NCPower: Power Modelling for NVM-based Neuromorphic Chip

Zhehui Wang[1], Huaipeng Zhang[1], Tao Luo[1], Weng-Fai Wong[2], Anh Tuan Do[1]
Paramasivam Vishnu[1], Wei Zhang[3] and Rick Siow Mong Goh[1]
[1]Agency for Science, Technology and Research (A*STAR), Singapore
[2]National University of Singapore [3]The Hong Kong University of Science and Technology
wang_zhehui@ihpc.a-star.edu.sg leto.luo@gmail.com

Abstract

Spiking neural networks (SNN) on non-volatile memory (NVM) based neuromorphic computing (NC) chips have been regarded as a promising solution in power constrained scenarios, such as Internet of Things (IoT), due to its low energy consumption. The high power efficiency of NC is due to various aspects including the non-von Neumann architecture of NC chip, low power NVM, and the event driven computation of SNN etc., and introduces a large space for low power design exploration. Therefore, a comprehensive quantitative study of the power modelling for such neuromorphic computing system is important for low power design. In this work, we propose NCPower, an energy consumption estimator for NVM-based neuromorphic chip. We systemically developed analytical models based on physical laws, and verify them by comparing the analytical results with measurement results from different neuromorphic chips. We integrated NCPower in a simulator, and analyzed the accuracy and energy consumption of both the traditional multi-spike based SNN and the new single-spike based SNN. It shows that the single-spike model has 7X energy efficiency over the multi-spike model, with similar accuracy under the CIFAR-10 dataset.

CCS Concepts

• **Hardware → Power estimation and optimization**.

Keywords

datasets, neural networks, gaze detection, text tagging

1 Introduction

Spiking neuron networks (SNN) is one promising technique in the field of artificial intelligence. The key component of SNN is the spiking neurons, which mimics the behavior of biology neurons. In SNN model, the input and output of the neuron is interpreted into spikes. Each neuron can independently trigger the output spikes regarding the frequency of input spikes and the weights. Given the larger amount of neurons in SNN model, a tremendous amount of computation power is required [6]. Today, the neuromorphic chip is an efficient way to accelerate the computation process, which adopts an non-Von Neumann architecture and shows very high parallelism [11]. One branch of the neuromorphic chip is based on the non-volatile memory (NVM), where its conductance is adjusted so that its value is proportional to the related weight [14]. NVM is a promising energy-efficient memory technology which adopted in both memory design [15] and in-memory logic design [8, 9].

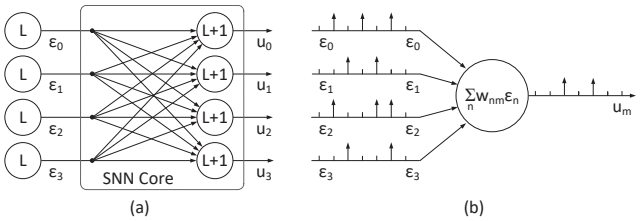

Figure 1: (a) The structure of SNN soft core; (b) The behavior of a single neuron in the soft core

The systematical estimation of NVM-based neuromorphic chip's energy consumption is very important during design stage. For example, one important configuration of the NVM resistor is the range of its resistance. On one hand, NVMs with lower resistance are less affected by the crosstalk noise in the analog circuits, which results in higher accuracy of the SNN model. On the other hand, the energy consumption of NVM resistors would increase if its resistance decrease. Hence, we need to precisely evaluate the energy consumption of the whole chip in order to make a proper trade-off between the energy and the model accuracy. Developers also want to find the ratio between the energy consumption of NVM resistors over the total energy consumption of the chip.

However, as an emerging technology, the power consumption of the NVM-based neuromorphic chip is hard to obtain from the EDA tools. Few paper have touched this topic before. Yakopcic *et al.* analyzed the power consumption of memristor based neuromorphic processor during training [17]. Stromatias *et al.* modelled the power consumption of large spiking neural networks running on multiple GPGPUs and FPGAs [12]. Dong *et al.* [3] modelled the energy consumption for emerging nonvolatile memory. Salkhordeh *et al.* [10] modelled the performance of hybrid DRAM-NVM main memories. None of these paper have discussed the power modelling of NVM-based neuromorphic chip during inference.

This paper proposes NCPower, an energy consumption estimator for NVM-based neuromorphic chip. We systemically develop analytical models for each module in the neuromorphic chip based on physical laws. Afterwards, we calibrate and verify the results of our estimator with published data from fabricated devices. To show the compatibility of the analytical model, we test this estimator on different neuromorphic chip designs using different NVM devices. One of the chips is our in-house developed Novena chip.

