Research Statement

Yingyan Lin, http://yingyan.web.engr.illinois.edu

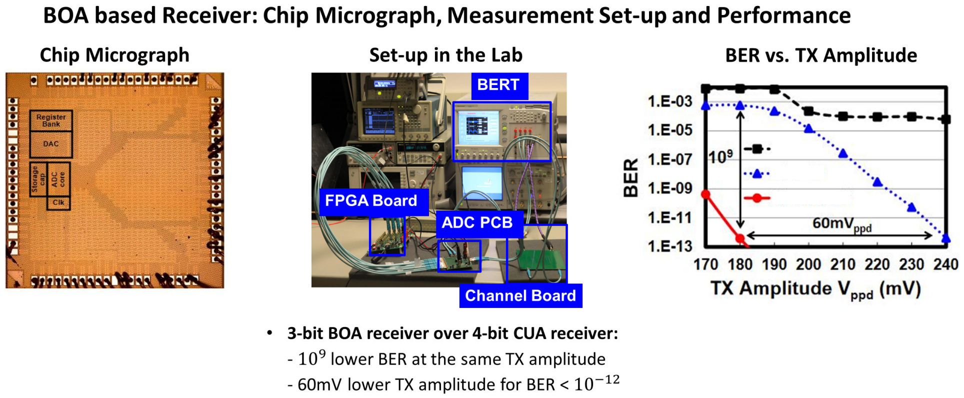
Machine learning (ML) algorithms are finding excellent utility in tackling the data deluge of the 21st Century at a large energy cost. Therefore, it is imperative to design energy-efficient ML systems to harness their full benefits. I have been passionate about addressing this fundamental problem, which involves diverse areas including devices, circuits, VLSI systems and architectures, and ML algorithms. Specifically, my PhD work has contributed to Systems on Nanoscale Information fabriCs (SONIC), a $30 million research center led by my PhD advisor Prof. Naresh R. Shanbhag and sponsored by SRC and DARPA.

Current Research

My PhD research focuses on the design of energy-efficient systems for information processing and transfer. Current ML systems are either centralized in a cloud, or distributed at the edge. In both platforms, there is a grand energy efficiency challenge as described next. In my PhD research, I investigate techniques to address this challenge.

Energy Efficiency Challenge in the Data Center: Data transfer due to inter-chip, inter-board, inter-shelf and inter-rack communications within data centers is one of the dominant energy costs in data centers. To address the energy efficiency challenge in data centers, I focus on reducing the energy of the I/O interface. Specifically, I am interested in analog-to-digital converter (ADC)-based multi-Gb/s serial link receivers, where the power dissipation is dominated by the ADC. In [1], I present an investigation on the use of link BER for designing a BER-optimal ADC (BOA) based serial link. Measured results (see Fig. 1) showed that a 3-bit BOA receiver outperforms a 4-bit CUA receiver at a BER $< 10^{-12}$, thereby supporting the claims in [2].

This work [1] was presented at IEEE International Solid-State Circuits Conference Student Research Preview in 2015 (*ISSCC SRP 2015*) and published at IEEE Transactions on Circuits and Systems I (*IEEE TCAS-I*) in 2016.



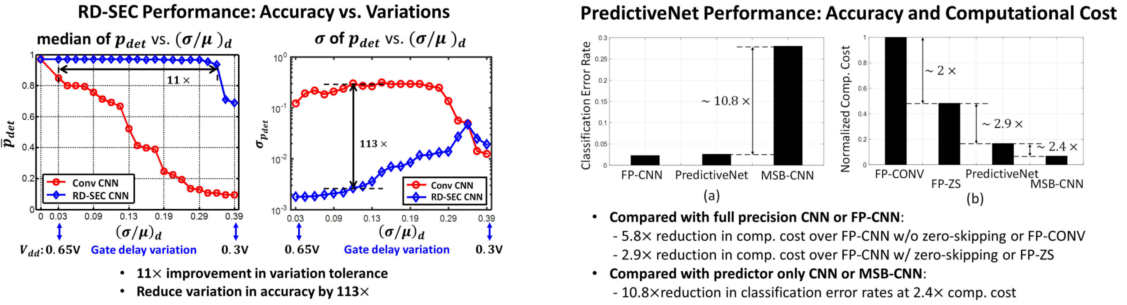
- **3-bit BOA receiver over 4-bit CUA receiver:**
  - 10^9 lower BER at the same TX amplitude
  - 60mV lower TX amplitude for BER < 10^{-12}

Figure 1. The chip micrograph, measurement set-up and performance of the BER-optimal ADC (BOA) based receiver.

Energy Efficiency Challenge at the Edge: The growing number of devices at the edge, such as smart phones, have limited energy, computational and storage resources, while many ML algorithms are computationally intensive. Therefore, the energy efficiency challenge is exacerbated if ML algorithms are to be embedded for local inference capability. To address the problem of resource-constrained computing at the edge, I tackle the issue of energy-efficient implementation of ML algorithms, particularly convolutional neural networks (CNNs). CNNs have recently gained considerable interest due to their record-breaking performance in many recognition tasks. However, their computational complexity hinders their application on power-constrained embedded platforms. I propose two techniques to enhance the energy efficiency of CNN design.

