Enabling An New Era of Brain-inspired Computing: Energy-efficient Spiking Neural Network with Ring Topology

Kangjun Bai
Virginia Tech
Blacksburg, Virginia, USA
kangjun@vt.edu

Jialing Li
Virginia Tech
Blacksburg, Virginia, USA
ljl@vt.edu

Kian Hamedani
Virginia Tech
Blacksburg, Virginia, USA
hkian@vt.edu

Yang Yi
Virginia Tech
Blacksburg, Virginia, USA
yangyi8@vt.edu

ABSTRACT

The reservoir computing, an emerging computing paradigm, has proven its benefit to multifarious applications. In this work, we successfully designed and fabricated an analog delayed feedback reservoir (DFR) chip. Measurement results demonstrate its rich dynamic behaviors and high energy efficiency. System performance, as well as the robustness, are evaluated. The application of video frame recognition is investigated using a hybrid neural network, which employs the multilayer perceptron (MLP) training model as the readout layer of our designed DFR system, and yields 98% classification accuracy. Compared to results of using the MLP training only, our hybrid training model exhibits much higher recognition rate and accuracy.

CCS CONCEPTS

• **Computer systems organization** → Architectures → Neural networks; **Hardware** → Integrated circuits

KEYWORDS

Spiking neuromorphic computing, delayed feedback reservoir, edge of chaos, video frame recognition

1 INTRODUCTION

The continued success of the development of the modern von Neumann computing system was firstly enabled by the increment of the transistor integration density, followed by the multi-core computing architecture. However, as the density of data continuously escalates, extracting valuable information becomes computationally expensive, even for supercomputers.

Inspired by the working mechanism of mammalian brains, neuromorphic computing has become one of the computing paradigms that is able to break through the bottleneck of traditional

computing systems. As powerful as mammalian brains, neuromorphic computers can efficiently solve complex tasks such as analyzing and classifying surrounded information. For instance, TrueNorth, the latest brain-inspired chip prototyped by IBM, classifies multiple objects within 240x400-pixel 3-color video input at 30-frame-per-second with merely 65mW of energy consumption, which is five orders of magnitude more efficient than von Neumann computers [1, 2].

The reservoir computing, introduced by Jaeger [3] and Maass [4] in the early 2000s, is with processing capabilities rely on the dynamical behavior of conventional recurrent neural networks (RNNs). In past decades, reservoir computing has been fully developed to reduce the training complexity of RNNs [5, 6]. Performance metrics of the reservoir computing outperform conventional RNNs in the nonlinear system identification, prediction, and classification. However, the development of hardware implementations in reservoir computing is hindered by the enormous amount of nonlinear neurons.

A key finding emerged that the best computational capability achieved by the system was found when the dynamic is in the transition regime of order and chaotic, namely the edge of chaos [7]. The computational performance in ordered, edge of chaos, and chaotic regimes are evaluated in [8, 9]. Results suggest that a system should stay close to critical dynamics (edge of chaos) leading to advanced information processing capabilities. This hypothesis has proven to be true through biological experiments showing our brain is indeed operating at the edge of chaos [10]. Meanwhile, the delay feature can be used to mimic the delay inherent in neurons making the sigmoid as a more suitable candidate for activation function in artificial neural networks. The edge of chaos and delay characteristics in the sigmoid activation function significantly improve the computing performance in targeted systems. These findings led researchers to merge the time delay property into the reservoir computing.

In this work, we present an analog delayed feedback reservoir (DFR) system that built on the edge of chaos with delay characteristics in the sigmoid activation function. Contributions of our work include the following:

- 1) The designed analog DFR chip has been successfully fabricated and tested, with careful consideration of its robustness and noise immunity;
- 2) Measurement results demonstrate rich dynamic behaviors and robustness of our fabricated analog DFR chip, which closely mimics the efficient neurological signal transmission of our mammalian brains;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

DAC '18, June 24–29, 2018, San Francisco, CA, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5700-5/18/06...\$15.00
<https://doi.org/10.1145/3195970.3196044>

- 3) The application of video frame recognition, with 48 images corresponding to different training and testing subjects from the Head Pose Image database [11], is employed for the performance evaluation;
- 4) Experimental results in video frame recognition illustrate its high classification accuracy by using the back-propagation training algorithm through a hybrid training model.

