# Artificial Intelligence, Communication, Imaging, Navigation, Sensing Systems

# Hybrid Intelligence in Design

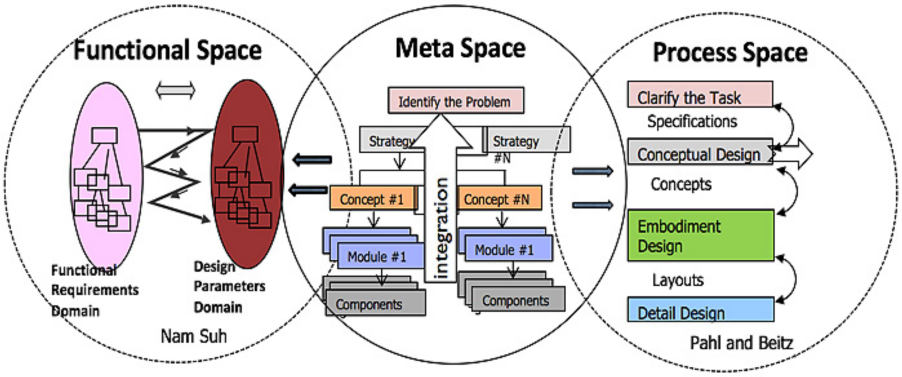
H. Akay, S-G. Kim

Sponsorship: MIT-Sensetime

One of the greatest challenges facing society is addressing the complexities of the big-picture, system-level, interdisciplinary problems in a holistic way. Human designers, architects, and engineers have come to rely on steadily improving computational tools to design, model, and analyze their systems of interest. The design of real-world systems (engineering, architecture, software, industrial, financial, and social systems) is, however, often a tumultuous endeavor fraught with great triumphs and, at times, significant regrets. Many believe that only human experts can conceptualize and orchestrate big projects upstream of designing systems.

There are two challenging issues in the current practice of a heuristic way of systems design. Firstly, it takes too long (decades) to become area experts through accumulating experience in many successes and some failures. Secondly, human experts also fail sometimes, especially at critical times. The questions one might ask at this stage are, "How could we teach junior engineers, architects, and scientists to design complex systems successfully without spending years of effort training on the job? Could we also assist human experts to minimize the probability of failure by leveraging recent developments in AI and big data?" While the resurgence of artificial intelligence and machine learning suggests ways to even more fully automate downstream tasks in the design process, we propose to go upstream of design, where all the key concepts are determined. Could machine intelligence help this early stage of designing beyond routine design and the optimization of pre-specified goals toward the generation of good, novel designs?

Our solution to the question above will be the use of Hybrid Intelligence: combining human intelligence, which grows through experience, and machine intelligence, which can learn from all the past successes and failures and does not forget them at all. Early-stage design across disciplines requires high-level intelligence based on one's intuition and experiential perceptions to understand challenges, constraints, and requirements in achieving the goals set. Instead of replacing humans with computational systems such as machine intelligence, we see humans and computers as working together within an ecosystem where each must bring their strengths to bear. We propose in the long run a fundamentally broad investigation of this likely convergence across the disciplines of Architecture, Structural Engineering, System Engineering, Mechanical Engineering, and Product Design. We call this approach Hybrid Intelligence because our concern is not with the intelligence of artifice, or the constraining of human designers, but rather with the effectiveness of their hybridized combination. Hybrid Intelligence for design is an umbrella term in which humans and computers collaborate from their strengths to find new processes for thinking, working, and designing.



▲ Figure 1: Illustration of the use of axiomatic design to reverse-map the past design successes and failures in terms of functional requirements and design parameters, which will then form hierarchical trees via design matrices.

## FURTHER READING

- S. G. Kim, S. M. Yoon, M. Yang, J. Choi, H. Akay, and E. Burnell, "AI for Design: Virtual Design Assistant," *CIRP Annals*.
- N. P. Suh, "The Principles of Design," *New York: Oxford University Press*, 1990.
- N. P. Suh, S. Kim, Bell, et al., "Optimization of Manufacturing Systems through Axiomatics," *CIRP Annals*, vol. 31, pp. 383-388, 1978.

# HAQ: Hardware-aware Automated Quantization

K. Wang, Z. Liu, Y. Lin, J. Li, S. Han

Sponsorship: IBM, QI, SONY, Intel, Samsung, Xilinx, Qualcomm, ARM, AMD, Amazon

Model quantization is a widely used technique to compress and accelerate deep neural network (DNN) inference. Emergent DNN hardware accelerators begin to support mixed precision (1-8 bits) to improve the computation efficiency further. This goal raises a great challenge to find the optimal bitwidth for each layer: it requires domain experts to explore the vast design space, trading off among accuracy, latency, energy, and model size, which is both time-consuming and sub-optimal. The conventional quantization algorithm ignores the different hardware architectures and quantizes all the layers uniformly.

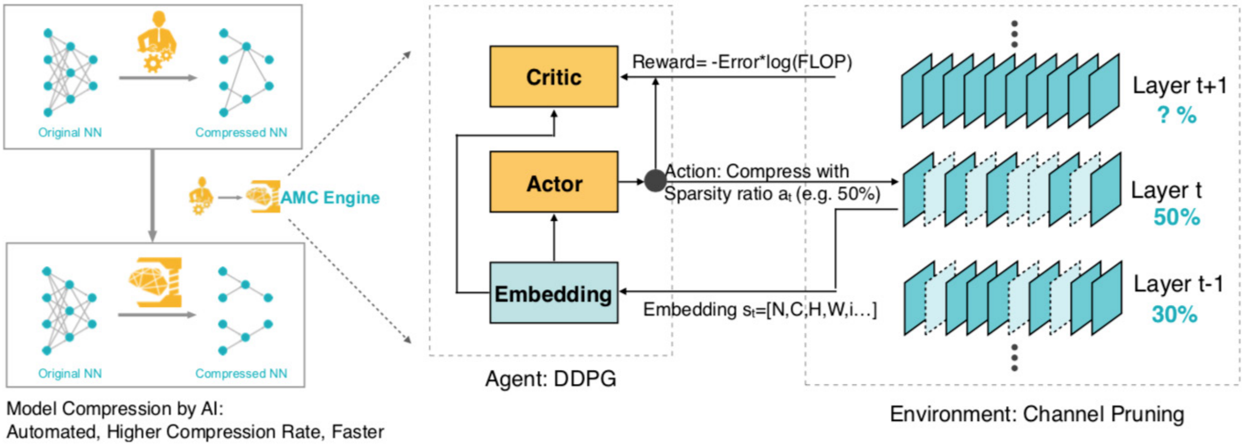
In this paper, we introduce the Hardware-aware Automated Quantization (HAQ) framework, which leverages the reinforcement learning to determine the quantization policy automatically, and we take the hardware accelerator's feedback in the design loop. Rather than relying on proxy signals such as FLOPs and model size, we employ a hardware simulator to generate direct feedback signals (latency and energy) to the RL agent. Compared with conventional methods, our framework is fully automated and can specialize the quantization policy for different neural network architectures and hardware architectures. Our framework effectively reduced the latency by 1.4-1.95x and the energy consumption by 1.9x with negligible loss of accuracy compared with the fixed bit width (8 bits) quantization. Our framework reveals that the optimal policies on different hardware architectures (i.e., edge and cloud architectures) under different resource constraints (i.e., latency, energy, and model size) are drastically different. We interpreted the implications of different quantization policies, which offer insights for both neural network architecture design and hardware architecture design.

# AMC: AutoML for Model Compression and Acceleration on Mobile Devices

Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, S. Han

Sponsorship: IBM, QI, SONY, Intel, Samsung, Xilinx, Qualcomm, ARM, AMD, Amazon

Model compression is a critical technique to efficiently deploy neural network models on mobile devices, which have limited computation resources. Conventional model compression techniques rely on hand-crafted heuristics and rule-based policies that require domain experts to explore the large design space, which is usually sub-optimal and time-consuming. In this paper, we propose AutoML for Model Compression (AMC), which leverages reinforcement learning to provide the model compression policy. This learning-based policy outperforms conventional rule-based policy by having a higher compression ratio, better preserving the accuracy, and freeing human labor. Under 4x floating point operations per second (FLOPs) reduction, we achieved 2.7% better accuracy than the hand-crafted compression policy for VGG-16 on ImageNet. We applied this automated compression pipeline to MobileNet and achieved a 1.81x speedup of measured inference latency on an Android phone and 1.43x speedup on the Titan XP GPU, with only 0.1% loss of ImageNet accuracy.



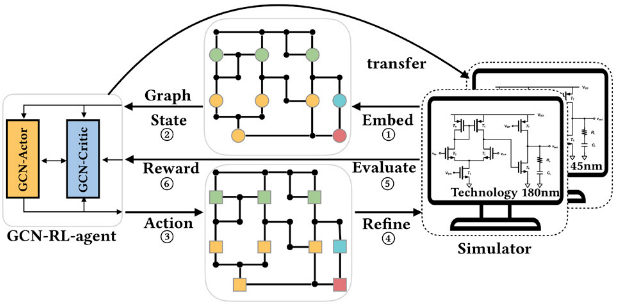
▲ Overview of AutoML for Model Compression (AMC) engine.

# Transferable Automatic Transistor Sizing with Graph Neural Networks and Reinforcement Learning

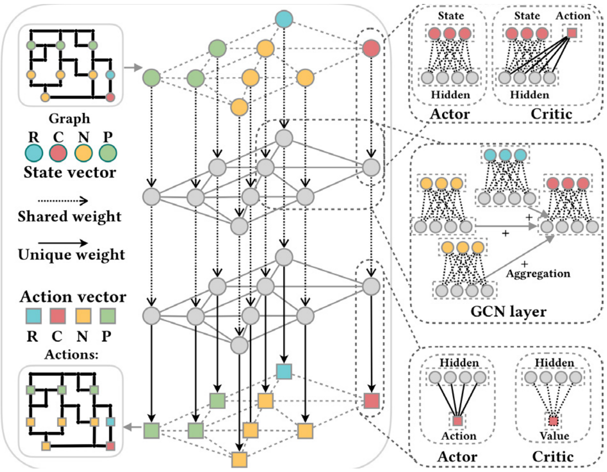
H. Wang, K. Wang, J. Yang, L. Shen, N. Sun, H.-S. Lee, S. Han
Sponsorship: IBM, QI, SONY, Intel, Samsung, Xilinx, Qualcomm, ARM, AMD, Amazon

Automatic transistor sizing is challenging due to the large design space, complex performance trade-offs, and fast technology advancement. Although much work has focused on transistor sizing targeting one circuit, limited research has explored transferring knowledge from one circuit to another to reduce redesign overhead. We propose leveraging a Reinforcement Learning (RL) algorithm to conduct knowledge transfer between different technology nodes and schematics. Inspired by the fact that circuits are graphs, we also propose to learn on the schematic graph with Graph Convolutional Neural Networks (GCN). The GCN-RL agent extracts features on the schematic graph, whose vertices are transistors and edges are wires. By learning the schematic information, our method consistently achieves higher Figures of Merit (FoMs) on four different circuits than conventional black box optimization methods (Bayesian Optimization, Evolutionary Algorithms). Experiments on transfer learning between five technology nodes and two circuit schematics demonstrate that with the same number of simulations, RL with transfer learning can achieve much higher FoMs than agents without knowledge transfer.

To the best of our knowledge, we are the first to leverage RL to transfer knowledge between technology nodes and schematics and to leverage GCN to learn on the schematic graph. Our work makes three main contributions. First, we leverage the schematic graph information in the optimization loop (open-box optimization) to build a GCN based on the circuits schematic graph to open the optimization black box effectively and embed the domain knowledge of circuits to improve performance. We use RL as an optimization algorithm; it consistently achieves better performance than a human expert, random search, Evolution Strategy, Bayesian Optimization, and MACE. Third, we use knowledge transfer with GCN-RL between technology nodes and circuit schematics to reduce the required number of simulations and shorten the design cycle.



▲ Figure 1: Overview: (1) Circuit schematic embedded into the graph (nodes: transistors, edges: wires) generates state vector for transistor. (2) Feed graph and states to RL agent. (3) RL agent processes vertices and generates sizes for transistor. (4) Environment refines actions. (5) circuit simulation (6) computes FoM value as reward, updating RL agent.



▲ Figure 2: Actor-Critic based GCN-RL Agent. Actor's first layer is fully-connected (FC) layer whose weight is shared among transistors. Critic's first layer is shared FC layer with transistor-specific encoder to encode different actions. Actor's last layer has transistor-specific decoder to decode hidden activations to different actions. Critic has shared FC layer to compute Q-values.

## FURTHER READING

- H. Wang, J. Yang, H. S. Lee, and S. Han, "Learning to Design Circuits," *arXiv* preprint arXiv:1812.02734, 2018.
- Y. He, J. Lin, Z. Liu, H. Wang, L. J. Li, and S. Han, "AMC: AutoML for Model Compression and Acceleration on Mobile Devices," *Proc. European Conference on Computer Vision (ECCV)*, pp. 784-800, 2018.