ICONS 2020, July 28–30, 2020, Oak Ridge, TN, USA

ACM ISBN 978-1-4503-8851-1/20/07…$15.00
https://doi.org/10.1145/3407197.3407619

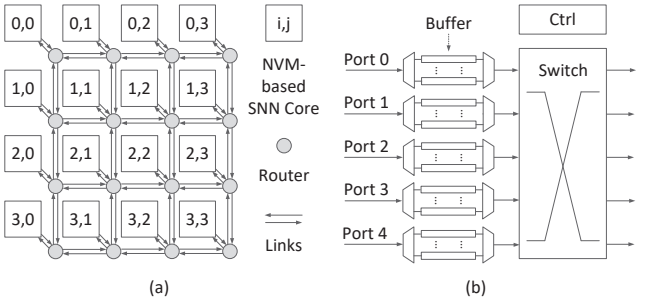

Figure 2: (a) The overview of the neuromorphic chip with 4×4 hard cores; (b) The structure of the 5-port router

NCPower is also integrated into a SNN simulator so that we can analyze the energy consumption and accuracy of the neuromorphic chip concurrently. It can generate the real-time heat map of the chip, and show the variations of power consumption over time. Using NCPower, developers can check if a certain location of the chip will be over-heated, and try to alleviate the thermal issues by optimizing soft core and hard core mapping strategy. Developers can also check if the power consumption would exceed the power limit at certain time, and try to avoid the power wall of the chip by optimizing the scheduling policy.

2 Modelling of Energy Consumption

The SNN model can be decomposed into multiple soft cores, whose structure is shown in Figure 1 (a). A soft SNN core contains multiple neurons, which receive input spikes from other neurons and generate output spikes based on the incoming spikes. Typically, the spikes from neurons in level L can reach all the neurons in level L+1. As shown in Figure 1 (b), each neuron can receive and generate spikes independently. The relationship between incoming spikes and the output spike is expressed in Equation 1.

$$u_m(t+1) = u_m(t) + \sum_n w_{nm} \cdot \epsilon_n(t) + \beta_m \tag{1}$$

$u_m(t)$ is the membrane potential of neuron m at time step t. $\epsilon_n(t)$ is the spiking from neuron n in level L. it equals 1 if the spike exits, and equals 0 if not. w_{nm} is the weight of the spike ϵ_n for neuron m in layer L+1. β_m is the bias value for neuron m. If the membrane potential at time step t+1 is larger than a threshold, then the potential is reset to a constant, which typically equals 0.

The overview of the neuromorphic chip is depicted in Figure 2 (a). The central computation unit in the neuromorphic chip is the hard core. In neuromorphic chip, there is a matrix of hard cores. Each hard core can emulate the behavior of a specific soft core. The data transmission in the neuromorphic chip is handled by the network-on-chip (NoC). Each hard core is connected to a router, which is connected into a mesh-based NoC. Spikes from one hard core is delivered to another hard core via the NoC.

$$E_{total} = \sum_i \sum_j \sum_k (E^{ijk}_{XBAR} + E^{ijk}_{RTR}) \tag{2}$$

The total energy consumption of the neuromorphic chip can be expressed in Equation 2. E^{ijk}_{XBAR} and E^{ijk}_{RTR} are the energy consumption of the hard core and its adjacent router at location (i,j) of the neuromorphic chip, at cycle k.

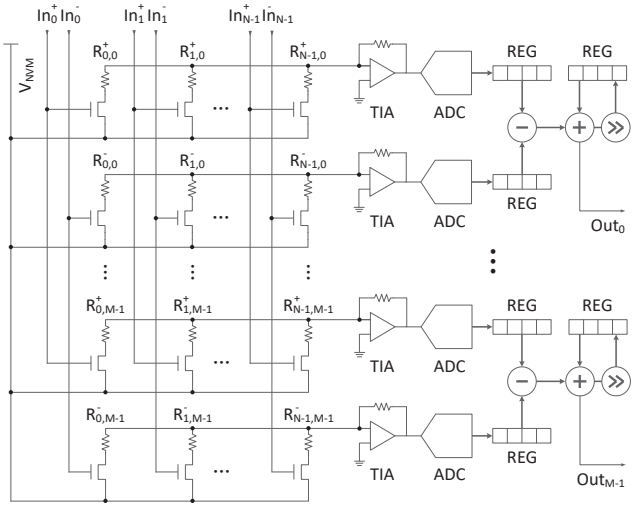

Figure 3: The overview structure of the NVM-based hard core, which consists of a matrix of NVM resistors, TIA (trans-impedance amplifier), ADC (analog-to-digital converter), REG (register), and arithmetic unit

2.1 Network-on-Chip

Router is the major component in NoC. The structure of router is shown in Figure 2 (b). Different routers have different number of ports depending on their locations in the network. For example, a router in the center of the network has five ports. Among them, four ports are used to transmit packets from four directions, and the fifth port is used to transmit packets from its adjacent hard core. At each input port of the router, there is an array of buffers, which temporarily store the packets. A switch transmits the packet in the buffer to its destination router. The control unit monitors the hardware resources and decides which packet in the buffer can be delivered first, typically based a round-robin mechanism [1].

