# 2022 Microsystems Annual Research Report

## CICS Edition

MTL MICROSYSTEMS TECHNOLOGY LABORATORIES

MIT.nano

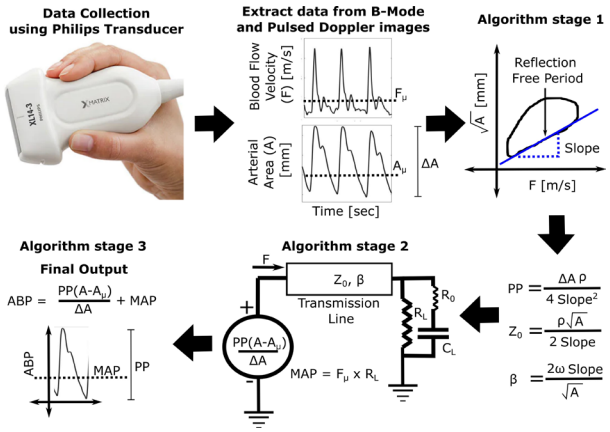
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

# Absolute Blood Pressure Waveform Monitoring Using an Ultrasound Probe

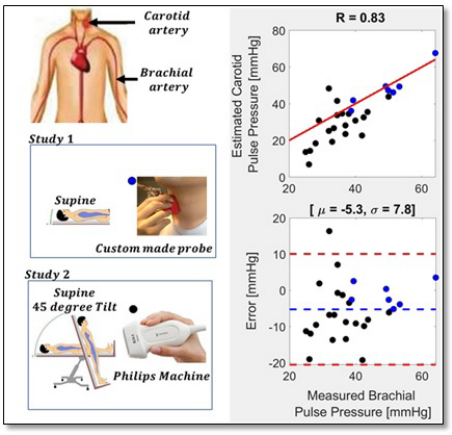
A. Chandrasekhar, A. Aguirre, H.-S. Lee, C. G. Sodini
Sponsorship: MEDRC-Philips, Analog Devices Inc., CICS

Accurate measurement of the absolute blood pressure (ABP) waveform assists clinical decision-making as it helps physicians titrate the cardiovascular therapy for a patient. In an intensive care unit (ICU), physicians use an invasive radial catheter to measure BP, whereas, outside an ICU, one may resort to isolated spot measurement of the BP via a brachial arm cuff device. These cuff devices measure only the systolic and diastolic values of the BP waveform, and unlike an arterial catheter used in an ICU, they do not output the shape of the ABP waveform. Morphology of the ABP waveform is significant as it reflects the hemodynamics of the underlying vasculature, and hence there is a need for a non-invasive and easy-to-use device that can output the ABP waveform. Ultrasound-based devices are a feasible alternative for monitoring the BP waveform as these devices can accurately measure pressure-dependent parameters like the blood flow velocity and arterial diameter waveforms. In this project, we are developing an algorithm (see Figure 1) to convert ultrasound data into an ABP waveform, and in this report, we present the preliminary results on estimating pulse pressure (PP) from the ultrasound data.

Two studies were performed to investigate the proposed PP estimation algorithm (See Algorithm Stage 1 in Fig. 1). In study 1, signals illustrated in Figure 1 were recorded from the carotid artery using a custom-designed ultrasound probe while the subject was supine, whereas, in study 2, the above-mentioned signals were recorded using a Philips ultrasound-transducer (XL-143) while the subject rested supine and in tilted posture. Gold standard PP was measured from the brachial artery using an Omron BP monitor as a reference. The Bland-Altman plot in Figure 2 shows that the estimated PP can track the gold standard PP values.




▲ Figure 1: Step by step algorithm to estimate pulse pressure and ABP waveform from ultrasound signals.




▲ Figure 2: Pulse pressure estimated via ultrasound signals recorded at the carotid artery.

## FURTHER READING


- J. Seo, "A Non-invasive Central Arterial Pressure Waveform Estimation System Using Ultrasonography for Real-time Monitoring," Dissertation, Massachusetts Institute of Technology, Cambridge, 2018.