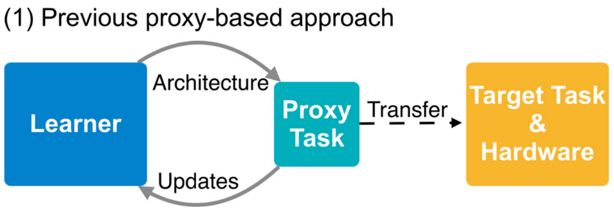
# ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware

H. Cai, L. Zhu, S. Han

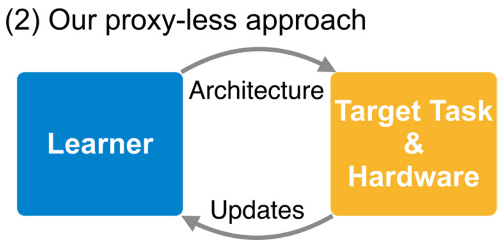
Sponsorship: MIT Quest for Intelligence, MIT-IBM Watson AI lab, SenseTime, Xilinx, Snap Research, AWS

Neural architecture search (NAS) has a great impact by automatically designing effective neural network architectures. However, the prohibitive computational demand of conventional NAS algorithms (e.g., $10^4$ GPU hours) makes it difficult to directly search the architectures on large-scale tasks (e.g., ImageNet). Differentiable NAS can reduce the cost of GPU hours via a continuous representation of network architecture but suffers from the high GPU memory consumption issue (grow linearly w.r.t. candidate set size). As a result, they need to utilize *proxy* tasks, such as training on a smaller dataset, or learning with only a few blocks, or training just for a few epochs. These architectures optimized on proxy tasks are not guaranteed to be optimal on the target task.

In this paper, we present *ProxylessNAS* that can *directly* learn the architectures for large-scale target tasks and target hardware platforms. We address the high memory consumption issue of differentiable NAS and reduce the computational cost (GPU hours and GPU memory) to the same level of regular training while still allowing a large candidate set. Experiments on CIFAR-10 and ImageNet demonstrate the effectiveness of directness and specialization. On CIFAR-10, our model achieves 2.08% test error with only 5.7M parameters, better than the previous state-of-the-art architecture AmoebaNet-B, while using 6X fewer parameters. On ImageNet, our model achieves 3.1% better top-1 accuracy than MobileNetV2, while being 1.2X faster with measured GPU latency. We also apply ProxylessNAS to specialize neural architectures for hardware with direct hardware metrics (e.g., latency) and provide insights for efficient CNN architecture design.



▲Figure 1: ProxylessNAS directly optimizes neural network architectures on target task and hardware. Benefiting from the directness and specialization, ProxylessNAS can achieve remarkably better results than previous proxy-based approaches.



▲Figure 2: ProxylessNAS consistently outperforms MobileNetV2 under various latency settings. On ImageNet, with only 200 GPU hours, our searched CNN model for mobile achieves the same level of top-1 accuracy as MobileNetV2 1.4 while being 1.8X faster.

## FURTHER READING

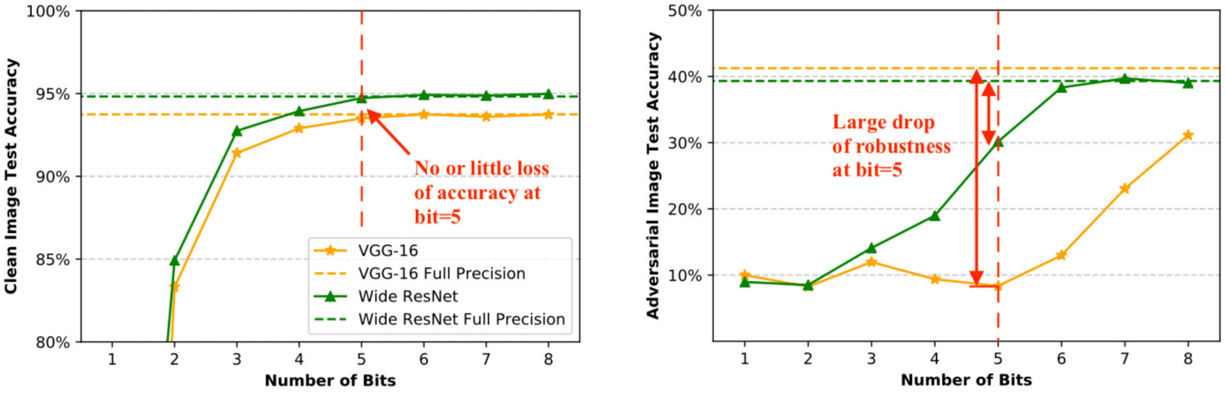
- M. Tan, et al., "Mnasnet: Platform-aware Neural Architecture Search for Mobile," arXiv preprint arXiv:1807.11626, 2018.
- M. Sandler, et al., "Mobilenetv2: Inverted Residuals and Linear Bottlenecks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- C. Han, L. Zhu, and S. Han. "ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware," ICLR, 2019.

# Defensive Quantization: When Efficiency Meets Robustness

J. Lin, C. Gan, S. Han
Sponsorship: IBM, QI, Google, Facebook

Neural network quantization is becoming an industry standard to efficiently deploy deep learning models on hardware platforms such as CPU, GPU, TPU, and FPGAs. However, we observe that the conventional quantization approaches are vulnerable to adversarial attacks. This paper aims to raise awareness about the security of the quantized models, and we designed a novel quantization methodology to optimize the efficiency and robustness of deep learning models jointly.

We first conduct an empirical study to show that vanilla quantization suffers more from adversarial attacks. We observe that the inferior robustness comes from the error amplification effect, where the quantization operation further enlarges the distance caused by amplified noise. Then we propose a novel defensive quantization (DQ) method by controlling the Lipschitz constant of the network during quantization, such that the magnitude of the adversarial noise remains non-expansive during inference. Extensive experiments on CIFAR-10 and Street View House Number datasets demonstrate that our new quantization method can defend neural networks against adversarial examples and even achieves superior robustness to their full- precision counterparts while maintaining the same hardware efficiency as vanilla quantization approaches. As a by-product, DQ can also improve the accuracy of quantized models without adversarial attack.



**(a)** Quantization preserves the accuracy till 4-5 bits on clean image.

**(b)** Quantization no longer preserves the accuracy under adversarial attack (same legend as left).

▲ Figure 1: Quantized neural network is more vulnerable to adversarial attack.

# Scalable Free-space Optical Neural Networks

L. Bernstein, A. Sludds, R. Hamerly, D. Englund

Sponsorship: P NSERC, NSF, ORISE IC Postdoctoral Fellowship at MIT (U.S. DOE / ODNI), U.S. ARO through the ISN at MIT

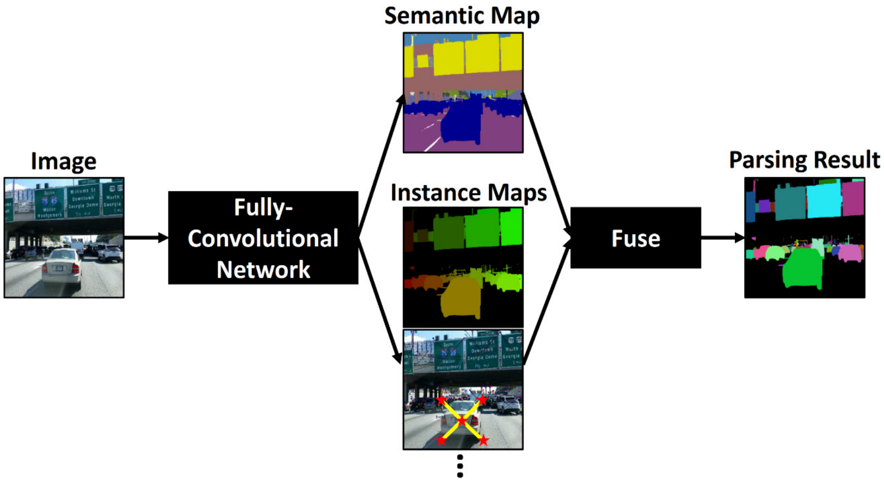
The transformative impact of deep neural networks (DNNs) in many fields has motivated the development of hardware accelerators to improve speed and power consumption. We present a novel photonic approach based on homodyne detection where inputs and weights are encoded optically and can be reprogrammed and trained on the fly. This architecture is naturally adapted to free-space optics where both fully-connected and convolutional networks can be implemented and scaled to millions of neurons. By utilizing passive optical fan-out and performing arithmetic coherently with optical interference, this scheme circumvents fundamental limits of irreversible electronic processing. We study the effect of detector shot noise on neural-network accuracy to establish a "standard quantum limit" for this system. This bound, which can be as low as 50 zJ/FLOP, suggests performance below the Landauer (thermodynamic) limit is theoretically possible with photonics.

## DeeperLab: Single-shot Image Parser

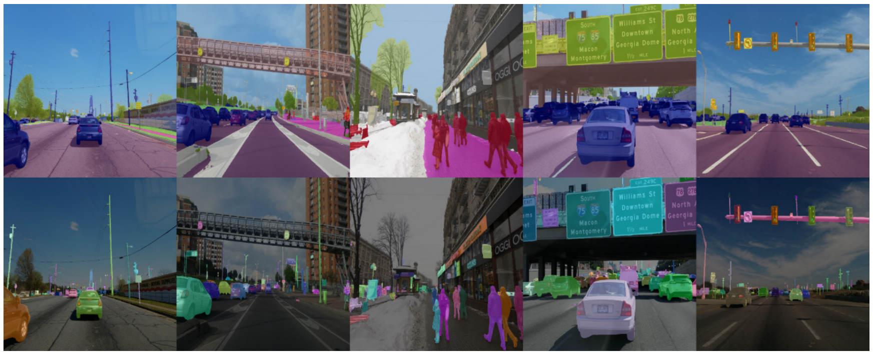
T.-J. Yang, M. D. Collins, Y. Zhu, J.-J. Hwang, T. Liu, X. Zhang, V. Sze, G. Papandreou, L.-C. Chen
Sponsorship: MIT, Google

Image parsing is the process of partitioning an image into multiple semantically meaningful regions (called semantic segmentation), such as car and road, and telling different countable instances apart (called instance segmentation), such as car A and car B. It is a long-lasting unsolved problem in computer vision and a basic component of many applications, such as autonomous driving. Recent approaches to image parsing typically employ separate standalone neural networks for the semantic and instance segmentation tasks and require multiple passes of inference.

Instead, the proposed DeeperLab image parser performs image parsing with a significantly simpler, more fully convolutional approach that jointly addresses the semantic and instance segmentation tasks and requires only one pass of inference (i.e., one-shot), resulting in a streamlined system that better lends itself to fast processing. For quantitative evaluation, we use both the instance-based panoptic quality (PQ) metric and the proposed region-based parsing covering (PC) metric, which better captures the image parsing quality on non-countable classes and larger object instances. We report experimental results on the challenging Mapillary Vistas dataset, in which our single model achieves 31.95% (val) / 31.6% PQ (test) and 55.26% PC (val) with 3 frames per second (fps) on a graphics processing unit (GPU) or near real-time speed (22.6 fps on GPU) with reduced accuracy.



▲ Figure 1: Illustration of DeeperLab



▲ Figure 2: Visualization on Mapillary Vistas Validation Set

### FURTHER READING

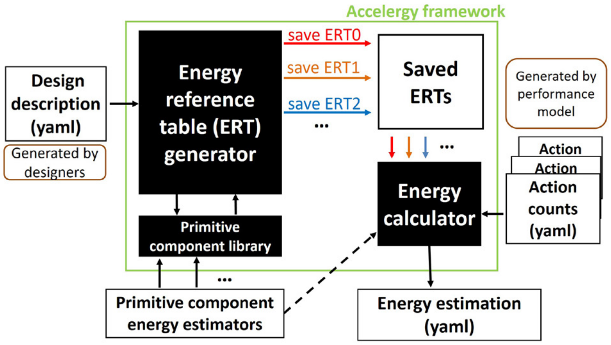
- T.-J. Yang et al., “DeeperLab: Single-shot Image Parser,” arXiv, 2019.
- L.-C. Chen et al., “Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation,” *ECCV*, 2018.
- G. Papandreou et al., “Personlab: Person Pose Estimation and Instance Segmentation with a Bottom-up, Part-based, Geometric Embedding Model,” *ECCV*, 2018.

# Architecture-level Energy Estimation of Accelerator Designs

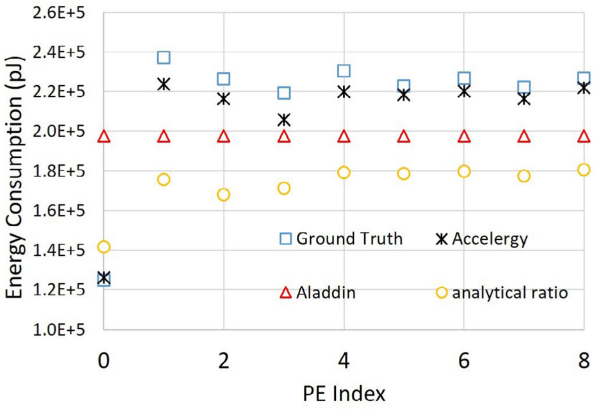
Y. N. Wu, J. S. Emer, V. Sze
Sponsorship: DARPA

With Moore's law slowing down and Dennard scaling ending, energy-efficient domain-specific accelerators, such as deep neural network (DNN) processors for machine learning and programmable network switches for cloud applications, have become a promising direction for hardware designers to continue bringing energy-efficiency improvements to data and computation intensive applications. To ensure fast exploration of accelerator design space, architecture-level energy estimators, which perform energy estimations without requiring complete hardware description of the designs, are critical to designers. However, using existing architecture-level energy estimators to obtain accurate estimates for accelerator designs is hard, as accelerator designs are diverse and sensitive to data patterns (e.g., sparsity in DNNs).