$$E_{RTR} = E_{SW} + E_{LINK} + E_{CTRL} + E_{BUF} \tag{3}$$

According to [7], the energy consumption of routers is expressed in Equation 3, which includes the energy consumption of the switch E_{SW}, the link drivers E_{LINK}, the control unit E_{CTRL} and the buffers E_{BUF}. Since the links among routers are passive devices, their energy consumption is included in the link drivers.

$$E_{SW} = e_{sw} \cdot \gamma \cdot F \tag{4}$$

$$E_{LINK} = e_{link} \cdot \gamma \cdot F \tag{5}$$

The energy consumption of the switch and the link drivers are proportional to the number of bits transferred via the router. They are expressed in Equation 4 and Equation 5. e_{sw} and e_{kink} are the energy consumption per bit of the switch and the link driver, respectively. γ is the number of transaction processed in the router, during a cycle period. F is the packet size in bits, and $\gamma \cdot F$ is the total number of bits transmitted.

$$e_{link} = 0.5 \cdot \sqrt{S_{CORE}} \cdot C_0 \cdot V^2_{LINK} \tag{6}$$

The energy consumption per bit of the link driver is expressed in Equation 6. S_{CORE} is the area of a single hard core, and $\sqrt{S_{CORE}}$ is the average router-to-router distance. C_0 is the capacitance per unit

Table 1: Parameters of Analog Circuits

Notation	Device	Description	Value	Ref.
V_{TIA}	TIA	Supply voltage	0.2 V	[20]
I_{TIA}	TIA	Supply current	6 nA	[20]
t_{acq}	ADC	Acquisition time	100 ns	[20]
t_{conv}	ADC	Conversion time per bit	100 ns	[20]
V_{ADC}	ADC	Supply voltage	0.4 V	[20]
$\overline{I_{ADC}}$	ADC	Average supply current	0.15 μA	[20]
V_{NVM}	NVM	Supply voltage	0.2 V	[5]
S_{NVM}	NVM	Area overhead per resistor	0.4 μm^2	[5]

length of the link. V_{LINK} is the supply voltage of the link driver.

$$E_{CTRL} = e_{ctrl} \cdot \gamma \tag{7}$$

The energy consumption of the control unit is proportional to the number of transactions completed in a cycle period. It is expressed in Equation 7. e_{ctrl} is the energy consumption per transaction of the control unit.

$$E_{BUF} = (e^b_{rd} + e^b_{wr}) \cdot \gamma \cdot F + P^b_s \cdot t_0 \cdot \sum_p B_p \tag{8}$$

The energy consumption of the buffer unit is expressed in Equation 8. e^b_{rd} and e^b_{wr} are the read and write energy consumption per bit of the buffer unit. P^b_s is the static power consumption of the buffer unit. t_0 is the duration time of one cycle. B_p is the capacity of the buffer at input port p.

2.2 NVM-based Hard Core

The overview of the NVM-based hard core is shown in Figure 3. It is able to emulate the behavior of a soft core. In general, it receives spikes from neurons in layer L, and generates output spikes of neurons in layer $L+1$, using digital-analog hybrid circuits. NVM resistors are the key components in the hard core. Suppose layer L has N neurons and layer $L+1$ has M neurons, a hard core contains a $N \times 2M$ matrix of NVM resistors. Each weight needs two NVM resistors, one for positive value and one for negative value. When we run the emulation, the supply voltage of the NVM matrix is fixed and the resistances of the NVM resistors are pre-tuned based on the weights in the SNN network. The tuning mechanism guarantees that the current going through each NVM is proportional to the corresponding weight, if the input spike exists.

$$I_{nm} = \alpha \cdot w_{nm} \cdot \epsilon_n \tag{9}$$

The relationship between the current and the weight is expressed in Equation 9. ϵ_n stands for the existence of spike from neuron n in layer L. It equals 1 if the spike exists and equals 0 if not. w_{nm} is the weight of spike from neuron n in layer L to neuron m in layer $L+1$. α is the coefficient between current I_{nm} and weight w_{nm}.

$$t_0 = t_{acq} + t_{conv} \cdot b + t_{cal} \cdot b \tag{10}$$

The during time of each cycle t_0 can be expressed in Equation 10. There are three stages within one cycle. First, it takes t_{acq} to charge the capacitance of the trans-impedance amplifier (TIA), and stabilize its output voltage. Next, the analog-to-digital converter (ADC) converts the analog signal to the digital signal. Given b as the width of the digital signal, the total converting time of ADC is $t_{conv} \cdot b$.