# Modeling the Arterial System to Improve Ultrasound Measurements of Hemodynamic Parameters

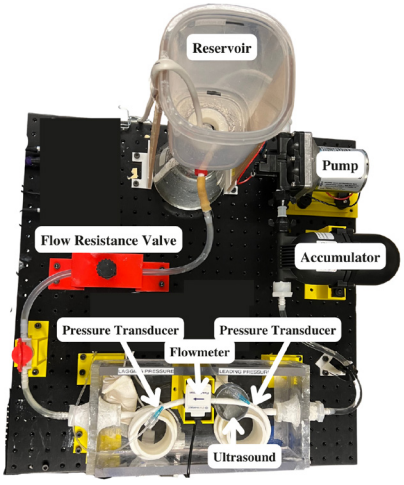
J. Harabedian, A. Chandrasekhar, C. G. Sodini
Sponsorship: Analog Devices

One of the most crucial parameters for monitoring cardiovascular disease risk is one's arterial blood pressure (ABP). Clinicians use a radial arterial catheter to measure ABP in an intensive care unit (ICU). Although this method is considered the gold standard, its invasive nature makes it undesirable and inaccessible outside an ICU. One solution to this problem is to take advantage of ultrasonic measurements, which are noninvasive and extremely accessible. However, developing an algorithm to convert ultrasound data into a legitimate ABP waveform requires an extensive amount of patient data. The limitation is that this data is difficult to obtain and impossible to fully control.

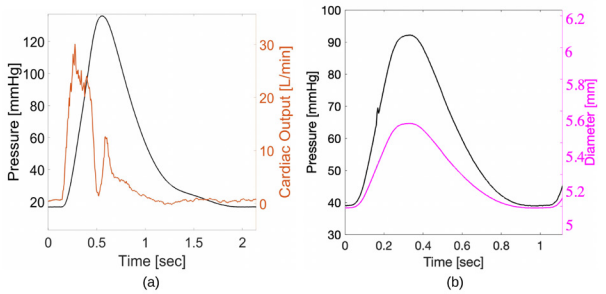
The solution presented here is to use a flow phantom: a physical, hydraulic system that mimics arterial blood flow. The phantom provides pressure waveforms, which come directly from a catheterized tube, and flow velocity waveforms, from an ultrasonic flow meter, that closely match the morphology of patient data. Developing a physical model of the arterial system allows for control of parameters that are considered uncontrollable in humans (e.g., arterial compliance, cardiac output, critical closing pressure) and enables data collection for a number of parameter combinations that would otherwise be unobtainable.

A flow phantom can be made using a pump, accumulator, compliant tubing, flow control valve, and a reservoir, representing the heart, large arterial compliance, large artery, arteriole resistance, and the ground pressure, respectively. To collect data from the phantom there are two pressure transducers, a flowmeter and the ultrasound device. Figure 1 shows the setup of the entire system. The pressure transducers and the flowmeter can collect control measurements that can be tested against the ultrasound device and the developed ABP estimation algorithm. Figure 2 shows example outputs from these measurement devices.

Experimental validation shows that the flow phantom does in fact mimic the hemodynamic behavior of arterial blood flow. This was confirmed by controlling various parameters of the system (e.g., flow resistance, ground pressure, cardiac output) and comparing its response against known hemodynamic responses.




▲ Figure 1: Physical flow phantom system, consisting of the pump, accumulator, compliant tubing, flow control valve and reservoir. There are also the measure-ment devices: two pressure transducers, one flowmeter and in-house ultrasound device.



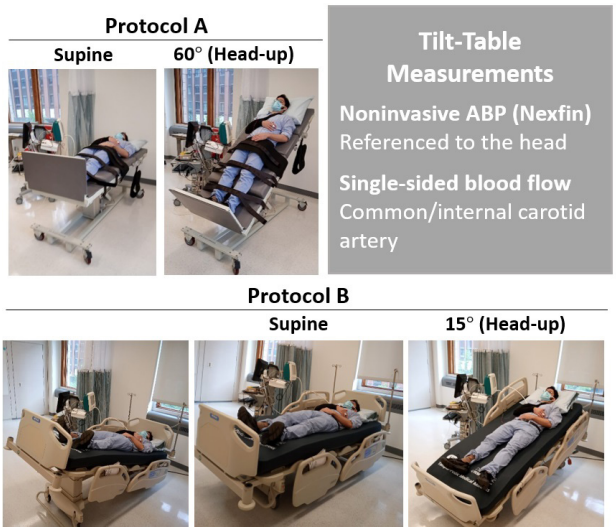

▲ Figure 2: Examples of typical measurements from flow phantom. 2a shows pressure and volumetric flow, with .5 Hz heart beat and pump on for .3 seconds. 2b shows pressure and diameter, with 1 Hz heartbeat and pump on for .2 seconds.