To solve this problem, we present Accelergy (Figure 1), an architecture-level energy estimation methodology for accelerator designs. Accelergy interprets a design in terms of its components (e.g., an arithmetic logic unit (ALU) design consists of multipliers and adders). Since accelerator design space is very diverse, Accelergy allows users to define their own components to describe the designs. At the same time, to reflect the energy differences brought by special data processing (e.g., zero-gating in DNN accelerators), Accelergy allows users to define special actions types related to the components (e.g., read and gated read actions for SRAM). To illustrate the usage of Accelergy methodology, we implemented a sample framework for energy estimations of DNN accelerators. The framework provides a set of primitive components for users to describe the design or construct their new components. To further enhance flexibility, Accelergy provides an interface to communicate with different primitive component estimators for system-level estimations of designs that involve emerging technologies (e.g., optical DNN). Accelergy achieves 95% accuracy on total energy estimation with a well-known accelerator design – Eyeriss. Accelergy can also produce accurate energy breakdown across components estimations comparing to other estimation methodologies (Figure 2).


Accelergy framework
Design description (yaml)
Generated by designers
Energy reference table (ERT) generator
save ERT0
save ERT1
save ERT2
...
Saved ERTs
Generated by performance model
Primitive component library
...
Energy calculator
Action
Action
Action counts (yaml)
Primitive component energy estimators
Energy estimation (yaml)

▲Figure 1: System diagram of Accelergy. Accelergy takes in design description and run time action counts as inputs and generates the energy estimation as the output.



▲Figure 2: Energy estimation comparison on the energy breakdown across the processing engines (PEs); in Eyeriss PE array (only selected PE are shown).

## FURTHER READING

- Y.-H. Chen, T. Krishna, J. Emer, and V. Sze, "Eyeriss: An Energy-efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE J. of Solid-State Circuits*, vol. 52, pp. 127-138, 2016.
- T.-J. Yang, Y.-H. Chen, J. Emer, and V. Sze, "A Method to Estimate the Energy Consumption of Deep Neural Networks," *Asilomar*, 2017.

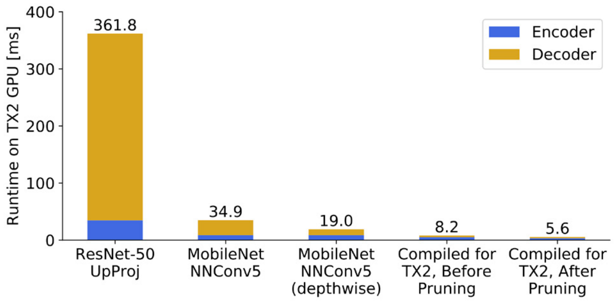
# FastDepth: Fast Monocular Depth Estimation on Embedded Systems

D. Wofk, F. Ma, T.-J. Yang, S. Karaman, V. Sze

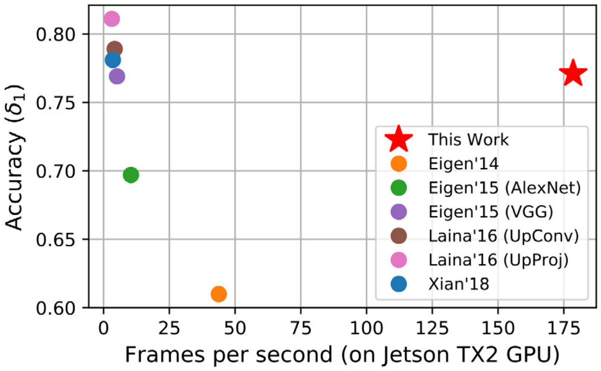
Sponsorship: Analog Devices, Inc.

Depth sensing is a critical function for many robotic tasks such as localization, mapping, and obstacle detection. There has been significant and growing interest in performing depth estimation from a single red-green-blue image, due to the relatively low cost and size of monocular cameras. However, state-of-the-art single-view depth estimation algorithms are based on fairly large deep neural networks that have high computational complexity and slow runtimes on embedded platforms. This poses a significant challenge when cameras perform real-time depth estimation on an embedded platform, for instance, mounted on a micro aerial vehicle.

Our work addresses this problem of fast depth estimation on embedded systems. We investigate efficient and lightweight encoder-decoder network architectures. To further improve their computational efficiency in terms of real metrics (e.g., latency), we apply resource-aware network adaptation (NetAdapt) to automatically simplify proposed architectures. In addition to reducing encoder complexity, our proposed optimizations significantly reduce the cost of the decoder network (Figure 1). We perform hardware-specific compilation targeting deployment on the NVIDIA Jetson TX2 platform. Our methodology demonstrates that it is possible to achieve accuracy similar to that of prior work on depth estimation, but at inference speeds that are an order of magnitude faster (Figure 2). Our proposed network, FastDepth, runs at 178 fps on a TX2 GPU and at 27 fps when using only the TX2 CPU, with active power consumption under 10 W.



▲ Figure 1: Impact of optimizations on our lightweight encoder-decoder network architecture for depth estimation. Our approach achieves significant reduction in inference runtime of encoder and decoder. Stacked bars represent encoder-decoder breakdown; total runtimes appear above bars.



▲ Figure 2: Accuracy vs. runtime (in fps) on NVIDIA Jetson TX2 GPU for various depth estimation algorithms. Top right represents desired characteristics: high throughput and high accuracy. Our work is an order of magnitude faster than prior work, maintaining comparable accuracy.

## FURTHER READING

- T.-J. Yang, A. Howard, B. Chen, X. Zhang, A. Go, M. Sadler, V. Sze, and H. Adam, "NetAdapt: PlatformAware Neural Network Adaptation for Mobile Applications," *European Conference on Computer Vision (ECCV)*, 2018.
- V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

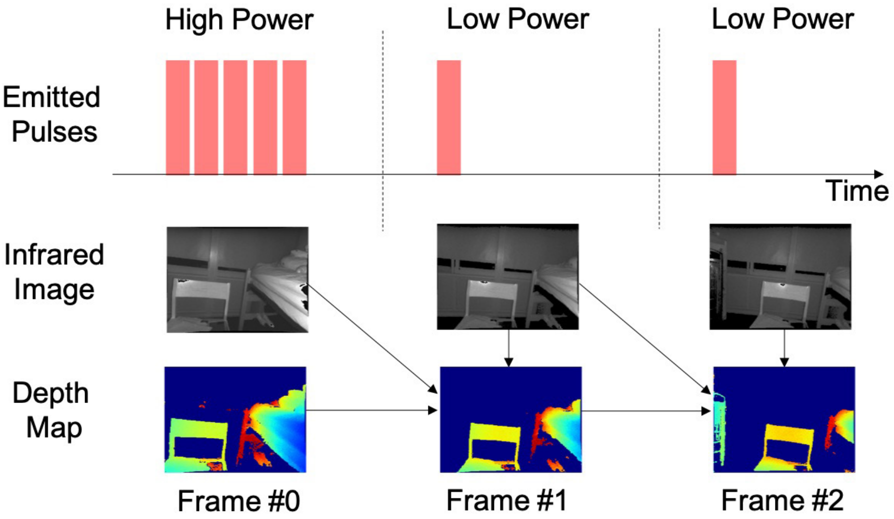
# Low-power Adaptive Time-of-Flight Imaging for Multiple Rigid Objects

J. Noraky, C. Mathy, A. Cheng, V. Sze

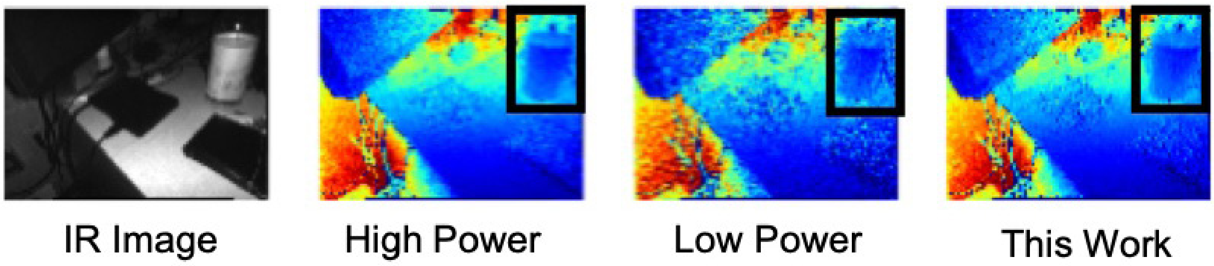
Sponsorship: Analog Devices, Inc.

Time-of-Flight (ToF) cameras are becoming increasingly popular for many mobile applications. To obtain accurate depth maps, ToF cameras must emit many pulses of light, which consumes a lot of power and lowers the battery life of mobile devices. However, lowering the number of emitted pulses results in noisy depth maps. To obtain accurate depth maps while reducing the overall number of emitted pulses, we propose an algorithm that adaptively varies the number of pulses to infrequently obtain high-power depth maps and uses them to help estimate subsequent low- power ones as shown in Figure 1. To estimate these depth maps, our technique uses the previous frame by accounting for the 3D motion in the scene. We assume that the scene contains independently moving rigid objects and show that we can efficiently estimate the motions.

In contrast to our previous work, this approach uses only the data from the ToF camera and does not need RGB images to estimate the 3D motion in the scene. The resulting algorithm estimates 640 × 480 depth maps at 30 frames per second on an embedded processor. We evaluate our approach on data collected with a pulsed ToF camera and show that we can reduce the mean relative error of the low-power depth maps by up to 65% (see Figure 2) and the number of emitted pulses by up to 80%.



▲Figure 1: We adaptively vary the number of pulses a ToF camera emits. For low-power depth maps, our depth estimation algorithm uses data from the previous frame along with the current one.



▲Figure 2: Our estimated depth map is less noisy than the low-power one. Best viewed in color.

## FURTHER READING

- J. Noraky and V. Sze, "Low Power Depth Estimation for Time-of-Flight Imaging," *IEEE International Conference on Image Processing*, 2017.
- J. Noraky and V. Sze, "Depth Estimation of non-Rigid Objects for Time-of-Flight Imaging," *IEEE International Conference on Image Processing*, 2018.
- J. Noraky and V. Sze, "Low Power Depth Estimation of Rigid Objects for Time-of-Flight Imaging," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*..

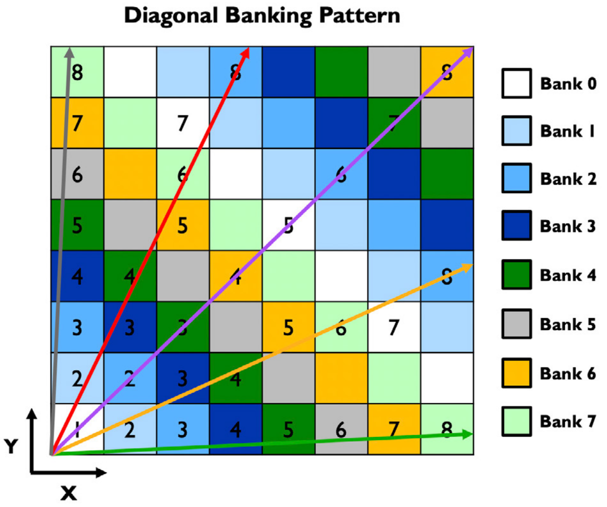
# Fast Shannon Mutual Information Accelerator for Autonomous Robotics Exploration

P. Z. X. Li, Z. Zhang, S. Karaman, V. Sze

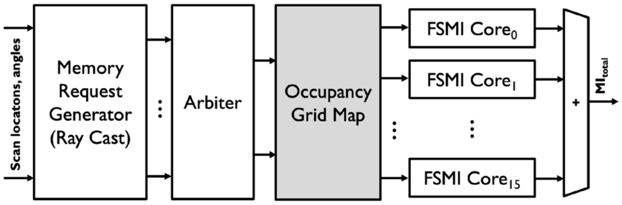
Robotic exploration problems arise in various contexts, ranging from search and rescue missions to underwater and space exploration. In these domains and beyond, exploration algorithms that can rapidly reduce uncertainty can provide significant benefits, for instance, by shortening time and reducing resources required for exploration. Unfortunately, principled algorithms based on rigorous information-theoretic metrics, such as maximizing Shannon mutual information (MI) along the exploration path, are computationally extremely demanding.

We propose a novel computing hardware architecture to efficiently compute Shannon MI on an occupancy grid map, which is the standard probabilistic representation for a 2D environment. The proposed architecture consists of multiple MI computation cores, each evaluating the MI between a single sensor beam and the occupancy grid map. We find that parallelization alone is not sufficient for high-throughput computation due to the limited bandwidth of the memory. In fact, it is critical to consider 1) memory management of the occupancy grid map storage and 2) data delivery from the occupancy grid map to MI cores. Thus, our key contributions consist of 1) a novel memory architecture that diagonally partitions the occupancy grid map into multiple banks to minimize the memory access conflicts among multiple cores (Figure 1); 2) a fast and fair memory request arbiter that ensures effective utilization of all MI computation cores; and 3) an energy-efficient, high-throughput MI computation core.