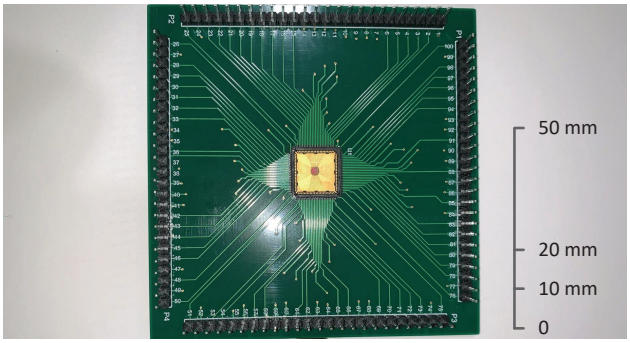

Figure 4: The tested Novena chip for validation, which is fabricated in 40 nm technology

Finally, it takes $t_{cal} \cdot b$ for the arithmetic unit to process the b-bit data from ADC, and to decide if the neurons would generate a spike.

$$E_{XBAR} = E_{NVM} + E_{TIA} + E_{ADC} + E_{REG} + E_{ARITH} \tag{11}$$

As expressed in Equation 11, the energy consumption of the NVM-based hard core includes the energy consumption on the NVM resistors E_{NVM}, the TIA E_{TIA}, the ADC E_{ADC}, the register E_{REG}, and the arithmetic unit E_{ARITH}.

Since resistance can only be positive, we need two NVM resistors R^+_{nm} and R^-_{nm} for one weight w_{nm}, in a complementary way. If $w_{nm} > 0$, then signal In^+_n is activated, which equals ϵ_n. At the same time, signal In^-_n is disabled. If $w_{nm} < 0$, then signal In^-_n is activated. At one time, current I_{nm} only flows through one of the two complementary NVM resistors, either via R^+_{nm} or R^-_{nm}.

$$E_{NVM} = V_{NVM} \sum_n \sum_m I_{nm} \cdot t_{acq} \tag{12}$$

The energy consumption of NVM resistors is expressed in Equation 12. Since I_{nm} only flows through one of the two complementary NVM resistors, the energy consumption on the other NVM resistor is zero. V_{NVM} is the supply voltage of the NVM resistors. After t_{acq}, the current is cut off.

$$E_{TIA} = V_{TIA} \cdot (\sum_n \sum_m I_{nm} + 2M \cdot I_{TIA}) \cdot t_{acq} \tag{13}$$

As expressed in Equation 13, the energy consumption of TIA includes the energy on the feedback resistor and the energy on the amplifier unit. V_{TIA} and I_{TIA} are the supply voltage and supply current of the TIA, respectively. Theoretically, the current of the NVM resistor will flow through the feedback resistor, where the current signal is converted to the voltage signal. At the same time, the amplifier stabilizes the voltage signal. In the hard core, currents flowing through the same row of NVM registers share the same ADC. For example, current I_{00} to current $I_{N-1,0}$ share the same ADC in row 0. According to physical laws, the current flowing through the feedback resistor of the ADC is the summation of all those currents flowing through NVM resistors related to that ADC.

$$E_{ADC} = 2M \cdot V_{ADC} \cdot \overline{I_{ADC}} \cdot t_{conv} \cdot b \tag{14}$$

The energy consumption of ADC is expressed in Equation 14. V_{ADC} and $\overline{I_{ADC}}$ are the supply voltage and average supply current of the ADC. The supply current I_{ADC} depends on the input voltage signal of the ADC. Therefore, we use the average current $\overline{I_{ADC}}$ to estimate

Table 2: Energy Consumption Breakdown (nJ)

Model	Resistance Range	NVM	TIA	ADC	REG	ARITH	BUF	LINK	SW	CTRL	Total
Multi-spike	1 kΩ-10 kΩ	$4.5x10^5$	$4.5x10^5$	$2.0x10^4$	$1.3x10^4$	$1.1x10^4$	$3.5x10^3$	$3.7x10^3$	$2.7x10^2$	$1.1x10^3$	$9.6x10^5$
	10 kΩ-100 kΩ	$4.4x10^4$	$4.4x10^4$	$2.0x10^4$	$1.3x10^4$	$1.1x10^4$	$3.5x10^3$	$3.6x10^3$	$2.6x10^2$	1.1x103	$1.4x10^5$
	100 kΩ-1 MΩ	$3.9x10^3$	$4.0x10^3$	$1.9x10^4$	$1.2x10^4$	$1.1x10^4$	$3.4x10^3$	$3.4x10^3$	$2.5x10^2$	$1.1x10^3$	$5.9x10^4$
Single-spike	1 kΩ-10 kΩ	$1.2x10^4$	$1.2x10^4$	$3.4x10^1$	$2.2x10^1$	$1.9x10^1$	$2.3x10^1$	$8.3x10^1$	4.3	$2.3x10^1$	$2.5x10^4$
	10 kΩ-100 kΩ	$1.3x10^3$	$1.3x10^3$	$3.4x10^1$	$2.2x10^1$	$1.9x10^1$	$2.5x10^1$	$9.5x10^1$	4.9	$2.6x10^1$	$2.9x10^3$
	100 kΩ-1 MΩ	$1.3x10^2$	$1.3x10^2$	$3.4x10^1$	$2.2x10^1$	$1.9x10^1$	$2.7x10^1$	$1.0x10^2$	5.3	$2.8x10^1$	$4.9x10^2$