# Model-based Noninvasive Intracranial Compliance and Vascular Resistance Estimation

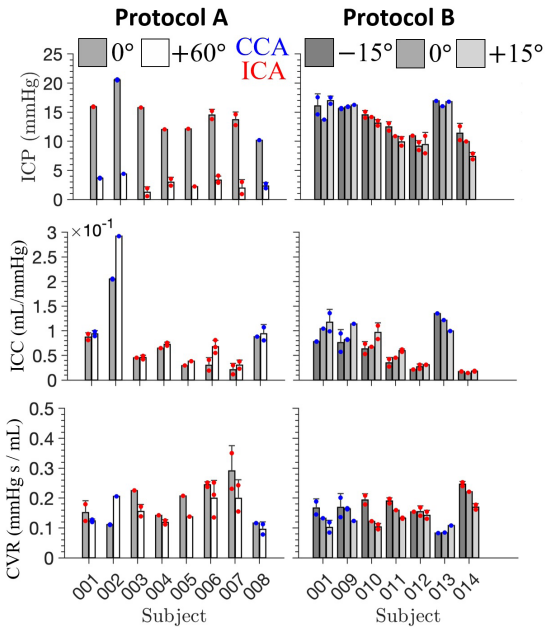
S. M. Imaduddin, C. G. Sodini, T. Heldt
Sponsorship: Analog Devices, Inc. via MIT Medical Electronic Device Realization Center

Existing neuromonitoring methods used for patients with severe head injury tend to be highly invasive and carry a risk of tissue damage and infection. In particular, fluid infusion/withdrawal studies via indwelling catheters are needed to determine intracranial compliance (ICC) – an index of the propensity of rise in intracranial pressure (ICP) in response to changes in cranio-spinal volume. Despite their potential to serve as early indicators of intracranial hypertension, ICC measurements are rarely performed owing to time-consuming, invasive measurement protocols. In addition, measurements of cerebrovascular resistance (CVR) to blood flow are useful in assessing cerebral autoregulation and tracking pathological vascular narrowing such as in moyamoya disease. Like ICC, however, CVR is not regularly obtained at the bedside as the requisite measurements – arterial blood pressure (ABP), cerebral arterial blood flow (CBF), and ICP – are rarely monitored simultaneously.

We previously developed a noninvasive, model-based approach for ICP estimation. Recently, we augmented our approach to additionally estimate ICC and CVR. In particular, subjects' ABP and CBF are related to the ICP, ICC, and CVR via a Windkessel-like model. Measurements of the ABP and CBF are then used to estimate the clinically interpretable model parameters in a noninvasive, patient-specific fashion. Ultrasound-based CBF measurements were made in extracranial (common/internal carotid) arteries. Vessel diameters were estimated with B-mode images and combined with color flow velocity measurements to yield the CBF. Tilt-table studies were carried out to validate the proposed method. We found that our system successfully tracked tilt-induced changes in ICP, ICC, and CVR, paving the way towards convenient and safe neuromonitoring across a wide spectrum of pathologies, patient ages, and disease severities.




▲ Figure 1: Illustration of tilt-table protocols for method validation. Two protocols were established. Protocol A involved transitions from supine to 60° head-up position. Protocol B involved both head-up and head-down transitions of 15°, respectively. Acquired measurements are also listed.Closed electronics box with force and accelerometer analog inputs from the casing and data streaming to a tablet for data collection and display. Bottom: Electronics box components consisting of the DC power source of two series 9V batteries, the force signal amplifier, and the multi-channel analog-to-digital converter.