2 DESIGN ASPECTS

In mammalian brains, the cerebral cortex is the key region that involves in processing memory, perception, languages, etc [12]. As is well known, the conduction time between neurons accounts for the delay in biological systems. With the everlasting enthusiasm of neuroscience, it was found that the cerebral cortex is operating at the edge of chaos regime due to the existence of the delay [13]. The concept of computation at the edge of chaos was reported in [7] as early as 1990. The hypothesis was first proposed in the field of neuroscience, by a group of researchers from the Cambridge University, claiming that our brain is operating at the edge of chaos regime. From the mathematical point of view, the dynamic of a system can be controlled by varying one or several parameters [14-16]. This led us to think how the dynamic is varied in mammalian brains.

In the reservoir computing, both high dimensionality and short-term memory are achieved. With its capability of mapping information onto higher dimensional spaces, the reservoir computing system facilitates the process of linear separation, as illustrated in Fig. 1.

Figure 1: Classification with (a) nonlinear separation in low dimensional spaces; (b) linear separation in high dimensional spaces.

From the chaos theory, the Lyapunov analysis has demonstrated that the high dimensionality could be achieved with a simple delay [17, 18]. It has been proven that the best performance of reservoir computing is found when operating at the edge of chaos regime [14, 19]. The same conclusion has drawn for spiking neural circuit models [20]. In the reservoir computing, the dynamic behavior varies accordingly as the delay starts to increase. In [21], phase portraits are used to visualize the transition of intrinsic dynamic behavior from a theoretical point of view. From the hardware realization of the time delay reservoir, the result from phase portrait has also shown that the time delay reservoir could operate at the edge of chaos [22, 23].

3 SYSTEM INTEGRATION AND RESULTS DISCUSSION

The analog DFR system has been designed and successfully fabricated through the standard Global Foundries 130nm CMOS technology. In this section, the detailed system-level integration as well as evaluations of performance and robustness are demonstrated.

3.1 Overview of Analog DFR System

Fig. 2 illustrates an overview of our designed analog DFR system that is composed of a temporal encoding scheme[24, 25], a sigmoid nonlinear activation function, and a dynamic delayed feedback loop with embedded delay calibration module.

Figure 2: Computing architecture of analog DFR system.

As demonstrated in Fig. 2, the analog DFR system contains an input layer and a reservoir layer. Analog input signals are first processed through a signal regulator, followed by leaky integrate-and-fire (LIF) neurons and a signal integrator to convert analog input signals into digitized spike trains. The nonlinear activation function within the reservoir layer experiences a nonlinear transformation of processed input signals, such that the input can be linearly separated for classification, as depicted in Fig. 1(b). Nonlinearly transformed spike trains travel along the dynamic delayed feedback loop at a specific time to mimic the recurrent connection, which represents the time-shifted nonlinear signal, to enable the temporal information processing.

The dynamic behavior of the analog DFR system can be controlled by altering the total delay time within the system. Such property is realized by an integrate-and-fire (IF) delay neuron. The delay time constant, τ , can be regulated by the equalized

calibration current through each IF delay neuron, which is controlled by the delay calibration module with the operational trans-conductance amplifier (OTA) within the reservoir layer.

To minimize the information lost during the signal processing, the analog signal is recovered from the spike train before injecting into the gain controller. To create the persistent dynamic memory, a feedback loop is employed to integrate the feedback signal from the previous state with the new incoming input, as depicted in Fig. 3, such that the new incoming data carries the information from its previous state(s). The expression of N^{th} output, $S_{(i)}$, could be simplified as:

$$S_{(i)} = f_n \left[I_{(i)} + \sum_{1}^N I_{(i-1)} \cdot Av^{(N-N_{(i)})} \right] \quad (1)$$

where the function, $f_n[\]$, represents the nonlinear transformation of input signals; $I_{(i)}$ is the N^{th} input data; Av is the finite gain of the gain controller; N and $N_{(i)}$ define the total number of I/O terms and the N^{th} I/O term, respectively.

Figure 3: Simplified block diagram of persistent dynamic memory integration.

Within the readout layer, weights are then being trained by the backpropagation training algorithm through the hybrid neural network, which employs the multilayer perceptron (MLP) training model as the readout layer of our designed analog DFR system, with the processed information from the reservoir layer.

3.2 Preliminary Performance Evaluation

The die photo of the fabricated analog DFR chip is demonstrated in Fig. 4. The overall design area of the analog DFR chip occupies $1.5\text{mm} \times 1.5\text{mm}$, while each analog DFR system module takes up to $180\mu\text{m} \times 60\mu\text{m}$.

Figure 4: Die photo of fabricated analog DFR chip.