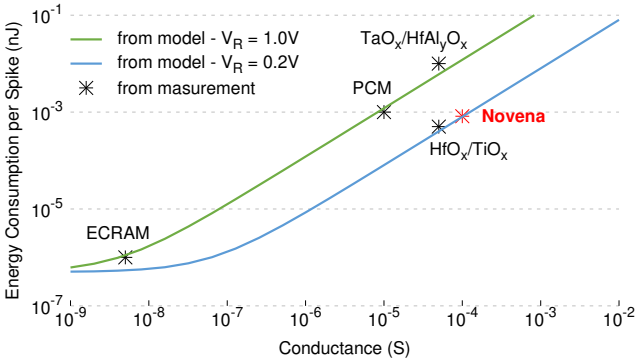

Figure 5: Comparison of data from analytical models in the NCPower and data from fabricated devices

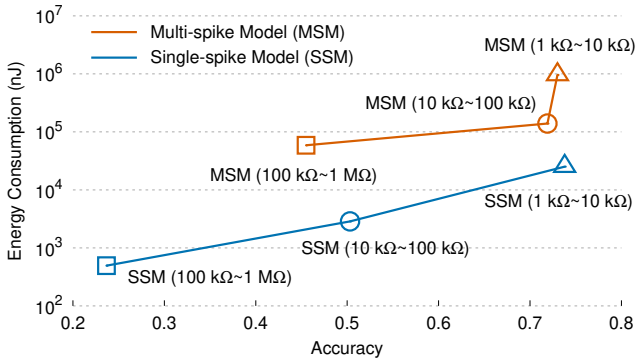

Figure 6: Energy consumption and accuracy of the multi-spike model (MSM) and the single-spike model (SSM)

the energy consumption of ADC, which is defined as the supply current of ADC when the input voltage signal is $1/2 \cdot V_{ADC}$.

Each ADC is connected to a b-bit register, which stores the output data of ADC. In the NVM-based hard core, the spikes from neurons in layer L are represented by two complementary currents: one for positive weights and one for negative weights. Therefore, we need to find the difference between these two currents by a digital arithmetic unit. Once the ADCs for these two complementary currents finish the converting process and store the results into registers, a subtractor would calculate the difference of these two values. The result is then added to the membrane potential u_m, which is stored in a third b-bit register. Finally, to emulate leakage, the potential is divided or subtracted by a constant using shift add/subtract operations. The result is written back to the register.

$$E_{REG} = 3M \cdot (e^r_{rd} + e^r_{wr} + P^r_s(t_0 - t_{acq})) \cdot b \quad (15)$$

The energy consumption of registers is expressed in Equation 15. e^r_{rd} and e^r_{wr} are the read and write energy consumption per bit of the register. P^r_s is the static power consumption of the register.

$$E_{ARITH} = 2M \cdot e_{add} \cdot b + M \cdot e_{sft} \cdot \Delta b \quad (16)$$

The energy consumption of the arithmetic unit is expressed in Equation 16. e_{add} and e_{sft} are the energy consumption per bit of the add operation and shift operation, respectively.

3 Analysis and Comparison

We integrate the estimator in SNN simulator, and analyze the accuracy and energy consumption of two SNN models: the traditional multi-spike based SNN and the new single-spike based SNN[4]. Traditionally, the multi-spike model converts each pixel of the input image into multiple spikes. The value of the pixel is represented by the frequency of the spikes. Different from the traditional multi-spike model, the single-spike model first converts the input image into a tensor by a convolutional layer. Afterwards, each element in the tensor is converted into a spike if the value of that element is greater than 0. Compared with the multi-spike model, the single-spike model usually takes less cycles in inference because it interprets the input images more efficiently. NCPower is accurate for both models because the basic working principles of these two models on the neuromorphic chip are equivalent.