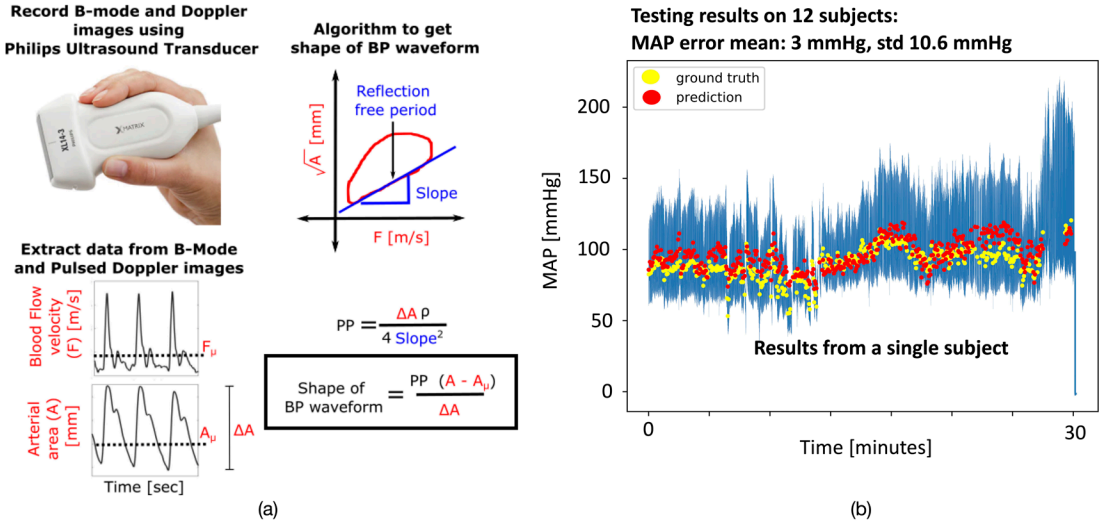

▲ Figure 2: Parameter estimation summary. Results obtained with recordings at common carotid (CCA) and internal carotid (ICA) arteries are shown in blue and red, respectively. ICP estimates decreased as subjects progressively moved to head-up positions. ICC estimates increased while CVR estimates decreased on average.

# Absolute Blood Pressure Measurement using Machine Learning Algorithms on Ultrasound- based Signals

H. Wang, A. Chandrasekhar, J. Seo, A. Aguirre, S. Han, C. G. Sodini, H.-S. Lee
Sponsorship: MIT J-Clinic, Philips, Analog Devices, MIT-IBM Watson AI Lab, NSF CAREER Award

In an intensive care unit (ICU), physicians can use an invasive arterial catheter to measure the blood pressure (BP) waveform with high resolution. In non-ICU settings, arterial catheters are not used, and clinicians must rely upon isolated spot measurements from a non-invasive arm-cuff device that cannot measure the absolute BP waveform. In this project, we are developing an algorithm to convert data from an ultrasound-based device to absolute BP waveforms. Such a device may offer a quantitative method to perform rapid hemodynamic profiling of patients in an emergency room, step down clinical ward, or outpatient clinic who cannot undergo invasive BP measurements.

We propose a non-invasive way to get BP waveform with blood flow velocity and arterial area obtained from non-invasive ultrasound signals (Figure 1a). One key drawback of the ultrasound-based device is that the output BP waveform has an arbitrary reference, so we have to estimate the mean arterial pressure (MAP) and leverage transmission line model to calculate the absolute BP value. Hence, we propose to use a machine learning model containing transformer encoder layers to regress the MAP accurately. The input features are flow velocity and the shape of BP waveform. Since the number of subjects is limited in the training set, we propose to use contrastive loss to guide the feature extraction and improve generalization. The contrastive loss encourages the features of beats of the same subject to be similar. When we enlarge the contrastive loss , the feature vector will be trained to contain as little subject-specific information as possible. Therefore, the model can generalize better to unseen subjects. On a collected dataset, the proposed method improves the MAP error standard deviation from baseline 10.6 mmHg (only training the MAP regressor) to 9.2 mmHg. Our algorithm has large potential to make affordable BP measurements accessible to everyone.




▲ Figure 1: (a) The whole pipeline of using machine learning-based algorithms to get BP waveforms from ultrasound data, and (b) MAP prediction results for a single subject.

## FURTHER READING

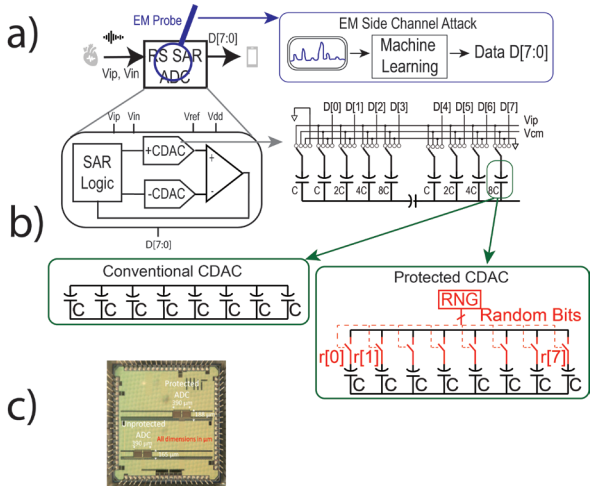

- V. Novak, and L. Mendez, "Cerebral Vasoregulation in Diabetes" (version 1.0.0), *PhysioNet*, 2020. https://doi.org/10.13026/m40k-4758.

