond the scope of this paper.

Table II shows the performances of edge detection ASICs implemented in 65 nm [48] and 45 nm [49] CMOS technologies. Both accelerators are optimized to run the Canny algorithm and are suitable for edge applications thanks to their low power consumption. We consider a VO\textsubscript{2}-ONN with a crossbar size of 500 nm to achieve low power operations and we vary the load capacitance to set the oscillating frequency. A single ONN running at 31 MHz would process a 512x512 image in 42 ms and would be x100 slower than Soares’s ASIC [49]. By reducing the capacitance load to 500 fF and parallelizing at least 10 ONNs, ONN could compete with state-of-the-art to achieve 0.42 ms/image.

Again, the peripheral circuits’ energy could become domi-
TABLE II
Edge detection benchmark

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Frequency</th>
<th>Mean Power</th>
<th>Image size</th>
<th>Time /image</th>
<th>Energy/pixel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee 2018 [48]</td>
<td>ASIC (65 nm)</td>
<td>500 MHz</td>
<td>5.48 mW</td>
<td>1280x720</td>
<td>2.2 ms</td>
</tr>
<tr>
<td>Soares 2020 [49]</td>
<td>ASIC (45 nm)</td>
<td>350 MHz</td>
<td>6.7 mW</td>
<td>512x512</td>
<td>0.42 ms</td>
</tr>
<tr>
<td>ONN1</td>
<td>10 VO₂-oscillators C=5 pF</td>
<td>31 MHz</td>
<td>13 μW + 330 µW (periph.)</td>
<td>512x512</td>
<td>42 ms</td>
</tr>
<tr>
<td>ONN2</td>
<td>10 VO₂-oscillators C=500 fF</td>
<td>310 MHz</td>
<td>13 µW + 3.3 mW (periph.)</td>
<td>512x512</td>
<td>4.2 ms</td>
</tr>
</tbody>
</table>

(a) Input image  (b) Sobel  
(c) Canny  (d) ONN

Fig. 19. a) 64x64 8-bits gray scale image [1]. b), c) and d) are the output images using Sobel, Canny and ONN edge detection methods, respectively.

VII. Conclusion

In this work, we derived the performance scaling laws of VO₂-ONNs at device, circuit, and architecture levels. We first studied ONNs used as associative memories and we derived that the ONN memory capacity scales as 0.15N when trained with the Hebbian learning rule, similarly to Hopfield Neural Networks. Next, we presented the trade off between the ONN size, SNR and frequency due to the thermal noise produced by the coupling resistors. We also showed that the constant ONN settling time leads to a favorable linear energy scaling when increasing the coupling resistance values. Assisted by TCAD simulations, we then proposed some design guidelines at device and circuit levels to build competitive VO₂-ONNs with respect to state-of-the-art chips. Finally, we applied our methods to an image edge detection application using a scaled VO₂-ONN that is suitable for low-power edge devices.

Acknowledgment

This work is supported by the European Union’s Horizon 2020 research and innovation program, EU H2020 NEURONN (www.neuronn.eu) project under Grant No. 871501.

Data Availability

The Matlab source codes will be made available by the authors upon acceptance of the manuscript.

References