This architecture (Figure 2) was optimized for 16 MI computation cores and was implemented on a field-programmable gate array. We show that it computes the MI metric for an entire map of 20m × 20m at 0.1m resolution in near real time, at 2 frames per second, which is approximately two orders of magnitude faster, while consuming an order of magnitude less power than an equivalent implementation on a Xeon CPU.



▲ Figure 1: Diagonal partitioned occupancy grid map minimizes memory read conflicts at each cycle, indexed by the numbers. Since each bank has two (read) ports, the same color cannot appear more than twice in each row or column of the map.



▲ Figure 2: Overview of the top-level architecture consisting of memory request generator (Bresenham ray-casting), fast and fair arbiter, diagonally partitioned occupancy grid map, and fast MI computation cores.

## FURTHER READING

- P. Z. X. Li, Z. Zhang, S. Karaman, and V. Sze, "High-throughput Computation of Shannon Mutual Information on Chip," *Robotics: Science and Systems (RSS)*, Freiburg, Germany, 2019.
- Z. Zhang, T. Henderson, S. Karaman, and V. Sze, "FSMI: Fast Computation of Shannon Mutual Information for Information-theoretic Mapping," *International Conference on Robotics and Automation (ICRA)*, Montreal, Canada, 2019.

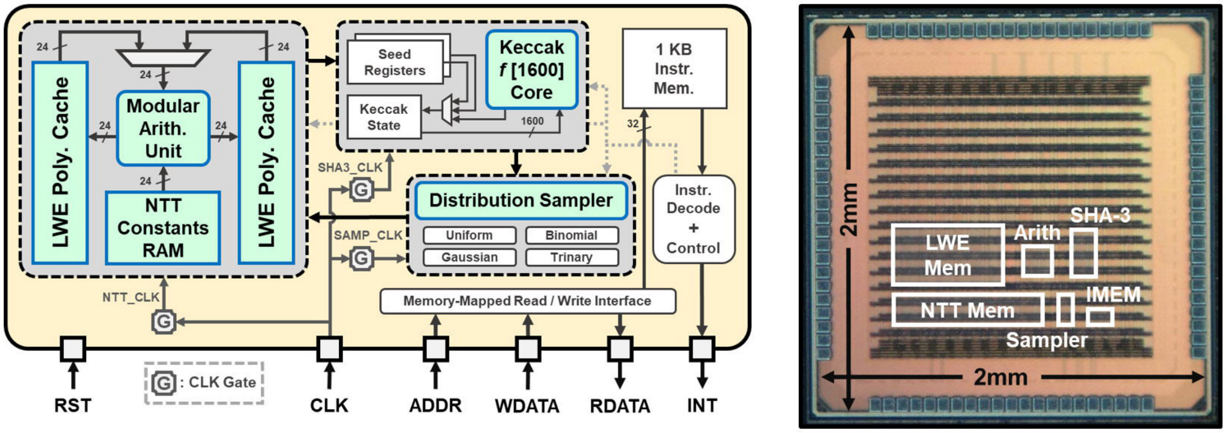
# An Energy-efficient Configurable Lattice Cryptography Processor for the Quantum-secure Internet of Things

U. Banerjee, A. P. Chandrakasan

Sponsorship: Texas Instruments

Modern public-key cryptography protocols, such as Rivest-Shamir-Adleman and elliptic-curve cryptography (ECC) will be rendered insecure by Shor's algorithm when large-scale quantum computers are built. Therefore, cryptographers are working on quantum-resistant algorithms, and lattice-based cryptography has emerged as a prime candidate. However, the high computational complexity of these algorithms makes it challenging to implement lattice-based protocols on resource-constrained Internet of things (IoT) devices, which need to secure data against both present and future adversaries. To address this challenge, we present a lattice cryptography processor with configurable parameters that enables energy savings of up to two orders of magnitude and 124k-gate reduction in system area through architectural optimizations. This is also the first ASIC implementation that demonstrates multiple lattice-based protocols proposed in the National Institute of Standards and Technology's post-quantum standardization process.

Figure 1 shows a block diagram of our system along with the chip micrograph. The chip was fabricated in a 40-nm low-power CMOS process and supported voltage scaling from 1.1V down to 0.68V. Our lattice cryptography processor occupies 106k NAND Gate Equivalents and uses 40.25KB of SRAM. When executing the Kyber-768 and NewHope-1024 key exchange schemes, our design is 28x and 37x more energy-efficient, respectively, than Cortex-M4 software, after accounting for voltage scaling. Moreover, post-quantum key exchange using our processor is 30x more energy-efficient than state-of-the-art pre-quantum ECC-based key exchange at the same pre-quantum security level. Through architectural and algorithmic optimizations, this work demonstrates practical hardware-accelerated quantum-resistant lattice-based cryptographic protocols that can be used to secure resource-constrained IoT devices of the near future.



▲ Figure 1: System block diagram and chip micrograph.

## FURTHER READING

- U. Banerjee, A. Pathak, and A. P. Chandrakasan, "An Energy-efficient Configurable Lattice Cryptography Processor for the Quantum-secure Internet of Things," *IEEE International Solid- State Circuits Conference (ISSCC)*, pp. 46-48, Feb. 2019.
- U. Banerjee, C. Juvekar, A. Wright, Arvind, and A. P. Chandrakasan, "An Energy-efficient Reconfigurable DTLS Cryptographic Engine for End-to-End Security in IoT Applications," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 42-44, Feb. 2018.

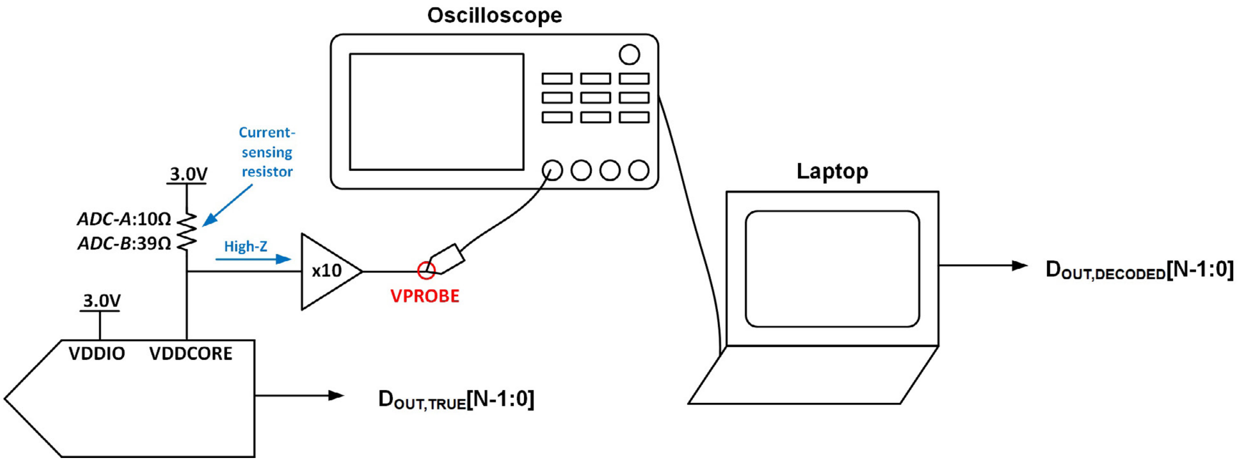
# Power Side-channel Attack on Successive Approximation Register Analog-to-digital Converters

T. Jeong, H.-S. Lee, A. P. Chandrakasan
Sponsorship: Analog Devices, Inc., Korea Foundation for Advanced Studies

When sensing hardware is used to acquire a private signal, there must be no information leakage throughout the entire signal chain. Applications that require such security include biomedical and military sensor platforms. Industrial and infrastructure monitoring sensing hardware must also be secure to prevent potentially harmful activities of adversaries. By using well-established cryptographic primitives, communication links for sensing hardware can be protected from hackers. However, once hackers physically access the sensing hardware, the sensor-interface circuit can leak critical information via its power side-channel.

Both analog and digital circuit blocks of the sensor-interface circuit can leak through a power side-channel as their operations depend on the sensor output value. Since the first discovery of the encryption engine's power side-channel leakage, countermeasures against digital circuit's power side-channel attacks have been researched in the cryptographic hardware community. However, unlike digital circuit blocks that can be protected by countermeasures, analog/mixed-signal circuit blocks are now vulnerable to side-channel attacks as their exploitations have not been recognized yet.

In this work, we have developed practical power side-channel attack scenarios that make analog/mixed-signal circuit blocks become the security loophole of the entire system. We chose analog-to-digital converters (ADCs) as our target block of study. We focused our research on successive approximation register (SAR) ADCs because they are more power-efficient than other ADC types in the performance range (resolution, sampling rate) that is suitable for most sensor platforms. To experiment with power side-channel attack on SAR ADCs, we devised an attack method and mounted it on two SAR ADC products from different manufacturers. The experimental results show that SAR ADCs' input waveforms could be faithfully reconstructed from their current traces.



▲ Figure 1: Overview of SAR ADC Power Side-channel Attack Experimental Setup.

## FURTHER READING

- P. C. Kocher, J. Jaffe, and B. Jun, "Differential Power Analysis," *Advances in Cryptology - CRYPTO '99, 19th Annual International Cryptology Conference*, pp. 388– 397, Santa Barbara, CA, Aug. 15-19, 1999.

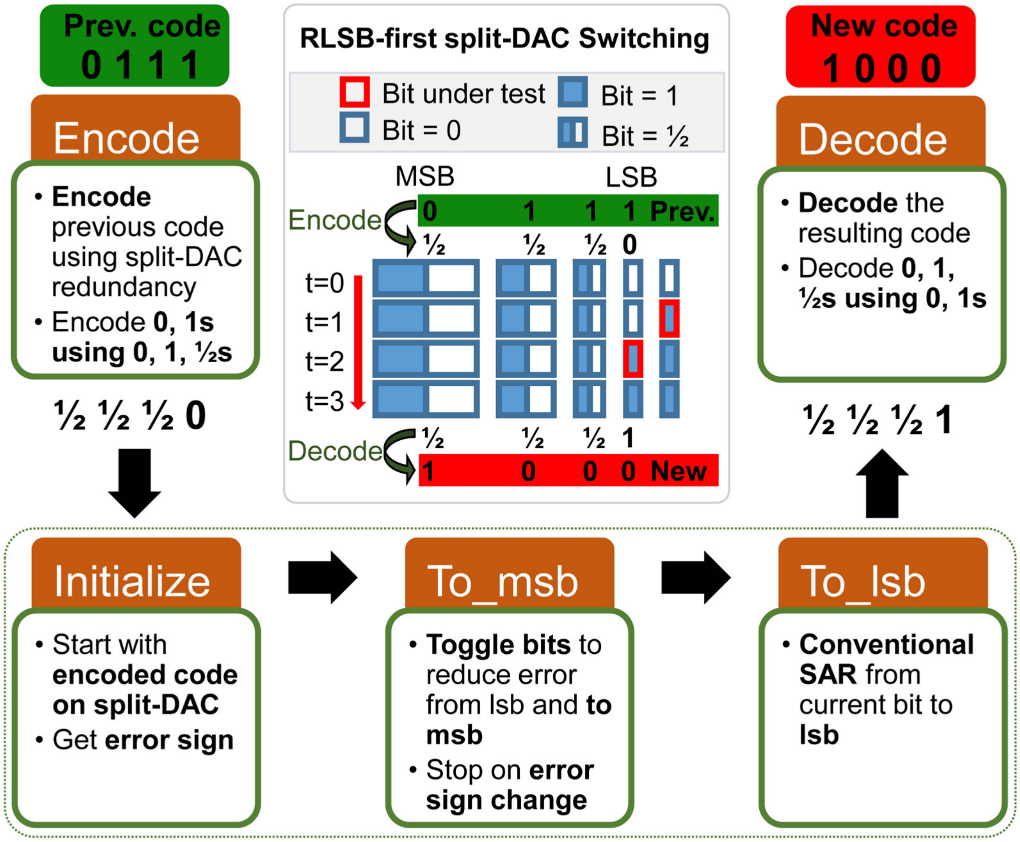
# Energy-efficient SAR ADC with Background Calibration and Resolution Enhancement

H. S. Khurana, A. P. Chandrakasan, H.-S. Lee

Sponsorship: Center for Integrated Circuits and Systems

Many signals, for example, medical signals, do not change much from sample to sample most of the time. Conventional switching schemes for SAR ADCs do not exploit this signal characteristic and test each bit starting with the MSB. Previous work called least-significant-bit (LSB)-first saves energy and bit-cycles by starting with a previous sample code and searching for the remainder by testing bits from the LSB end. However, certain code transitions consume unnecessary energy, even when the code change over the previous code is small.

This work addresses it with a new algorithm called Recode then LSB-first (RLSB-first) that reduces the switching energy and bit-cycles required for all cases of small code change across the full range of possible previous sample codes. RLSB-first uses split-DAC to systematically encode the previous code before LSB-first. RLSB-first lowers switching energy by up to 2.5 times and uses up to 3 times fewer bit-cycles than LSB-first. In addition to an energy-efficient SAR ADC, this work aims to use the savings for background calibration and resolution enhancement.