# Randomized Switching Successive Approximation Register (RS-SAR) ADC Protections for Power and Electromagnetic Side Channel Security

M. Ashok, E. V. Levine, A. P. Chandrakasan
Sponsorship: MITRE Innovation Program, NSF Graduate Research Fellowship Program, MathWorks Engineering Fellowship

Analog to digital converters (ADCs) are necessary in most Internet of Things (IoT) devices, to link the physical analog world to digital computation. In many of these applications, the ADC is processing sensitive data such as biomedical signals or private conversations, which should not be accessible to an attacker. Physical side channel attacks (SCAs) have been used to reconstruct information processed within digital integrated circuits in a variety of applications, through power or electromagnetic (EM) traces. These attacks correlate unintentional leakage of information in the current consumption to the operations and data processed by a circuit, allowing for complete reconstruction of private data, as seen in Figure 1a. Specifically, EM SCAs allow fully non-invasive and localized attacks, by simply placing a probe above a packaged chip, eliminating the effectiveness of some global protections.

In this work, we propose the RS-SAR ADC, which decorrelates the data processed by the ADC from the power and EM side channel leakage. In the capacitive DAC, the parallel unit capacitors corresponding to the more significant bits are independently controlled with random bits, as shown in Figure 1b. This controlrandomization leads to variable timing of the binary search conversion, eliminating the attacker's ability to determine which of the digital data bits the various measured current spikes correspond to. When tested on a chip fabricated in TSMC 65-nm complementary metal-oxide-semiconductor (CMOS) (Figure 1c, provided through TSMC University Shuttle), the protected ADC has 82x the attack error of the unprotected ADC for power SCA and 32x the attack error for EM SCA.




▲ Figure 1: (a) Motivation for SAR ADC side channel security, along with (b) proposed RS-SAR architecture and (c) die micrograph. (From first Reading).

## FURTHER READING

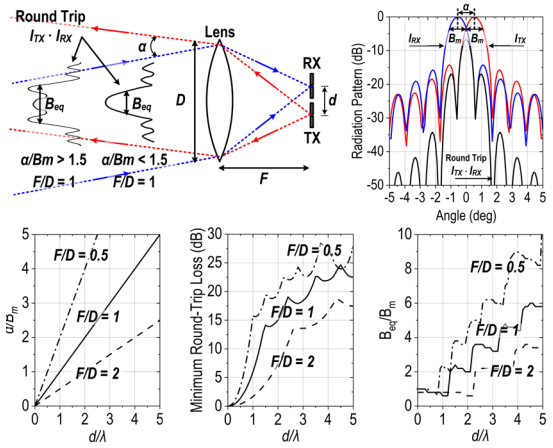

- M. Ashok, E. V. Levine, and A. P. Chandrakasan, "Randomized Switching SAR (RS-SAR) ADC Protections for Power and Electromagnetic Side Channel Security," presented at *IEEE Custom Integrated Circuits Conference*, Newport Beach, CApp. 1-2, 2022.
- T. Jeong, A. P. Chandrakasan, and H. Lee, "S2ADC: A 12-bit, 1.25MS/s Secure SAR ADC with Power Side-Channel Attack Resistance," *IEEE Custom Integrated Circuits Conference*, pp. 1-4, 2020.

# A 140-GHz FMCW TX/RX-Antenna-Sharing Transceiver with Low-Inherent-Loss Duplexing and Adaptive Self-Interference Cancellation

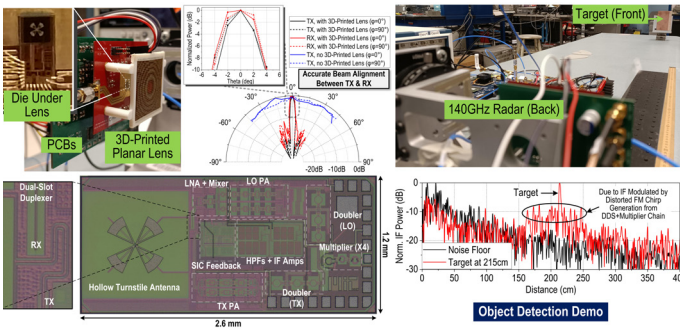
X. Chen, M. I. W. Khan, X. Yi, X. Li, W. Chen, J. Zhu, Y. Yang, K. E. Kolodziej, N. M. Monroe, R. Han