3.1 Validation of NCPower

We collect parameters from published paper. The parameters related to analog circuits are listed in Table 1. Typically, the successive approximation (SAR) ADC is used. The listed parameters of NVM resistors are based on HfOx NVM, which are fabricated as 1T1R cell arrays. The parameters of digital circuits and metal interconnects are collected from the ITRS report [16]. We verified NCPower by comparing the results generated from our analytical models with the experiment results of the real NVM-based neuromorphic chip. One of the chip is our in-house developed chip called Novena (shown in Figure 4), where we measured the data directly. The data of other neuromorphic chips using different NVM devices are collected from publish paper, including ECRAM [13], PCM [2], $TaO_x/HfAl_yO_x$ [18] and HfO_x/TiO_x [19]. The comparison on energy consumption per spike is shown in Figure 5. From the figure we can see that our analytical model could well match experiment results, in both low and high conductance ranges. In average, NCPower has ±23% error on energy estimation. One reason of such error is that we do not have enough detailed configurations of these devices. Their exact architectures may have little variations. In this experiments, we only input the basic configurations they shown into the estimator.

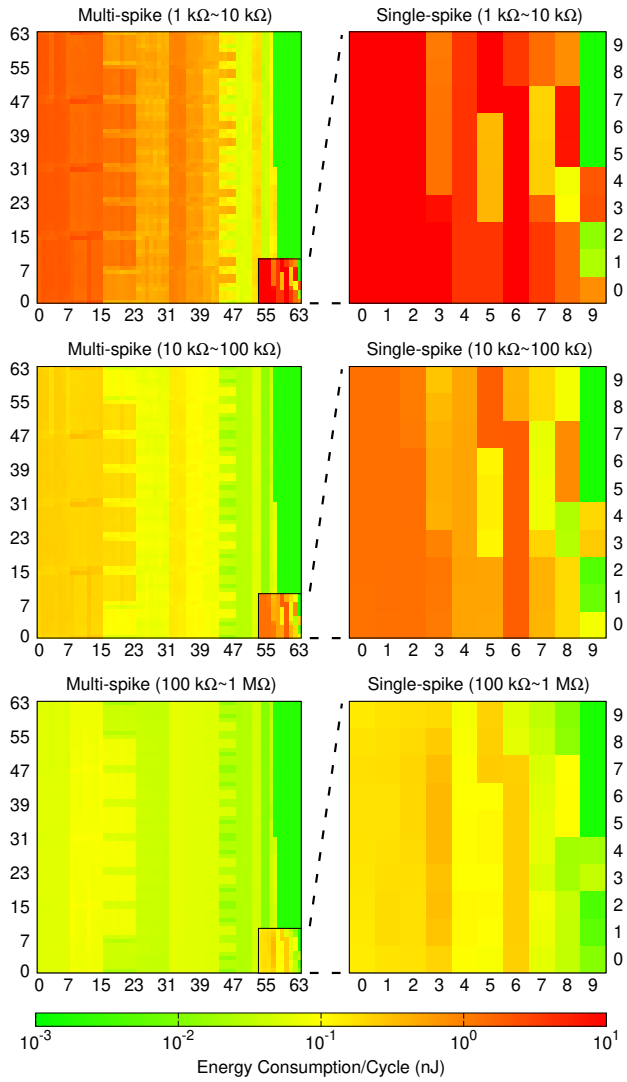

Figure 7: The energy consumption per cycle on different cores. The chip for the multi-spike model have more cores than the single-spike model. For fair comparison, the result of the single-spike model is plotted in its actual size, at the right bottom corner of the multi-spike model

3.2 Trade-off on Resistance Range

NCPower can help us to make choice on the resistance range of the NVM resistors. Figure 6 shows the energy consumption and the accuracy of the multi-spike model and the single-spike model, in three different resistance ranges. In this figure we have two observations. First, the accuracy of the multi-spike model at resistance range 1 kΩ-10 kΩ is similar to the accuracy at resistance range 10 kΩ-100 kΩ, but their energy consumption is different. We can save the energy consumption by one order of magnitude if we choose the larger resistance range. Second, the accuracy of the single-spike model with resistance range 1 kΩ-10 kΩ is comparable to the multi-spike model with resistance range 1 kΩ-10 kΩ and 10 kΩ-100 kΩ. If we choose the single-spike model instead of the multi-spike model,