Prev. code
0 1 1 1
Encode
• Encode previous code using split-DAC redundancy
• Encode 0, 1s using 0, 1, ½s
½ ½ ½ 0
RLSB-first split-DAC Switching
Bit under test
Bit = 1
Bit = 0
Bit = ½
MSB
LSB
Encode
0 1 1 1 Prev.
½ ½ ½ 0
t=0
t=1
t=2
t=3
Decode
½ ½ ½ 1
1 0 0 0 New
New code
1 0 0 0
Decode
• Decode the resulting code
• Decode 0, 1, ½s using 0, 1s
½ ½ ½ 1
Initialize
• Start with encoded code on split-DAC
• Get error sign
To_msb
• Toggle bits to reduce error from lsb and to msb
• Stop on error sign change
To_lsb
• Conventional SAR from current bit to lsb

▲ Figure 1: Algorithm for RLSB-first.

## FURTHER READING

- F. M. Yaul and A. P. Chandrakasan, "11.3 A 10b 0.6nW SAR ADC with Data-dependent Energy Savings using LSB-first Successive Approximation," *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 198-199, 2014.
- H. S. Khurana, A. P. Chandrakasan, and H. Lee, "Recode then LSB-first SAR ADC for Reducing Energy and Bit-cycles," *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1-5, Florence, Italy, 2018.

# An 8-bit Multi-GHz Flash ADC with Time-based Techniques

X. Yang, H.-S. Lee

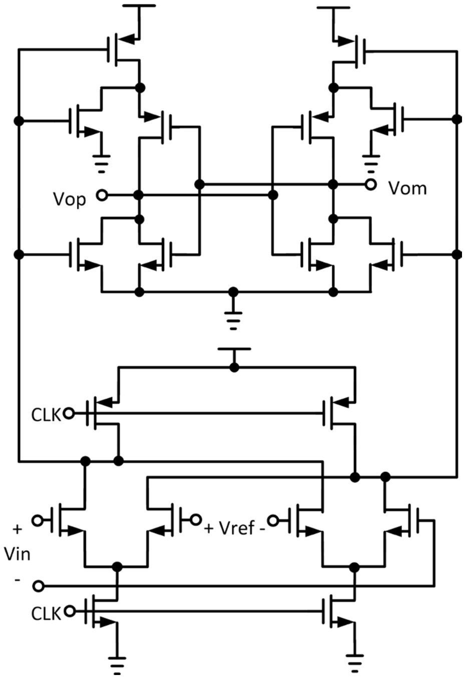
Sponsorship: MIT Center for Integrated Circuits and Systems

High-speed and medium-to-low-resolution flash analog-to-digital converters (ADCs) are widely used in applications such as 60-GHz receivers, serial links, and high-density disk drive systems. In this project, we propose an 8-bit, multi-gigahertz flash ADC with two major innovations: the time-based comparator offset calibration and the time-based 4x interpolation.

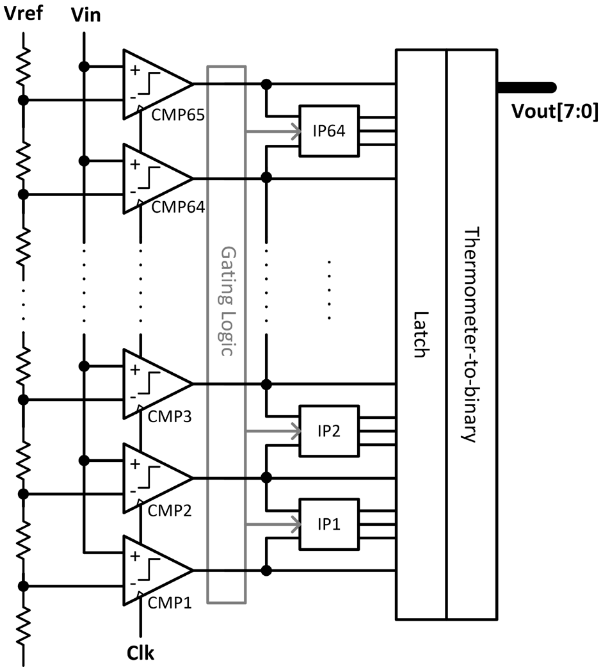
A high-speed, low-power comparator with low noise and offset requirements is a key building block. Figure 1 shows the two-stage dynamic comparator used in our design. With the scaling of CMOS technology, the offset voltage of the comparator keeps increasing due to greater transistor mismatches, making offset calibration a necessity. Traditional offset calibration methods that use digitally-controlled capacitor banks or extra input transistor pairs add extra parasitics to the comparators and slow down the operation. In this work, the proposed time-based comparator offset calibration put no additional load on the comparators and avoids the speed penalty of traditional methods.

The number of comparators in a flash ADC grows exponentially with resolution. This is a major drawback of flash ADCs. Time-domain interpolation is a popular technique that utilizes the timing information from adjacent comparators to resolve extra bits of resolution without adding comparators. Figure 2 shows the proposed flash ADC. Sixty-five comparators are used to achieve the six most significant bits (MSBs). Sixty-four interpolators are inserted between the comparators to obtain two extra bits by comparing the delay from neighboring comparators. The input capacitance of this design is ¼ of the conventional 8-bit flash ADC. Therefore, a higher operating speed can be achieved. We introduce gating logic so that only one interpolator is enabled during operation, which reduces power consumption significantly.

The prototype ADC is realized in 65-nm CMOS technology. At 2.8 GS/s, the prototype measures an SNDR of 43.3 dB at Nyquist input frequency and achieves a state-of-the-art figure-of-merit.



▲Figure 1: Schematic of the two-stage dynamic comparator.



▲Figure 2: Flash ADC architecture, with 65 comparators and 64 2-bit interpolators.

## FURTHER READING

- M. Miyahara, Y. Asada, D. Paik, and A. Matsuzawa, "A Low-noise Self-calibrating Dynamic Comparator for High-speed ADCs," *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC),* pp. 269-272, Nov. 2008.
- Y.-S. Shu, "A 6b 3GS/s 11mW Fully Dynamic Flash ADC in 40nm CMOS with Reduced Number of Comparators," *Symp. on VLSI Circuits Dig. Tech. Papers,* pp. 26-27, 2012.

# A Sampling Jitter-tolerant Continuous-time Pipelined ADC in 16-nm FinFET

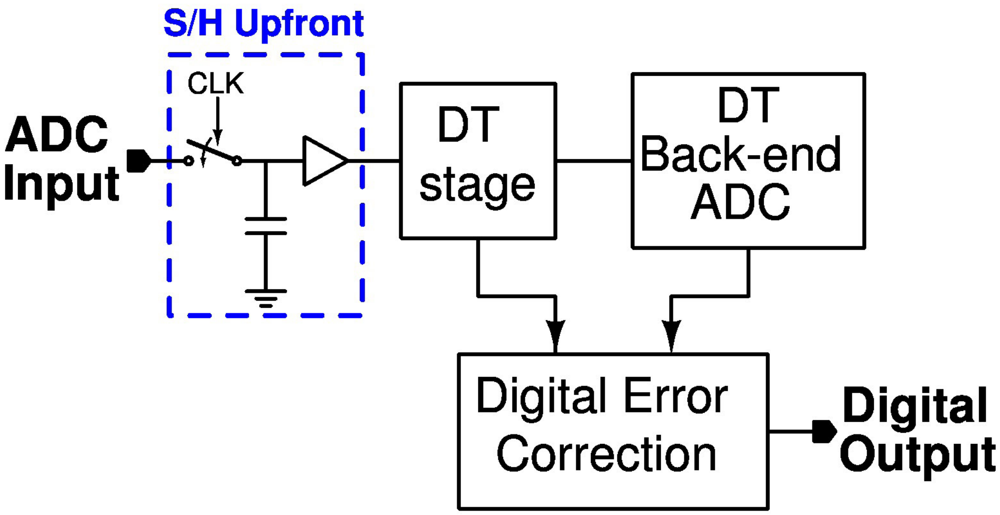
R. Mittal, G. Manganaro, A. P. Chandrakasan, and H.-S. Lee

Sponsorship: Analog Devices, Inc.

Analog-to-digital converters (ADCs) interface real-world analog signals with digital systems, and hence they are an essential part of any electronic system. Although there have been steady improvements in the performance of ADCs, the improvements in conversion speed have been less significant because the speed-resolution product is limited by the sampling clock jitter. The effect of sampling clock jitter has been considered fundamental. However, it has been shown that continuous-time delta-sigma modulators may reduce the effect of sampling jitter. Since delta-sigma modulators rely on relatively high oversampling, they are unsuitable for high-frequency applications such as 5G baseband processors. Therefore, ADCs with low oversampling ratio are desirable for high-speed data conversion.

In conventional Nyquist-rate ADCs, the input is sampled upfront (Figure 1). Any jitter in the sampling clock directly affects the sampled input and degrades the signal-to-noise ratio (SNR). For fast varying input signals, the sampling jitter severely limits the maximum attainable SNR. It is well known that for a known rms sampling jitter $\sigma_{t,}$ the maximum achievable SNR is limited to $1/(2\pi f_{in}\sigma_t)$, where $f_{in}$ is the input signal frequency. Typically, reducing the rms jitter below 100 fs is difficult. This challenge limits the maximum SNR to just 44 dB (which is equivalent to 7 bits) for a 10-GHz input signal. Therefore, unless the effect of sampling jitter is reduced, the performance of an ADC would be greatly limited for high-frequency input signals.

In this project, we propose a hybrid ADC with reduced sensitivity to sampling jitter. We are designing this ADC in 16-nm FinFET technology to give a proof-of-concept for improved sensitivity to the sampling clock jitter.



▲ Figure 1: A conventional discrete-time pipelined ADC with a sample-and-hold up front. This fundamentally limits the maximum achievable SNR for a given clock jitter.

## FURTHER READING

- R. van Veldhoven, "A Tri-mode Continuous-time/spl Sigma//spl Delta/modulator with Switched-capacitor Feedback DAC for a GSM-EDGE/CDMA2000/UMTS Receiver," *Solid-State Circuits Conference, Digest of Technical Papers. ISSCC. 2003 IEEE International*, pp. 60-477, 2003.

# Studies on Long-term Frequency Stability of OCS Molecular Clock

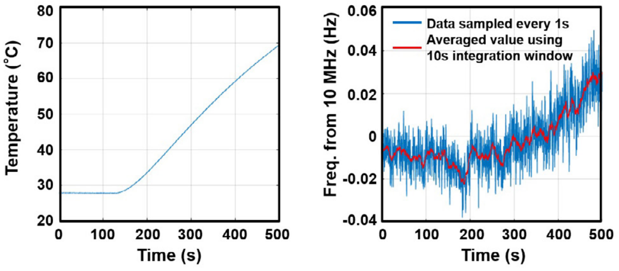
M. Kim, C. Wang, Z. Hu, R. Han

Sponsorship: Texas Instruments, NSF

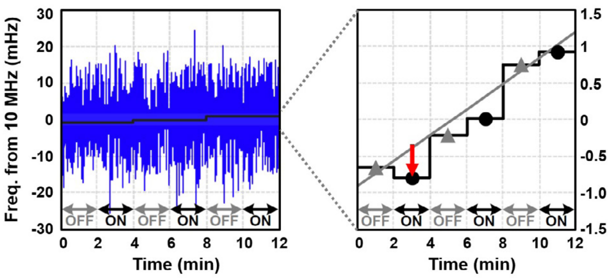
Miniature clocks with high long-term stability are critical to navigation, sensing, and communication networks. Crystal/micro-electro-mechanical systems (MEMS) oscillators with typical stability of $10^{-4}$ to $10^{-8}$ are not well suited for high-precision systems. Small-volume atomic clocks improved the stability to $10^{-11}$ to $10^{-12}$ by probing hyperfine transitions of Cs and Rb atoms at microwave frequencies, but their complicated electro-optical implementation leads to exceedingly high cost. Recently, complementary metal-oxide semiconductor (CMOS) molecular clocks emerged as a promising alternative to miniature clocks with high long-term stability. By probing the rotational lines of gaseous carbonyl sulfide (OCS) molecules at 267.530 GHz and then calibrating the clock's 10 MHz output frequency according to the measured terahertz transition frequency of OCS, the molecular clock achieved Allan deviation of $1\times10^{-11}$ with fully electronic operations. To verify the clock's robustness to external environmental variations, two critical metrics related to the long-term stability of THz OCS clocks were studied: temperature and magnetic field.

The intrinsic frequency OCS transition line is very robust to the temperature change having the temperature coefficient of a few parts per trillion per kelvin. However, the clock's sensitivity to temperature is increased by the baseline tilting, which is mainly caused by the reflection of the THz wave at the waveguide vacuum sealing window. Also, with the presence of a magnetic field, the rotational energy levels associated with different magnetic quantum numbers deviate from their degenerate value at zero field due to the Zeeman effects. While first-order Zeeman effects of all transition sub-levels maintain the symmetry of the transition line and introduce no shift, the clock shift caused by the second-order Zeeman effects is, by theory, $4\times10^{-13}$.

In our preliminary testing, the temperature coefficient of the clock is ~$1.3\times10^{-10}$/°C without ovenized temperature stabilization and temperature compensation, and the upper limit of the magnetic-induced shift in response to a 75-Gauss external magnetic field is $4\times10^{-11}$. This study verifies the molecular clock's high robustness under temperature variations and strong-magnetic conditions.



▲Figure 1: Measured instantaneous temperature of the gas cell and deviation of the clock output frequency from its mean value near 10 MHz, as the gas cell is heated up.