3.3 Results Discussion

3.3.1 Nonlinear Activation Function

The designed sigmoid nonlinear activation function could be expressed as:

$$f(x) = \frac{\alpha}{1 + e^{n(x-\beta)}} \quad (2)$$

where α is the scaling factor, which controls the amplitude of output signals; n is the nonlinear exponent, and β determines the time-shift of input signals. By varying the input signal from GND to V_{DD} , the experimental characteristic of the sigmoid nonlinear activation function is plotted in Fig. 5. In this experiment, the function, $f(x)$, fits the experimental nonlinearity with a nonlinear exponent of $n = 12$.

Figure 5: Characteristic of nonlinear activation function.

In the hardware realization, the robustness of the nonlinear transformation could be enhanced by increasing the nonlinearity of the activation function, which is determined by its intrinsic time coefficient. Increasing the intrinsic time coefficient often results in a lower computation efficiency. However, to facilitate the high-precision computation, such nonlinear transformation system is designed with a large intrinsic time coefficient and operated with a lower sampling frequency.

Figure 6: Error rate of nonlinear transformation under different sampling frequency.

As depicted in Fig. 6, the average error rate of the nonlinear transformation maintains below 1% if the sampling frequency is less than 5MHz. As the sampling frequency increases, the average error rate increases up to 3.5%.

3.3.2 Intrinsic Delay Property

In general, the delay time constant, τ , of the IF delay neuron could be expressed as:

$$\tau = C_{mem} \cdot \frac{V_{th}}{I_{cal}} \quad (3)$$

where C_{mem} is the membrane capacitance of the sensing capacitor in the IF delay neuron; V_{th} defines the firing potential of the IF delay neuron, and I_{cal} is the calibration current. The dynamic range of the delay time is characterized by the calibration current. As demonstrated in Fig. 7, the IF delay neuron has a superior dynamic range of delay times from 0.2 μ s to 1.4 μ s.

Figure 7: (a) Intrinsic delay calibration; (b) demonstration of analog spiking signal.

3.3.3 Dynamic Behavior

In this experiment, phase portraits are used to visualize the intrinsic dynamic behavior of our designed analog DFR system. As depicted in Fig. 8, it can be observed that our fabricated analog DFR chip indeed goes through a range of dynamic behaviors. Results confirm that our fabricated analog DFR chip successfully implemented the desired functionality.

As demonstrated in Fig. 8(b), the delayed signal, $I(t - \tau)$, repeatedly traces its initial path even in a long run, which results as periodic. As the delay increases, demonstrated in Fig. 8(d), the dynamic of the system exhibits an edge of chaos behavior, such that the delayed signal diverges its initial path without off tracking from the equilibrium point even in a long run. With testing results on dynamic behaviors, it is reasonable to conclude that our fabricated analog DFR chip has realized both delay and richness of dynamic behavior with merely 526 μ W of power consumption. This indicates that the designed analog DFR system closely mimics biological neurons of mammalian brains with the intrinsic conduction time delay.

Figure 8: Intrinsic dynamic behaviors of fabricated analog DFR chip with (a) $T = 0.64\mu s$; (b) $T = 1\mu s$; (c) $T = 1.2\mu s$; (d) $T = 1.4\mu s$.

3.3.3 System Robustness and Normalized Error

With 20 sampling points from the measurement, normalized errors, E_n , of the nonlinear activation function and the IF delay neuron are evaluated through:

$$E_n = \frac{|V_m - V_{pre}|}{\sqrt{U_m^2 + U_{pre}^2}} \quad (4)$$

$$E_n \in \left\{ E_{n(min)} : \frac{E_{n(max)} - E_{n(min)}}{Num} : E_{n(max)} \right\} \quad (5)$$

where V_m and V_{pre} represent actual measured and predicted values, respectively; U_m and U_{pre} are uncertainties of measured and predicted values, respectively. Normalized errors are then categorized based on the number of sampling points within each particular range of error. As illustrated in Fig. 9, average normalized errors of the nonlinear activation function and the IF delay neuron are 1.3% and 2.7%, which indicate that the average signal offset is merely $\pm 5.8mV$ and $\pm 6.2ns$, respectively.

Figure 9: Normalized error in (a) nonlinear activation function; (b) dynamic delay regulation.

4 APPLICATION EVALUATION

4.1 Experimental Setup

In this experiment, we evaluate the performance and reliability of our designed analog DFR system by using the application of video frame recognition. The hybrid neural network is applied by using the MLP training model as the readout layer of our designed analog DFR system. Weights are then trained by the backpropagation training algorithm. Since the training time increases dramatically with the design scale, the hidden layer of the MLP training model is built with two neuron layers that contain 80 and 40 neurons, respectively.