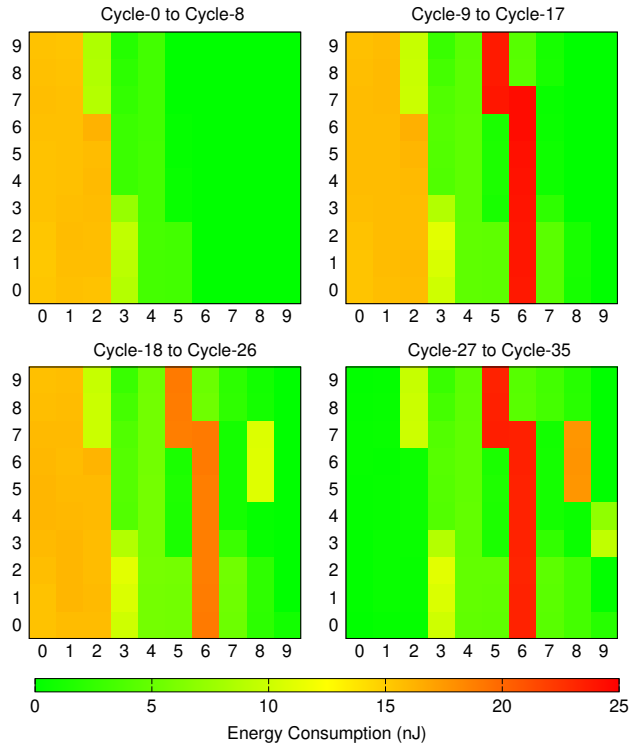

Figure 8: The accumulated energy consumption of different cores on the neuromorphic chip. The total inference time is divided into four quarters, and each sub-figure shows the accumulated energy over a quarter. The single-spike model is tested, with resistance range 1 kΩ-10 kΩ

the energy consumption can be saved by around two orders of magnitude. We can also use NCPower to further analyze the breakdown of the energy consumption. Table 2 shows the details. At resistance range 1 kΩ-10 kΩ and 10 kΩ-100 kΩ, around 87% of the energy consumption is on the NVM resistors and TIA. However, at resistance range 100 kΩ-1 MΩ, the energy consumption of NVM resistors and TIA are only 33% of the total energy consumption, in average of the two models. This is because with the increasing of resistance, the energy consumption of NVM and TIA decrease substantially, while that of the other components does not change a lot. This table explains why the increment of NVM resistance could effectively reduce the energy consumption of the chip.

3.3 Spatial Analysis

We can use NCPower to analyze neuromorphic chip's energy consumption on the space domain. Figure 7 shows the energy consumption per cycle on each core. We compare the multi-spike model and the single spike model, on three different resistance ranges of the NVM resistors. From the figure we can see that with the increasing of resistance range, the energy consumption in both cases decrease. In average, the energy consumption per cycle on each core of the multi-spike model is 79% less than that of the single-spike model. However, given the fact that the multi-spike model uses 3680 cores and the single-spike model uses only 95 cores, the total energy consumption of the single-spike model is much less than that of the

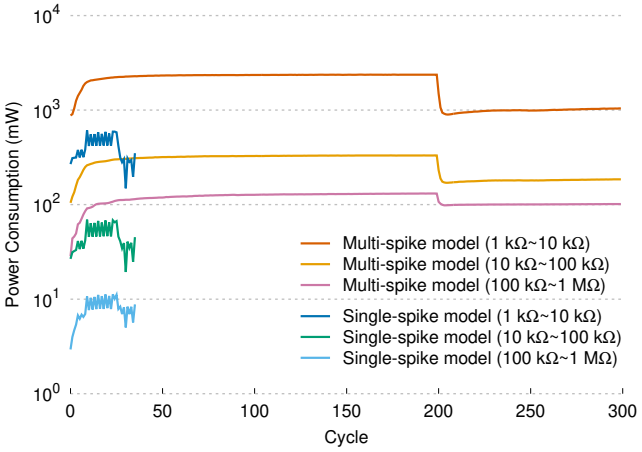

Figure 9: The power consumption of the neuromorphic chip over time when it is inferring a single image

multi-spike model. Cores do not consume equal amount of energy, and some cores are overheated. For example, in the multi-spike model, cores in the left part of the chip consume most of the energy. These cores are mapped to neurons in the first few layers of the SNN model. Usually, it takes several cycles of operation until neuron on those layers can accumulate enough potential and fire. Such property makes the neurons at the first few layers tend to consume more energy consumption than neurons in the other layers, resulting in hot spots on the chip. This phenomenon become acute if the soft cores in SNN is not properly mapped to hard cores. The mapping strategy can be optimized to avoid this problem, and NCPower can be a useful tool during the optimization process.

3.4 Temporal Analysis

We can also use NCPower to analyze neuromorphic chip's energy consumption on the time domain. Figure 7 shows the energy consumption of the single-spike model when it is inferring one image. We divide the total inference time into four quarters. From the figure we can observe the movement of hot sport from one part of the chip to the other part. In the first quarter, around 82% of the energy consumption is consumed on the first two columns of cores. In the later three quarters, eleven cores at column 5 and column 6 consume around 33% of the energy consumption. This is because the number of weights in those eleven cores is almost half of the total number of weights in the model. Figure 9 shows the variation of power consumption of the neuromorphic chip when it is inferring one image. We compare the multi-spike models and the single-spike model, in three different resistance ranges of the NVM resistors. From the figure we can see that with the increasing of resistance range, the power consumption in both cases decrease. In average, the power consumption of the single-spike model is one order of magnitude less than that of the multi-spike mode, and the single-spike model has 7X energy efficiency over the multi-spike model. In either case, we can observe a drop of power consumption at a certain point of the inference time. This is because at that point, the whole image is transmitted into the SNN model and there are no incoming spikes from the input image afterwards. Hence, neurons at the first few layers of SNN models do not consume energy any more. If the neuromorphic chip is inferring multiple images at the same time, the chip may consume tremendous amount of energy at a certain cycle, and hit the power wall of the chip. The scheduling policy can be optimized to avoid this problem, and NCPower can be a useful tool during the optimization process.