▲Figure 2: Measured frequency deviation with the external field turned on and off every 2 minutes.

## FURTHER READING

- C. Wang, X. Yi, J. Mawdsley, M. Kim, Z. Wang, and R. Han, "An On-chip Fully Electronic Molecular Clock based on Sub-terahertz Rotational Spectroscopy," *Nature Electron*., vol. 1, no. 7, pp. 421–427, 2018.
- M. Kim, C. Wang, Z. Hu, and R. Han, "Chip-scale Terahertz Carbonyl Sulfide (OCS) Clock: an Overview and Recent Studies on Long-term Frequency Stability of OCS Transitions," *IEEE Transactions on Terahertz Science and Technology,* 2019.

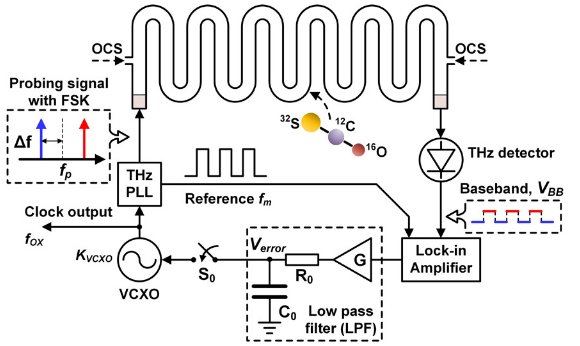
# Miniaturized, Ultra-stable Chip-scale Molecular Clock

C. Wang, X. Yi, J. Mawdsley, M. Kim, Z. Wang, R. Han
Sponsorship: NSF, MIT Lincoln Lab, MIT CICS, Texas Instruments

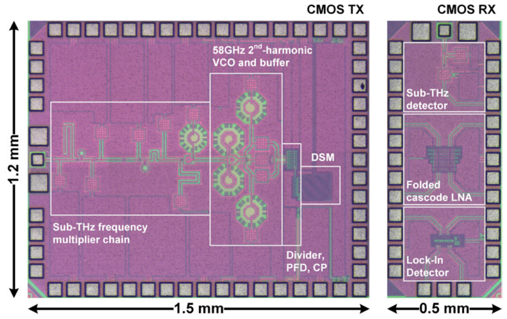
Mobile electronic devices require stable, portable, and energy-efficient frequency references (or clocks). However, current approaches using quartz-crystal and micro-electro-mechanical systems (MEMS) oscillators suffer from frequency drift. Recent advances in chip-scale atomic clocks, which probe the hyperfine transitions of evaporated alkali atoms, have led to devices that can overcome this issue, but their complex construction, cost, and power consumption limit their broader deployment. Here, we show that sub-terahertz rotational transitions of polar gaseous molecules can be used as frequency bases to create low-cost, low-power miniaturized clocks.

A molecular clock probing 231.061 GHz ($J=19\leftarrow18$) spectral line of carbonyl sulfide ($^{16}O^{12}C^{32}S$) is shown in Figure 1. Based on complementary metal–oxide semiconductor (CMOS) technology, a terahertz phase-locked loop with built-in frequency-shifting-keying (FSK), referenced to an 80-MHz crystal oscillator, and generates the probing signal. The OCS molecules are accessed within a compact WR4.3 waveguide gas cell. The relative frequency error through comparing the probing frequency and selected spectral line center is detected by envelope rectification and phase-sensitive detection in a CMOS receiver.

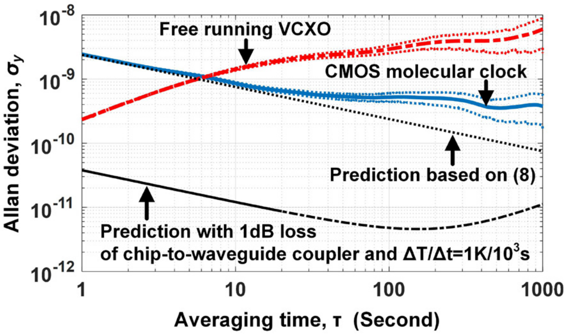
Finally, a type-I frequency locking feedback loop is established to stabilize the crystal frequency. Figure 2 shows the photograph of the CMOS molecular clock chipset. Figure 3 shows that with an averaging time of $10^3$ s, the clock stability (defined by Allan deviation) achieves $3.8\times10^{-10}$. Compared with chip-scale atomic clocks, our approach is less sensitive to external influences (temperature variation, electromagnetic field fluctuation, and mechanical vibration); offers faster frequency error compensation; and, by eliminating the need for alkali metal evaporation, offers faster start-up time and lower power consumption. Our work demonstrates the feasibility of monolithic integration of atomic-clock-grade frequency references in mainstream silicon-chip systems.



▲Figure 1: Schematic of CMOS molecular clock adopting Frequency-Shift-Keying (FSK) for spectral line center probing of OCS at 231.061GHz.



▲Figure 2: Photograph of CMOS chipset, including TX for probing signal generation and RX for demodulation.



▲Figure 3: Measured clock stability (characterized by Allan deviation) for averaging time between 1 s to 103 s. The total measurement time is 4 ×103 s. A performance prediction is plotted with reduced coupling loss.

## FURTHER READING

- C. Wang, X. Yi, J. Mawdsley, M. Kim, Z. Wang, and R. Han, "An On-chip Fully-electronic Molecular Clock Based on Sub-terahertz Rotational Spectroscopy," *Nature Electronics*, vol. 1, no. 7, pp. 1-7, Jul. 2018.
- C. Wang, X. Yi, M. Kim, Y. Zhang, and R. Han, "A CMOS Molecular Clock Probing 231.061-GHz Rotational Line of OCS with Sub-ppb Long-term Stability and 66-mW DC Power," *2018 Symposium on VLSI Circuits (VLSI)*, pp. 113-114, 2018.
- C. Wang, X. Yi, J. Mawdsley, M. Kim, Z. Hu, Y. Zhang, B. Perkins, and R. Han, "Chip-scale Molecular Clock," *IEEE J. of Solid-State Circuits (JSSC)*, vol. 54, no. 4, pp. 914-926, Apr. 2019.

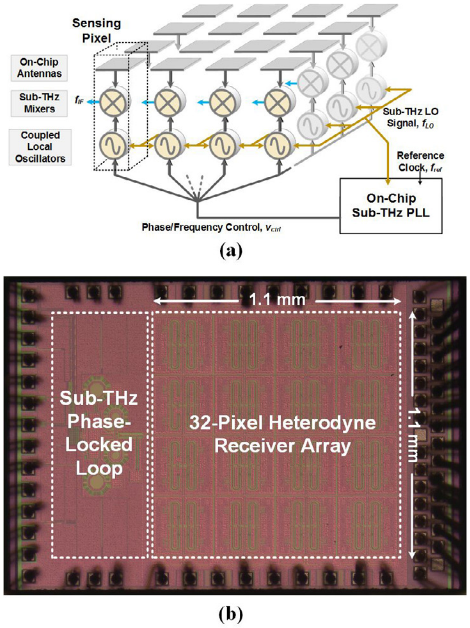
# A Dense 240-GHz 4×8 Heterodyne Receiving Array on 65-nm CMOS Featuring Decentralized Generation of Coherent Local Oscillation Signals

Z. Hu, C. Wang, R. Han

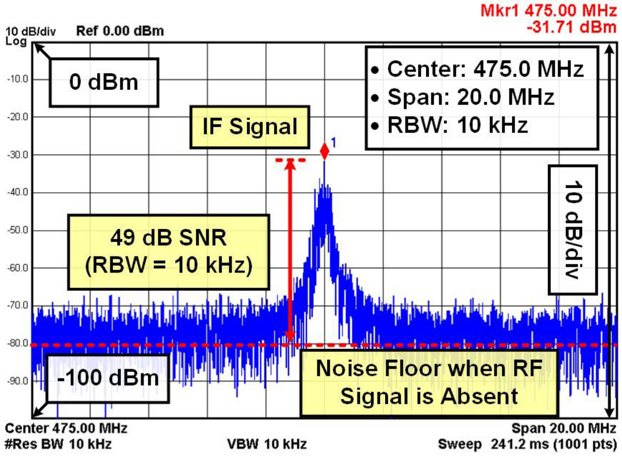
Sponsorship: NSF, MIT-SMART, TSMC

There is a growing interest in pushing the frequency of beam-steering systems towards the terahertz range, in which case narrow beams can be formed at chip scale. However, this calls for disruptive changes to traditional terahertz receiver architectures, e.g., square-law direct detector arrays (with low sensitivity and no phase information preserved) and small heterodyne mixer arrays (bulky and not scalable). Specifically, for the latter case, corporate feed (for generating and distributing the local oscillation (LO) signals), typically a necessary component, can be very lossy at large scale. Here, we report a highly scalable 240-GHz 4×8 heterodyne array achieved by replacing the LO corporate feed with a network that couples LOs generated locally at each unit. A major challenge for this architecture is that each unit should fit into a tight λ/2×λ/2 area to suppress side lobes in beamforming--it makes the integration of mixer, local oscillator, and antenna in a unit extremely difficult. This challenge is well addressed in our design, where highly compact units enable the implementation of two interleaved 4×4 phase-locked sub-arrays in an area of 1.2 mm$^2$.

The architecture of the entire array is shown in Figure 1(a). Its core component is a self-oscillating harmonic mixer (SOHM), which simultaneously (1) generates high-power LO signal and (2) down-mixes the radio frequency (RF) signal. Owing to the coupling, LOs generated in each unit are all locked to an external reference signal, so that the array is coherent. Die photo showing the placement of the array and the phase-locked loop (PLL) is given in Figure 1(b). A measured spectrum at 475-MHz (beyond the noise corner frequency) baseband signal is shown in Figure 2. The measured sensitivity (required incident RF power to achieve SNR=1 at baseband) over 1-kHz detection bandwidth is 58fW–a more than 4000× improvement over prior state-of-the-art large-scale square-law detector arrays in silicon.


Sensing Pixel
On-Chip Antennas
Sub-THz Mixers
Coupled Local Oscillators
$f_{IF}$
Sub-THz LO Signal, $f_{LO}$
Reference Clock, $f_{ref}$
On-Chip Sub-THz PLL
Phase/Frequency Control, $v_{ctrl}$
(a)
1.1 mm
1.1 mm
Sub-THz Phase-Locked Loop
32-Pixel Heterodyne Receiver Array
(b)

▲Figure 1: (a) Architecture of the entire array; (b) die photo of the chip.



▲Figure 2: Measured 475-MHz baseband spectrum.

## FURTHER READING

- C. Jiang, A. Mostajeran, R. Han, M. Emadi, H. Sherry, A. Cathelin, and E. Afshari, "A Fully Integrated 320 GHz Coherent Imaging Transceiver in 130-nm SiGe BiCMOS," *IEEE J. of Solid-State Circuits*, vol. 51, no. 11, pp. 2596-2609, 2016.
- K. Sengupta, D. Seo, L. Yang, and A. Hajimiri, "Silicon Integrated 280 GHz Imaging Chipset with 4×4 SiGe Receiver Array and CMOS Source," *IEEE Transactions on Terahertz Science and Technology*, vol. 5, no. 3, pp. 427-437, 2015.
- Z. Hu, C. Wang, and R. Han, "A 32-Unit 240-GHz Heterodyne Receiver Array in 65-nm CMOS with Array-wide Phase Locking," *IEEE J. of Solid-State Circuits*, vol. 54, no. 5, pp. 1216-1227, 2019.

# A PLL-free Molecular Clock based on Second-order Dispersion Curve Interrogation of a Carbonyl Sulfide Transition at 231 GHz

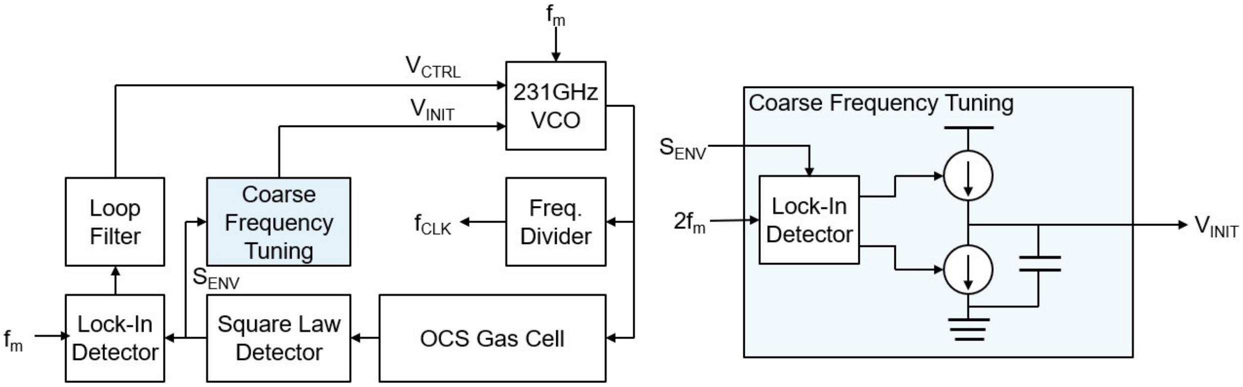
M. Kim, J. Mawdsley, C. Wang, R. Han

Sponsorship: Texas Instruments, NSF

Miniature clocks with high long-term stability are critical to navigation, sensing, and communication networks. Crystal/MEMS oscillators with a typical stability of $10^{-4}$ to $10^{-8}$ are not well suited for high-precision systems. Small-volume atomic clocks improved the stability to $10^{-11}$ to $10^{-12}$ by probing hyperfine transitions of Cs and Rb atoms at microwave frequencies, but their complicated electro-optical implementation leads to exceedingly high cost. Recently, CMOS molecular clocks that use a sub-THz spectrometer to probe the absorption lines of carbonyl sulfide molecules have emerged to achieve a low-cost miniature clock with high long-term stability.