First, I propose a new statistical error compensation (SEC) technique referred to as rank decomposed SEC (RD-SEC). When applied to a CNN architecture in near threshold voltage (NTV) regime, simulation results (see Fig. 2 (a)) showed that the proposed architecture enables robust CNN design in the NTV regime. Moving forward, I am working to provide analytical justification of RD-SEC, such as analytical bound of the estimation errors and optimal choice of the rank to minimize the overall estimation errors.

Our work [3] received the 2nd place Best Student Paper Award when I presented it at the IEEE International Workshop on Signal Processing Systems in 2016 (*SiPS 2016*).



- 11× improvement in variation tolerance
- Reduce variation in accuracy by 113×

- Compared with full precision CNN or FP-CNN:
  - 5.8× reduction in comp. cost over FP-CNN w/o zero-skipping or FP-CONV
  - 2.9× reduction in comp. cost over FP-CNN w/ zero-skipping or FP-ZS
- Compared with predictor only CNN or MSB-CNN:
  - 10.8×reduction in classification error rates at 2.4× comp. cost

(a) (b)

Figure 2. Performance of the proposed: (a) rank decomposed statistical error compensation (RD-SEC), and (b) *predictive* CNN (PredictiveNet) techniques.

Second, I propose a *predictive* CNN (PredictiveNet), which predicts the sparse outputs of the nonlinear layers thereby bypassing a majority of computations. Simulation results (see Fig. 2 (b)) showed that the proposed PredictiveNet can significantly reduce the computational cost while incurring marginal accuracy degradation. Encouraged by these excellent results, I am currently working to combine RD-SEC and PredictiveNet for an ultra energy-efficient CNN design.

Our work [4] has been submitted to IEEE International Symposium on Circuits and Systems 2017 (*ISCAS 2017*) for peer review.

Future Research

Looking forward, I am excited to continue working in the broad area of energy-efficient ML systems. I would like to explore the following three directions through active collaboration with faculty members in related areas.

Systems: Many ML algorithms are essentially optimization problems and try to minimize certain loss functions. This provides a system-level opportunity to improve energy efficiency: the original optimization problem can be reformulated by introducing extra architecture or circuit constraints for energy purposes. Such resource-constrained reformulation will enable systematic design of energy-efficient ML systems. Possible constraints include cost of data movements, precision requirements for data representation, and others. For example, imposing constraints that favor the reduction of estimation errors in my proposed RD-SEC technique could possibly eliminate the need for power-hungry implementations.

Architectures: The design of ML systems traditionally employs an expensive worst case design methodology to ensure reliable circuit operation, limiting the achievable energy efficiency. Therefore, new architectures should embrace the inherent robustness of ML algorithms, and bridge the gap between the statistical nature of performance metrics in ML systems and the stochastic device behavior in nanoscale fabrics. I would like to explore the possibility of completely eliminating the partition of data storage and processing units. On the other hand, I would like to investigate new SEC techniques by taking advantage of ML algorithms' inherent tolerance to errors and the inherent redundancy within the algorithms themselves. In fact, the RD-SEC technique I proposed is one such heuristic step.

Circuits: ML algorithms relax precision and linearity requirements of the underlying circuits and devices. I would like to leverage my solid background on circuits and devices, and study new circuit techniques that can take advantage of tolerated non-linearities for energy purposes. On the device side, emerging technologies have the potential for either aggressive energy savings but are subject to various hardware errors. I am particularly interested in investigating new SEC techniques to unfold their excellent energy efficiency.

Moving forward, I look forward to working with passionate graduate students and collaborating with professors from various fields and departments. My extensive background and research experience in devices, circuits, VLSI systems and architectures, and ML algorithms will help ensure my research is innovative and practical.

References

[1] Y. Lin, M. S. Keel, A. Faust, A. Xu, N. R. Shanbhag, E. Rosenbaum, and A. C. Singer, "A study of BER-optimal ADC-based receiver for serial links," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 5, pp. 693–704, 2016.

[2] Y. Lin, A. Xu, N. R. Shanbhag, and A. C. Singer, "Energy-efficient high-speed links using BER-optimal ADCs," in *Electrical Design of Advanced Packaging and Systems Symposium (EDAPS), 2011 IEEE*, 2011, pp. 1–4.

[3] Y. Lin, S. Zhang, and N. R. Shanbhag, "Variation-tolerant architectures for convolutional neural networks in the near threshold voltage regime," in *Signal Processing Systems (SiPS), IEEE Workshop on*, 2016.

[4] Y. Lin, C. Sakr, Y. Kim, and N. R. Shanbhag, "Predictivenet: An energy-efficient convolutional neural network via zero prediction," in *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*, 2017.