Figure 10: (a) Standard training database with three subjects; (b) down-sampled testing dataset with various salt-and-pepper noise levels.

Six images corresponding to different training subjects from the Head Pose Image database [11] are demonstrated in Fig. 10(a). Total of 48 images, which contain three different subjects with multiple rotation angles, are drawn. Among those images, 24 are used during the training stage, while the rest of images are used for the testing phase. For the training dataset, the horizontal angle of face alters from 0° to 75° with an increment of 15° per rotation, while the vertical angle of face maintains at 0° . In the testing dataset, the alteration of horizontal angles of face follow the training dataset but with additional 15° applied to the vertical angle of face. The reliability of the system is investigated by introducing various levels of salt-and-pepper noise to the down-sampled testing dataset, as depicted in Fig 10(b).

Figure 11: Hybrid training model for application evaluation.

Fig. 11 illustrates the test bench of our hybrid training model. During the operation, each input image is firstly down-sampled to 64×64 pixels, and each pixel of the image is mapped onto higher dimensional spaces for linear separation by the designed analog DFR system. As its output, a total of 4096 spike trains, which represent the information of each pixel of an image, are generated from the reservoir layer. These spike data patterns are then recorded and converted into analog signals, x_n . Within the hidden layer of the MLP training model, weighted inputs are mapped to each neuron in the following layer. Initially, weights are generated randomly. During the training stage, as each piece of data is processed, weights are calibrated based on the corrections that minimizing the error between actual and target outputs. The output layer, which contains three neurons, represents 000 to 111 in a digital format such that the system can classify input patterns up to eight different categories. In order to prevent overfitting, the cross-validation technique is applied during the training process.

4.2 Performance Analysis

The recognition rate is evaluated against different levels of salt-and-pepper noise with two different training models: (1) hybrid and (2) MLP-only.

Figure 12: Recognition rate with respect to various levels of salt-to-pepper noise.

As depicted in Fig. 12, the recognition rate maintains at 98% if the salt-and-pepper noise level is less than 10% by using the hybrid training model. The same trend can be found with the MLP-only training model, which is at a constant recognition rate of 78%. At low noise level, our hybrid training model exhibits much higher recognition rate than that of MLP-only training model. This trend can also be found at higher noise levels. As the noise level reaches 50%, the recognition rate of the hybrid training model is decreased to 93%. However, with the MLP-only training model at 50% of noise level, the recognition rate has dropped to 67%, which is 26% more than that of the hybrid training model. Hence, the hybrid training model is more robust against noise than the MLP-only training model does.

Figure 13: Recognition rate with respect to various dynamic behavior under different noise levels.

The delay plays an important role in the system. As discussed in section 3.3, our fabricated analog DFR chip has rich dynamic behaviors. In this task, the recognition rate is evaluated with various delay time. As demonstrated in Fig. 13, it can be observed that the recognition rate varies with respect to dynamic behaviors. When the designed analog DFR system operates at the edge of chaos regime ($T = 20ms$), the recognition rate is maintained at 98% with a noise level below 10%. As the noise level approaches to 50%, the recognition rate at the edge of chaos regime maintains above 93%. However, if the delay deviates from 20ms, the recognition rate is drastically affected by the change in the dynamic behavior of the system, which results in approximately 25% worse in the performance.

5 CONCLUSIONS

In this paper, we demonstrate the functionality of our designed DFR system, along with measurement results and the application on video frame recognition. Mimicking how brain processes information, the designed analog DFR system is embedded with a temporal encoder to achieve spiking based information processing. The designed analog DFR system has successfully fabricated with the Global Foundries 130nm CMOS technology, where high energy efficiency and small design area are accomplished. From measurement results, the fabricated analog DFR system exhibits its rich dynamic behaviors, which indicates the successful implementation of the proposed delayed feedback architecture that mimics biological neurons with merely 526μW of power consumption. The performance of the designed DFR system is also evaluated through the application on video frame recognition whereby hybrid training model is employed. The testing accuracy is found to be as high as 98%.

ACKNOWLEDGMENTS

This material is based upon work funded by Air Force Research Lab (AFRL), under AFRL Grant No. FA8750-16-2-0120. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of AFRL.