To generate the sub-THz probing signal within the lock-range, molecular clocks require a voltage-controlled crystal oscillator (VCXO) and a fractional-N phase-locked loop (PLL) as a frequency multiplier. However, eliminating the VCXO and PLL is necessary to further reduce the power consumption and form factor. In addition, using PLL leads to degraded in-band noise because of the high-frequency multiplication factor of the PLL.

This work proposes a molecular clock without a VCXO and a PLL. A sub-THz voltage-controlled oscillator (VCO) is directly controlled by a negative feedback loop and then locked to the center of the absorption line. For frequency initialization and coarse frequency tuning, the second-harmonic dispersion curve of the absorption line profile was utilized instead of a PLL. Since the polarity of the second-harmonic dispersion curve is positive only when the frequency of the probing signal is very close to the absorption line, detection of the absorption line does not depend on the signal strength. Also, the second harmonic signal is robust against spectral baseline variations. By eliminating the VCXO and PLL from the loop and using the proposed coarse frequency tuning method, the noise performance of the proposed molecular clock is expected to improve, and further miniaturization of an ultra-stable clock can be achieved.



▲Figure 1: Block diagram of the proposed architecture. A coarse frequency tuning circuit sets the $f_{VCO}$ within the lock range by changing $V_{INIT}$ while ensuring the 2nd harmonic of $f_m$ in the square law detector's output ($S_{ENV}$) is positive.

## FURTHER READING

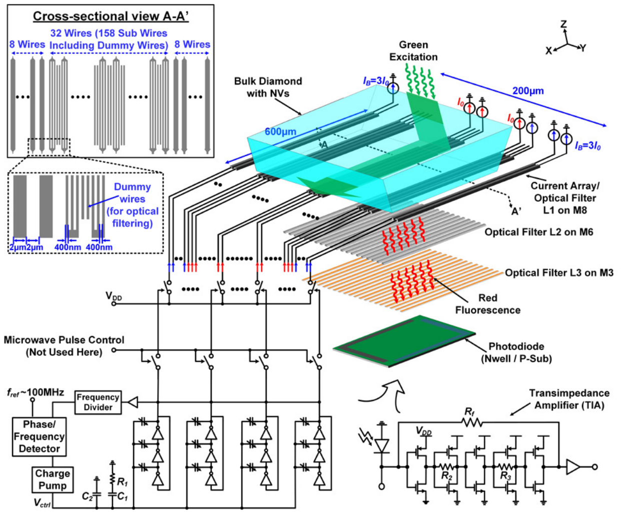
- C. Wang, X. Yi, J. Mawdsley, M. Kim, Z. Wang, and R. Han, "An On-chip Fully Electronic Molecular Clock based on sub-Terahertz Rotational Spectroscopy," *Nature Electron*., vol. 1, no. 7, pp. 421-427, 2018.
- C. Wang, X. Yi, M. Kim, Y. Zhang, and R. Han, "A CMOS Molecular Clock Probing 231.061-GHz Rotational Line of OCS with sub-ppb Long-term Stability and 66-mW DC Power," *Proc. Symp. VLSI Technol. Circuits*, pp. 113-114, Honolulu, HI, Jun. 2018.

# Chip-scale Scalable Ambient Quantum Vector Magnetometer in 65-nm CMOS

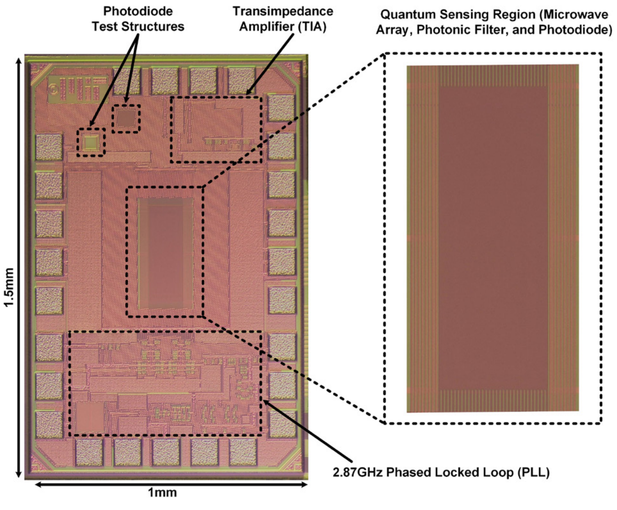
M. I. Ibrahim, C. Foy, D. R. Englund, R. Han
Sponsorship: NSF

Room-temperature coherent spin state control and detection of nitrogen-vacancy (NV) centers in diamond have enabled magnetic field sensing with high sensitivity and spatial resolution. However, current NV sensing apparatuses use bulky off-the-shelf discrete components, which increases the system scale and limits practical applications. To address this challenge, we developed a hybrid complementary metal-oxide semiconductor (CMOS)-NV platform to shrink this spin-based magnetometer to chip scale. In this work, we present a fully integrated CMOS-NV quantum sensor fabricated using a 65-nm CMOS process.

Magnetic field sensing is accomplished by the excitation and detection of the spin states of the NV. The frequency of the spin states is determined through optically detected magnetic resonance (ODMR). The magnetic field is proportional to the frequency splitting of the spin states (2.8 MHz/Gauss). Our CMOS-NV magnetometer system is composed of (i) a microwave generation and delivery system to control the NV's spin states and (ii) an optical system for the readout of spin states. We implement a highly scalable microwave delivery structure, which consists of an array of current-carrying conductors. We control the current flowing in each conductor to achieve a uniform magnetic-field profile. This uniform field enables coherent driving of the NV centers, which enhances the sensitivity. The on-chip optical readout follows the microwave manipulation of the NV spin ensembles. We implemented a CMOS-compatible, three-layer grating structure to filter out the green excitation. The filter reduces the shot noise of the photo-detector caused by the input green laser. The Talbot effect is used in the filter, where we place layers of gratings with positions aligned with the maxima and minima of the green and the red diffraction patterns generated from the preceding grating layer. We detect the spin-dependent red fluorescence of the NV centers using on-chip N-Well/P-sub photodiode. This work presents a hybrid NV-CMOS platform that can perform coherent spin control and readout of the NV ensemble's spin state: a highly advanced, scalable, and compact platform for quantum sensing.



▲Figure 1: Block diagram of the NV-based magnetic sensor in 65-nm CMOS.


Photodiode Test Structures
Transimpedance Amplifier (TIA)
Quantum Sensing Region (Microwave Array, Photonic Filter, and Photodiode)
1.5mm
1mm
2.87GHz Phased Locked Loop (PLL)

▲Figure 2: Chip die photo.

## FURTHER READING

- M. I. Ibrahim, C. Foy, D. R. Englund, and R. Han, "A Scalable Quantum Magnetometer in 65-nm CMOS with Vector-field Detection Capability," *IEEE Intl. Solid-State Circuit Conf. (ISSCC)*, San Francisco, CA, 2019.
- M. I. Ibrahim, C. Foy, D. Kim, D. R. Englund, and R. Han, presented at "Room-temperature Quantum Sensing in CMOS: On-chip Detection of Electronic Spin States in Diamond Color Centers for Magnetometry," *IEEE VLSI Circuits Symposium*, Honolulu, HI, 2018.
- "Nanoscale Imaging Magnetometry with Diamond Spins under Ambient Conditions," *Nature*, 455, no. 7213, pp. 648-651, 2008.

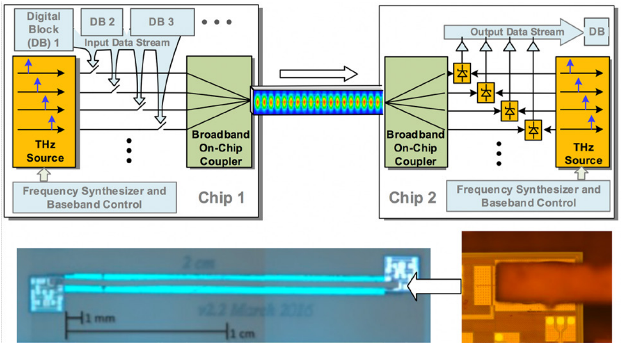
# Broadband Inter-chip Link using a Terahertz Wave on a Dielectric Waveguide

J. Holloway, R. Han

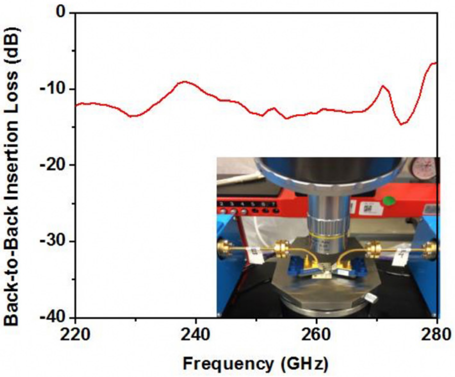
Sponsorship: Intel, Office of Naval Research, MIT Lincoln Laboratories

The development of data links between different microchips of an onboard system has encountered a speed bottleneck due to the excessive transmission loss and dispersion of the traditional inter-chip electrical interconnects. Although high-order modulation schemes and sophisticated equalization techniques are normally used to enhance the speed, they also lead to significant power consumption. Silicon photonics provides an alternative path to solve the problem, thanks to the excellent transmission properties of optical fibers; however, the existing solutions are still not fully integrated (e.g., off-chip laser source) and normally require process modification to the mainstream complementary metal-oxide semiconductor (CMOS) technologies. Here, we aim to utilize a modulated THz wave to transmit broadband data. Similar to the optical link, the wave is confined in dielectric waveguides, with sufficiently low loss (~0.1dB/cm) and bandwidth (>100GHz) for board-level signal transmission (Figure 1). In commercial CMOS/BiCMOS platforms, we have previously demonstrated high-power THz generation with modulation, frequency conversion, and phase-locking capabilities.

In addition, a room-temperature Schottky-barrier diode detector (in 130-nm CMOS) with <10pW/Hz$^{1/2}$ sensitivity (antenna loss excluded) is also reported. The prototype data link will leverage these techniques to achieve a ~100Gbps/channel transmission rate with <1pJ/bit energy efficiency. As the first step of this project, we have designed a new broadband chip-to-fiber THz wave coupler, passive channelizers, broadband THz modulators, and sub-harmonic carrier generation. In contrast to previous couplers using off-chip antennas, our THz coupler is entirely implemented using the metal backend of a CMOS process and requires no post-processing (e.g., wafer thinning). The structure is also fully shielded, which prevents THz power leakage into the silicon substrate. Conventional on-chip radiators using ground shield work are the resonance type (e.g., patch antenna) and have only <5% bandwidth. In comparison, our design is based on a traveling-wave, tapered structure, which supports broadband transmission. A proof-of-concept is shown in Figure 1: two on-chip couplers are connected with a 2-cm waveguide using Rogers 3006 dielectric material. The entire back-to-back setup exhibits only ~11dB insertion loss across over 60-GHz bandwidth (Figure 2). Additionally, our on-chip and on-interposer channelizers provide a compact and efficient means of reducing ISI while combining incoherent parallel data streams.


Digital Block (DB) 1
DB 2
DB 3
• • •
Input Data Stream
THz Source
Broadband On-Chip Coupler
Frequency Synthesizer and Baseband Control
Chip 1
Output Data Stream
DB
Broadband On-Chip Coupler
THz Source
Chip 2
Frequency Synthesizer and Baseband Control
2 cm
1 mm
1 cm

▲Figure 1: (Top) High-speed, energy-efficient inter-chip transmission using guided THz wave. (Bottom) A test structure with back-to-back THz integrated couplers separated by a 2-cm dielectric waveguide.



▲Figure 2: The measured back-to-back insertion loss using a two-port network analyzer in the WR-3 band.

## FURTHER READING

- C. Yeh, F. Shimabukuro, and P. H. Siegel, "Low-loss Terahertz Ribbon Waveguides," *Applied Optics*, vol. 44, no. 28, pp. 5937-5946, Oct. 2005.
- R. Han, C. Jiang, A. Mostajeran, M. Emadi, H. Aghasi, H. Sherry, A. Cathelin, and E. Afshari, "A 320GHz Phase-locked Transmitter with 3.3mW Radiated Power and 22.5dBm EIRP for Heterodyne THz Imaging Systems," presented at *IEEE Int. Solid-State Circuit Conf. (ISSCC)*, San Francisco, CA, 2015.
- R. Han, Y. Zhang, Y. Kim, D. Kim, H. Shichijo, E. Afshari, and K. K. O, "Active Terahertz Imaging using Schottky Diodes in CMOS: Array and 860-GHz Pixel," *IEEE J. of Solid-State Circuits (JSSC)*, vol. 48, no. 10, Oct. 2013.