High-resolution integrated radars are crucial in today's automotive, vital sign, and security-sensing applications. Compared to radars operating in the microwave/low-millimeter-wave and optical regimes, the sub-terahertz/terahertz (sub-THz/THz) spectrum shows great opportunities in both high-resolution and all-weather radar imaging capabilities. For isolation between the radar transmitter (TX) and receiver (RX), a bistatic configuration with separate TX and RX antenna positions is commonly adopted. However, in non-MIMO high-angular resolution systems, the radar transceiver should pair with a large lens/reflector for beam collimation. The bistatic arrangement then causes severe misalignment between the peaks of TX and RX beam patterns, as in Figure 1. Radar transceivers with a shared TX and RX antenna interface, or monostatic configuration, are therefore required in this scenario. Prior monostatic radars adopt hybrid/directional couplers for passive TX-RX duplexing, but at the cost of 3dB + 3dB insertion loss inherent to couplers.

As demonstrated in Figure 2, we present a 140-GHz frequency-modulated continuous-wave (FMCW) radar transceiver in 65-nm CMOS, featuring TX/RX antenna sharing that solves the TX/RX beam misalignment problem. A full-duplexing technique based on circular polarization and geometrical symmetry is applied to mitigate that 6dB inherent insertion loss, while maintaining high TX-to-RX isolation. In addition, a self-adaptive self-interference cancellation is implemented to suppress extra leakage due to antenna mismatch from a desired frontside radiation scheme. The TX/RX antenna sharing enables the pairing with a large 3D-printed planar lens and boosts the measured EIRP to 25.2dBm. The measured total radiated power and minimum single-sideband noise figure including antenna and duplexer losses are 6.2dBm and 20.2dB, respectively. The measured total TX-RX isolation is 33.3dB under 14-GHz wide FMCW chirps. Among all reported sub-THz transceivers with TX/RX antenna sharing, our work demonstrates the highest total radiated power and is the only work that has >30dB of TX-RX isolation while mitigating the inherent 6dB coupler loss.




▲ Figure 1: Architecture of cryptographic core and chip micrograph.




▲ Figure 2: Demo of the 140-GHz FMCW radar transceiver in 65-nm CMOS. The TX/RX antenna sharing enables the pairing with a large 3D-printed planar lens and accurate TX/RX beam alignment.

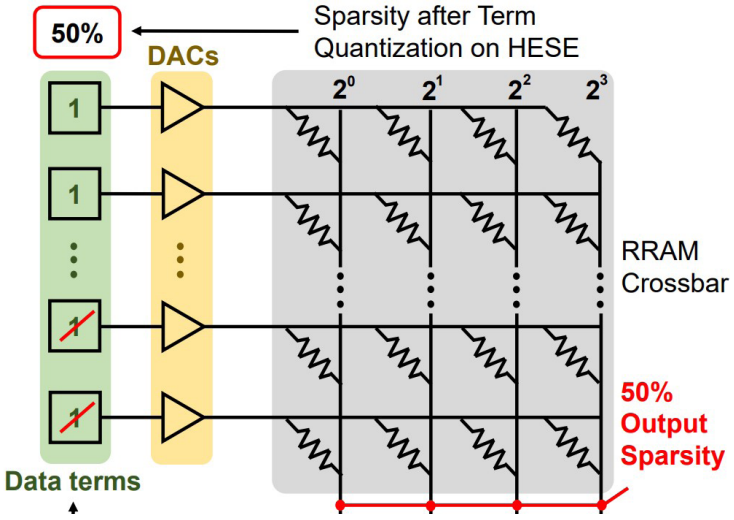
## FURTHER READING


- X. Chen, M. I. W. Khan, X. Yi, X. Li, W. Chen, J. Zhu, Y. Yang, K. E. Kolodziej, N. M. Monroe, and R. Han, "A 140GHz Transceiver with Integrated Antenna, Inherent-Low-Loss Duplexing and Adaptive Self-Interference Cancellation for FMCW Monostatic Radar," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, pp. 80-82, 2022, doi: 10.1109/ISSCC42614.2022.9731637.
- N. M. Monroe, G. C. Doqiamis, R. Stingel, P. Myers, X. Chen, and R. Han, "Electronic THz Pencil Beam Forming and 2D Steering for High Angular-Resolution Operation: A 98×98-Unit 265GHz CMOS Reflectarray with In-Unit Digital Beam Shaping and Squint Correction," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, pp. 1-3, 2022, doi: 10.1109/ISSCC42614.2022.9731671.
- X. Yi, C. Wang, X. Chen, J. Wang, J. Grajal, and R. Han, "A 220-to-320-GHz FMCW Radar in 65-nm CMOS Using a Frequency-Comb Architecture," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 2, pp. 327-339, Feb. 2021, doi: 10.1109/JSSC.2020.3020291.