REFERENCES

- [1] Modha, D.S., *Introducing a brain-inspired computer*. Published online at <http://www.research.ibm.com/articles/brain-chip.shtml>, 2017.
- [2] Merolla, P.A., et al., *A million spiking-neuron integrated circuit with a scalable communication network and interface*. Science, 2014. **345**(6197): p. 668-673.
- [3] Jaeger, H., *The "echo state" approach to analysing and training recurrent neural networks-with an erratum note*. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, 2001. **148**(34): p. 13.
- [4] Maass, W., T. Natschlager, and H. Markram, *Real-time computing without stable states: A new framework for neural computation based on perturbations*. Neural computation, 2002. **14**(11): p. 2531-2560.
- [5] Schrauwen, B., D. Verstraeten, and J. Van Campenhout. *An overview of reservoir computing: theory, applications and implementations*. in *Proceedings of the 15th European Symposium on Artificial Neural Networks*. p. 471-482 2007. 2007.
- [6] Lukoševičius, M. and H. Jaeger, *Reservoir computing approaches to recurrent neural network training*. Computer Science Review, 2009. **3**(3): p. 127-149.
- [7] Langton, C.G., *Computation at the edge of chaos: phase transitions and emergent computation*. Physica D: Nonlinear Phenomena, 1990. **42**(1-3): p. 12-37.
- [8] Bertschinger, N. and T. Natschlager, *Real-time computation at the edge of chaos in recurrent neural networks*. Neural computation, 2004. **16**(7): p. 1413-1436.
- [9] Li, J., et al. *Analog hardware implementation of spike-based delayed feedback reservoir computing system*. in *Neural Networks (IJCNN), 2017 International Joint Conference on*. 2017. IEEE.
- [10] Kitzbichler, M.G., et al., *Broadband criticality of human brain network synchronization*. PLoS computational biology, 2009. **5**(3): p. e1000314.
- [11] Gourier, N., D. Hall, and J.L. Crowley. *Estimating face orientation from robust detection of salient facial structures*. in *FG Net Workshop on Visual Observation of Deictic Gestures*. 2004.
- [12] Kandel, E.R., et al., *Principles of neural science*. Vol. 4. 2000: McGraw-hill New York.
- [13] Boedecker, J., et al., *Information processing in echo state networks at the edge of chaos*. Theory in Biosciences, 2012. **131**(3): p. 205-213.
- [14] Legenstein, R. and W. Maass, *Edge of chaos and prediction of computational performance for neural circuit models*. Neural Networks, 2007. **20**(3): p. 323-334.
- [15] Li, J., et al., *Enabling Sustainable Cyber Physical Security Systems Through Neuromorphic Computing*. IEEE Transactions on Sustainable Computing, 2017.
- [16] Zhao, C., et al. *Novel spike based reservoir node design with high performance spike delay loop*. in *Proceedings of the 3rd ACM International Conference on Nanoscale Computing and Communication*. 2016. ACM.
- [17] Ikeda, K. and K. Matsumoto, *High-dimensional chaotic behavior in systems with time-delayed feedback*. Physica D: Nonlinear Phenomena, 1987. **29**(1-2): p. 223-235.
- [18] Hegger, R., et al., *Identifying and modeling delay feedback systems*. Physical review letters, 1998. **81**(3): p. 558.
- [19] Legenstein, R. and W. Maass, *What makes a dynamical system computationally powerful*. New directions in statistical signal processing: From systems to brain, 2007: p. 127-154.
- [20] Mayor, J. and W. Gerstner, *Signal buffering in random networks of spiking neurons: Microscopic versus macroscopic phenomena*. Physical Review E, 2005. **72**(5): p. 051906.
- [21] Junges, L. and J.A. Gallas, *Intricate routes to chaos in the Mackey-Glass delayed feedback system*. Physics letters A, 2012. **376**(30): p. 2109-2116.
- [22] Li, J., et al. *A Deep Learning Based Approach for Analog Hardware Implementation of Delayed Feedback Reservoir Computing System*. in *Proceedings of 19th International Symposium In Quality Electronic Design (ISQED)*. 2018. IEEE.
- [23] Bai, K. and Y. Yi. *A Path to Energy-efficient Spiking Delayed Feedback Reservoir Computing System for Brain-inspired Neuromorphic Processors*. in *Proceedings of 19th International Symposium In Quality Electronic Design (ISQED)*. 2018. IEEE.
- [24] Zhao, C., et al., *Spike-time-dependent encoding for neuromorphic processors*. ACM Journal on Emerging Technologies in Computing Systems (JETC), 2015. **12**(3): p. 23.
- [25] Zhao, C., J. Li, and Y. Yi. *Making neural encoding robust and energy efficient: an advanced analog temporal encoder for brain-inspired computing systems*. in *Proceedings of the 35th International Conference on Computer-Aided Design*. 2016. ACM.