4 Conclusions

In this paper, we propose NCPower, a power estimator for neuromorphic chip. We verify it on different NVM based neuromorphic chips. The analytical result from NCPower could well match experiment results. We integrate it into a SNN simulator, and analyze the traditional multi-spike SNN and the new single-spike SNN. From NCPower we can see that the single-spike model has 7X energy efficiency over the multi-spike model, while keeping similar accuracy under the CIFAR-10 dataset. Using NCPower, developers can check if a certain location of the chip is over-heated, and try to alleviate the thermal issues by optimizing soft core and hard core mapping strategy. Developers can also check if the power consumption would exceed the power limit at certain time, and try to avoid the power wall of the chip by optimizing the scheduling policy.

Acknowledgement

This work was supported by the Singapore Government's Research, Innovation and Enterprise 2020 Plan (Advanced Manufacturing and Engineering domain) under Grant A1687b0033.

References

[1] N. Agarwal et al. 2009. GARNET: A detailed on-chip network model inside a full-system simulator. In *2009 ISPASS*. 33–42.
[2] Irem Boybat et al. 2018. Neuromorphic Computing with Multi-memristive Synapses. *Nature communications* 9, 1 (2018), 1–12.
[3] X. Dong et al. 2012. NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory. *IEEE TCAD* 31, 7 (2012), 994–1007.
[4] SK Esser et al. 2016. Convolutional networks for fast, energy-efficient neuromorphic computing. 2016. *arXiv preprint* 27 (2016).
[5] A. Grossi et al. 2016. Performance and Reliability Comparison of 1T-1R RRAM Arrays with Amorphous and Polycrystalline HfO2. In *2016 EUROSOI-ULIS*. 80–83.
[6] Daniele Ielmini et al. 2019. Emerging neuromorphic devices. *Nanotechnology* 31, 9 (2019), 092001.
[7] Sheng Li et al. 2009. McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures. In *2009 MICRO*. 469–480.
[8] Tao Luo et al. 2016. A racetrack memory based in-memory booth multiplier for cryptography application. In *2016 ASP-DAC*. IEEE, 286–291.
[9] Tao Luo et al. 2017. A novel two-stage modular multiplier based on racetrack memory for asymmetric cryptography. In *2017 ICCAD*. IEEE, 276–282.
[10] R. Salkhordeh et al. 2019. An Analytical Model for Performance and Lifetime Estimation of Hybrid DRAM-NVM Main Memories. *IEEE Trans. Comput.* 68, 8 (2019), 1114–1130.
[11] Catherine D Schuman et al. 2017. A survey of neuromorphic computing and neural networks in hardware. *arXiv preprint* (2017).
[12] Evangelos Stromatias et al. 2013. Power analysis of large-scale, real-time neural networks on SpiNNaker. In *The 2013 IJCNN*. IEEE, 1–8.
[13] Jianshi Tang et al. 2018. ECRAM as Scalable Synaptic Cell for High-speed, Low-power Neuromorphic Computing. In *2018 IEDM*. IEEE, 13–1.
[14] Yu Wang et al. 2015. Energy efficient RRAM spiking neural network for real time classification. In *GLSVLSI*. 189–194.
[15] Wujie Wen et al. 2016. A holistic tri-region MLC STT-RAM design with combined performance, energy, and reliability optimizations. In *DATE*. IEEE, 1285–1290.
[16] Linda Wilson. 2013. ITRS Report. *Semiconductor Industry Association* 1 (2013).
[17] Chris Yakopcic et al. 2015. SPICE analysis of dense memristor crossbars for low power neuromorphic processor designs. In *2015 NAECON*. IEEE, 305–311.
[18] Peng Yao et al. 2017. Face Classification Using Electronic Synapses. *Nature communications* 8, 1 (2017), 1–8.
[19] Shimeng Yu et al. 2013. A Low Energy Oxide-based Electronic Synaptic Device for Neuromorphic Visual Systems with Tolerance to Device Variation. *Advanced Materials* 25, 12 (2013), 1774–1779.
[20] D. Zhang et al. 2012. A 53-nW 9.1-ENOB 1-kS/s SAR ADC in 0.13-μm CMOS for Medical Implant Devices. *IEEE JSSC* 47, 7 (July 2012), 1585–1593.