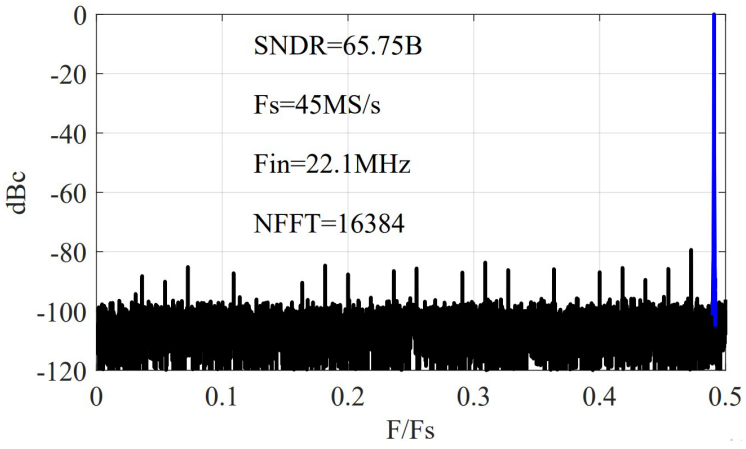
# A Bit-level Sparsity-aware SAR ADC with Direct Hybrid Encoding for Signed Expressions Leveraging Algorithm-circuit Co-design

R.-C. Chen, H. -T. Kung, A. P. Chandrakasan, H.-S. Lee
Sponsorship: CICS, DARPA

Machine learning is promising for many applications including image recognition and natural language processing. Machine learning accelerators are needed for these computation-intensive tasks. Analog neural networks are promising for breaking the memory wall for conventional machine learning accelerators. In this work, we propose the first sparsity-aware successive approximation register analog-to-digital converter (SAR ADC) with direct hybrid encoding for signed expressions (HESE) leveraging encoding algorithm-circuit co-design. ADCs are typically a bottleneck of analog neural networks. For a pre-trained convolutional neural network (CNN) inference, ANN with HESE SAR minimizes the non-zero terms and enables a reduction in energy along with the term quantization (TQ). The proposed SAR ADC directly produces the HESE signed-digit representation (SDR) using two thresholds per cycle as a 2-bit look-ahead. A proof-of-concept direct HESE SAR ADC is being fabricated by 65-nm technology. Measurements show that it provides the novel sparsity encoding with a Walden figure-of-merit of 15.2fJ/conv-step at a 45-MHz sampling rate. The core area is 0.072 mm^2. This opens the direction of direct sparsity encoding ADCs.




◀ Figure 1: An example crossbar of RRAM with 1-bit RRAM cells and 1-bit input values for computing dot products.




▶ Figure 2: Measured fast Fourier transform spectrum of the proposed direct HESE SAR.

## FURTHER READING

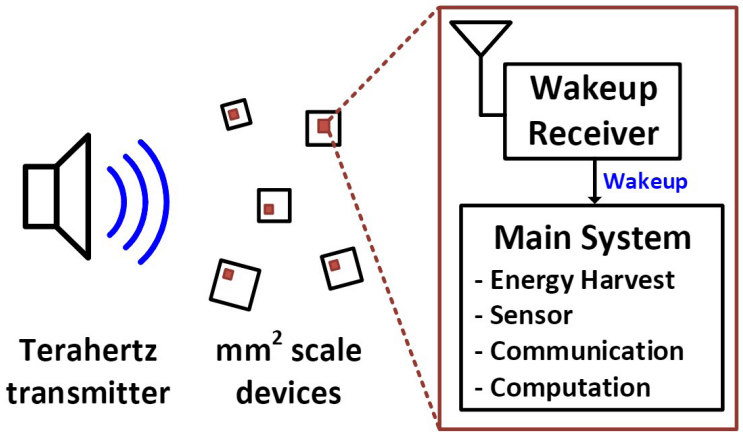

- Y. Peng, W. Huaqiang, G. Bin, T. Jianshi, Z. Qingtian, Z. Wenqiang, Y. J. Joshua, and Q. He, "Fully Hardware-implemented Memristor Convolutional Neural Network," Nature, vol. 577, no. 7792, pp. 641–646, 2020.
- H. T. Kung, B. McDanel, and S. Q. Zhang,"Term Revealing: Furthering Quantization at Run Time on Quantized Dnns," arXiv preprint arXiv:2007.06389, 2020.