# Low-energy Current Sensing with Integrated Fluxgate Magnetometers

P. Garcha, V. Schaffer, B. Haroun, S. Ramaswamy, J. Wieser, D. Buss, J. H. Lang, A. P. Chandrakasan

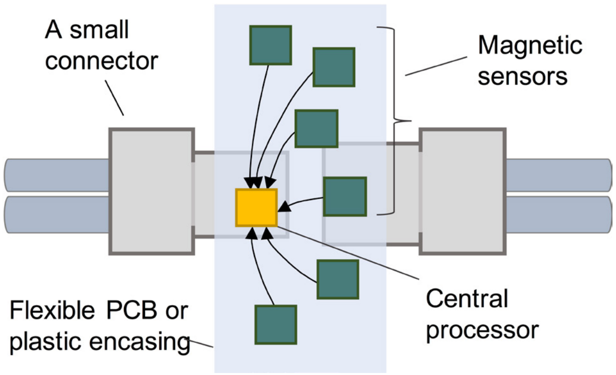
Sponsorship: Texas Instruments

The ability to sense current is crucial to many industrial applications including power line monitoring, motor controllers, battery fuel gauges, etc. We are developing smart connectors with current sensing abilities for use in the industrial Internet of things (IoT). These connectors can be used for 1) power quality management to measure real power, reactive power, and distortion and 2) machine health monitoring applications for continuous monitoring, control, prevention, and diagnosis.

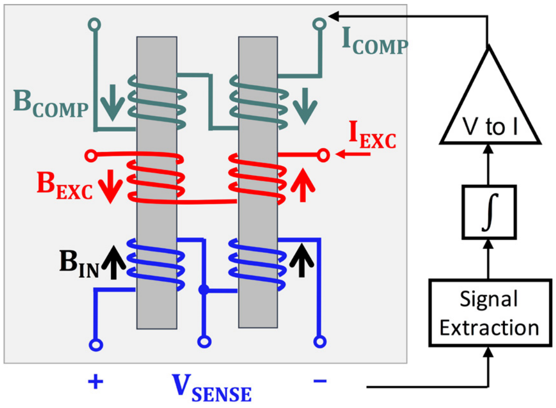
At the system level, the smart connectors need to 1) measure AC, DC, and multiphase currents; 2) reject stray magnetic fields; and 3) detect impending connector failure. On the sensor level, they need high accuracy and performance and a small area to fit inside the outer plastic encasing of the connectors. Therefore, the sensors must not use large external magnetic cores as field concentrators.

A good system solution is to use an array of integrated fluxgate (FG) sensors (Figure 1), which offer a better alternative than Hall/magneto-resistive sensors and shunt-sensing in terms of dynamic range (~10^5), sensitivity (200 V/T), linearity (0.1%), low temperature drift, and inherent isolation. But high power consumption is a drawback for FG sensors. FG sensors work by driving magnetic cores in and out of saturation and sensing the resulting voltage difference (Figure 2). They achieve high linearity by balancing the external magnetic fields within the core with an equivalent compensation current, which can be quite power-hungry. We need to reduce the energy needs of the FG sensors so they can be used in an array, especially in energy-constrained environments.

We propose a low-energy front-end design with bandwidth scalability and lower energy per measurement for FG sensors. We use a mixed-signal architecture with quick convergence techniques to enable duty cycling from >50 kHz bandwidth for machine health monitoring to <1 kHz for power quality management.



▲Figure 1: Proposed contactless current sensing approach for smart connectors, including a central processor and multiple sensors to measure multiphase currents and reject disturbances. The system can be wrapped around or plugged into the connectors.



▲Figure 2: FG sensor with two magnetic cores and three sets of coils: excitation, sense, and compensation. When excited, one core saturates before the other, producing voltage $V_{SENSE}$ as a function of ($B_{IN}$-$B_{COMP}$). Compensation provides feedback to improve linearity.

FURTHER READING

- HARTING, 04. Industrial Connector Han e-Catalogue [Online], Available: https://www.harting.com/sites/default/files/2018-02/DevCon_07_5_E_Kap04_Industrial-Connectors-Han.pdf, Feb. 2018.
- M. F. Snoeij, V. Schaffer, S. Udayashankar, and M. V. Ivanov, "Integrated Fluxgate Magnetometer for use in Isolated Current Sensing," *IEEE J. Solid-State Circuits*, vol. 51, no. 7, pp. 1684–1694, Jul. 2016.

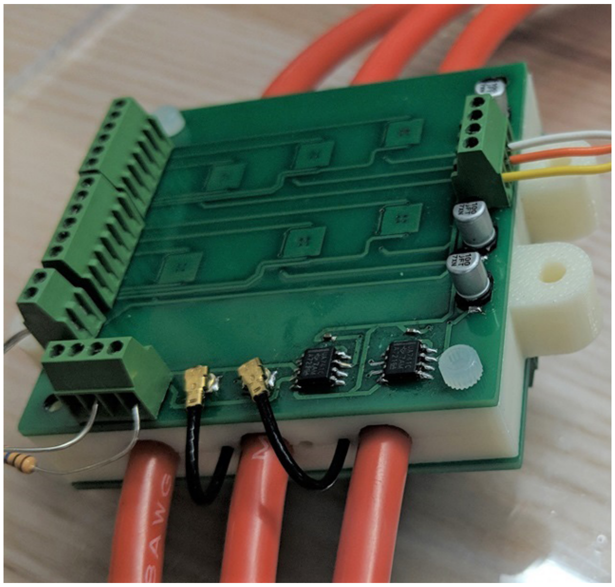
# Contactless Current and Voltage Detection using Signal Processing and Machine Learning

A. Casallas, J. H. Lang
Sponsorship: HARTING, Texas Instruments

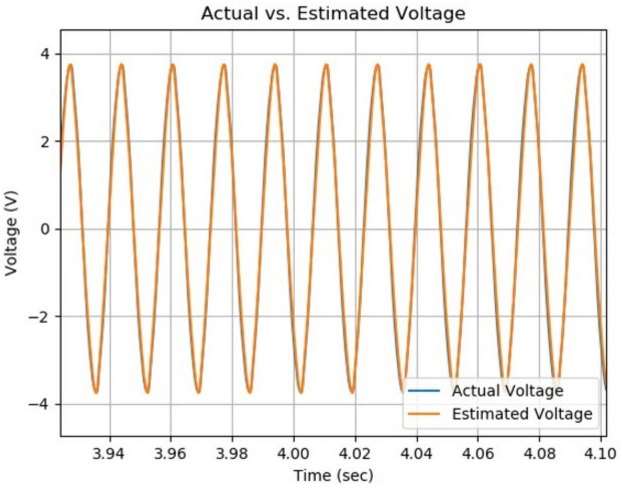
Measuring current and voltage in electrical systems is a critical task in industrial environments and can be used to monitor power quality and machine and process performance. Easily retrofitted contactless measurements are preferred, but they can require difficult installations and bulky hardware. In contrast, we are developing a contactless clip-on sensor that will estimate voltage and current in three-phase power cables. Our goal is to create a measurement system that uses less hardware than present state-of-the-art solutions while maintaining a high level of accuracy.

Current is estimated using an array of magnetic field sensors embedded in a yoke that fits around the cables, as shown in Figure 1. The measurements are filtered to remove magnetic fields from external sources, such as adjacent cables or eddy currents. This filtering employs a Best Linear Unbiased Estimate of cable currents that is based on a covariance matrix calculated from a probabilistic model of external magnetic fields detected by the sensor array. Additionally, we are using collected data to train neural networks and explore whether machine learning can generate a better estimate. To estimate voltage, we employ guarded electrodes in the yoke that fit snugly against the cables. We then sense cable voltage capacitively coupled to the electrodes and use a physical model of the electrode system to estimate the voltage differences between cables. A voltage estimate example is shown in Figure 2.

At present, our system can estimate voltage with an error of less than 1% and current with an error of less than 2%, even in the presence of electric and magnetic field interference. This performance is comparable to currently used contactless detection systems but uses significantly less hardware and should thus be less costly to manufacture. Furthermore, since our estimates produce full current and voltage waveforms, we can calculate quantities such as instantaneous power and power quality.



▲ Figure 1: A photograph of the yoke with magnetic field sensors to estimate current and active shielding hardware to estimate voltage.



▲ Figure 2: An example of an estimated voltage waveform displayed over a true voltage waveform measured using electrical contacts.

## FURTHER READING

- G. D'Antona, L. Di Rienzo, R. Ottoboni, and A. Manara, "Processing Magnetic Sensor Array Data for AC Current Measurement in Multiconductor Systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 50, no. 5, pp. 1289-1295, 2001.
- L. Di Rienzo and Z. Zhang, "Spatial Harmonic Expansion for use with Magnetic Sensor Arrays," *IEEE Transactions on Magnetics*, vol. 46, no. 1, pp. 53-58, 2010.

# SHARC: Self-healing Analog Circuits with RRAM and CNFETs

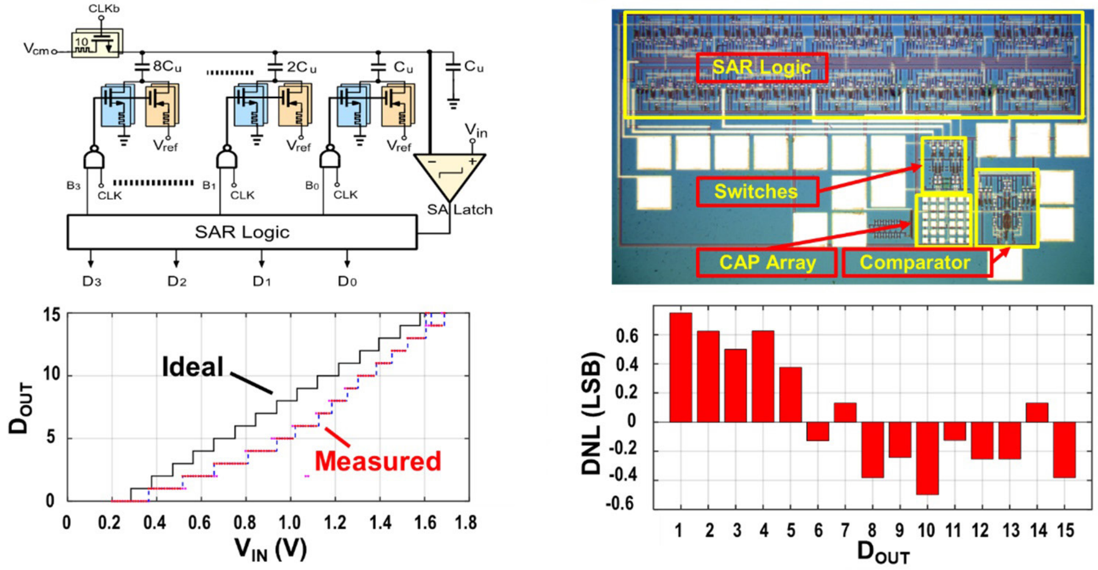
A. Amer, R. Ho, G. Hills, A. P. Chandrakasan, M. M. Shulaker

Sponsorship: Analog Devices, Inc., DARPA 3DSoC

Next-generation applications require processing of a massive amount of data in real time, exceeding the capabilities of electronic systems today. This has spurred research in a wide range of areas: from new devices to replace silicon field-effect transistors (FETs) to improved circuit implementations to new system architectures with dense integration of logic and memory. However, isolated improvements in any one area are insufficient. Rather, enabling these next-generation applications will require combining benefits across all levels of the computing stack: leveraging new devices to realize new circuits and architectures.

For instance, carbon nanotube (CNT) field-effect transistors (CNFETs) for logic and resistive random-access memory (RRAM) for memory are two promising emerging nanotechnologies for energy-efficient electronics. However, CNFETs suffer from inherent imperfections (such as of metallic CNTs (m-CNTs)), which have prohibited realizing large-scale CNFET circuits in the past. M-CNTs create shorts between the CNFET source and drain, which translates into (1) a 100x intrinsic gain reduction for analog circuits causing the failure of the whole system and (2) high power consumption and degraded noise margin for digital circuits. This work proposes a circuit design technique (called self-healing analog circuitry with RRAM correction (SHARC)) that integrates and combines the benefits of both CNFETs and RRAM to realize three-dimensional circuits that are immune to m-CNTs. Non-volatile RRAMs are 3D-integrated with CNFETs, whereas each CNFET is split into multiple minimum-width FETs (i.e., "sub-CNFETs"), with a RRAM cell in series fabricated directly under (or over) the source or drain contact of each sub-CNFET.

SHARC is a non-volatile technique that self-reconfigures the circuit by programming RRAMs. The sub-FETs including m-CNTs become connected in series to reset high-resistance RRAM that effectively removes those sub-CNFETs from the circuit, while CNFETs containing only semiconducting CNTs are connected in series with set low-resistance RRAM. Leveraging this technique, we experimentally demonstrate the first and largest CMOS CNFET mixed-signal systems robust to m-CNTs (by implementing SHARC in amplifiers and switches) such as a 4b-DAC and 4b-SAR ADC. SHARC can also be combined with additional existing circuit techniques to further improve performance for very-large-scale integrated circuits.



▲ Figure 1: 4-Bit SAR DAC with SHARC. (top) Schematic and die photo. (bottom) Measured characteristics show the ADC behavior with offset (35mV), non-linearity, and gain error whereas the DNL (-0.5 LSB→ 0.75 LSB).

## FURTHER READING

- Aya G. Amer, et al. "29.8 SHARC: Self-healing Analog with RRAM and CNFETs," *2019 IEEE International Solid-State Circuits Conference-(ISSCC), IEEE*, 2019.