# A Low-power THz Wakeup Receiver for an Ultra-miniaturized Platform

E. Lee, M. I. Ibrahim, U. Banerjee, R. T. Yazicigil, A. P. Chandrakasan, R. Han
Sponsorship: NSF, Korea Foundation for Advanced Studies

With the increasing demand for wirelessly connected devices, extending the lifetime of the communication nodes has become essential. As wireless communication is often one of the most power-hungry parts of an overall system, it is necessary to use devices with low-power wireless communication capabilities. A wakeup receiver (WuRX) is a circuit block that listens to the predefined token and turns on the node. The WuRX keeps the node in standby mode until a valid request, which helps to reduce unnecessary power consumption and, thus, lengthen the battery lifetime. Among various metrics of WuRXs, sensitivity and power consumption are two major axes that have led the progress of WuRXs in past decades. Several sub gigahertz/gigahertz works have achieved sensitivity-power tradeoff by co-designing off-chip components such as a high-Q antenna. While these have improved sensitivity and power performance, they are not suitable for ultra-miniaturized platforms due to the external components.

Pushing the carrier frequency to terahertz (THz) is key to reducing the form factor near the mm2- scale. Thanks to the small antenna aperture size requirement of the THz electromagnetic waves, antennas can be fabricated on a chip and integrated with the receiver's front-end without any external off-chip components. In this work, we aim at developing a sub-microwatt THz WuRX, operating at 261 GHz. We use an envelope detector first receiver architecture to avoid large power consumption of THz demodulation. To increase the sensitivity of the WuRX, we investigate a method to improve the noise equivalent power of the THz detector. The THz detector output is amplified, filtered, and recovered to the original data. The duty-cycling technique is also applied to reduce power consumption. In addition, we propose a secure wakeup protocol to prevent the battery-drainage attack, which is especially critical to battery size-limited miniaturized platforms. While this project is still in progress, this system will facilitate the use of THz for the ultra-miniaturized platform.




▲ Figure 1: Conceptual application scenario for the THz wakeup receiver.

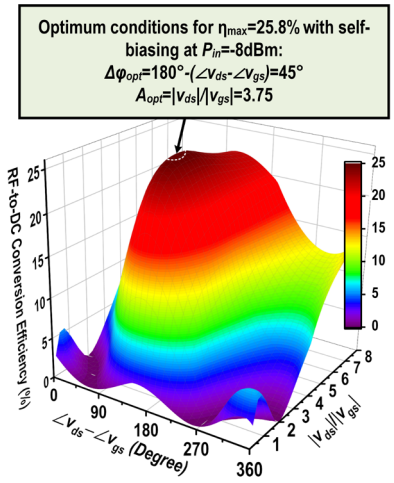
## FURTHER READING


- K. R. Sadagopan et al., "A 365nW −61.5 dBm Sensitivity, 1.875 cm2 2.4 GHz Wake-up Receiver with Rectifier-antenna Co-design for Passive Gain," *2017 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, pp. 180-183, 2017.
- H. Jiang et al., "A 22.3-nW, 4.55 cm2 Temperature-Robust Wake-Up Receiver Achieving a Sensitivity of −69.5 dBm at 9 GHz," *IEEE J. Solid-State Circuits*, vol. 55, no. 6, pp. 1530-1541, Jun. 2020.

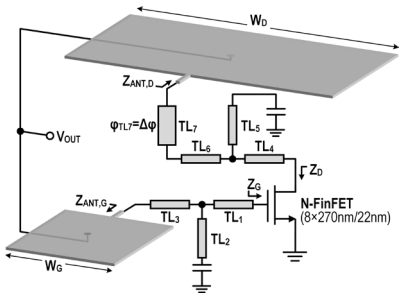
# A Dual-antenna, 263-GHz Energy Harvester in CMOS with 13.6% RF-to-DC Conversion Efficiency at -8dBm Input Power

M. I. W. Khan, E. Lee, N. M. Monroe, A. P. Chandrakasan, R. Han
Sponsorship: NSF (Grant No. SpecEES ECCS-1824360)

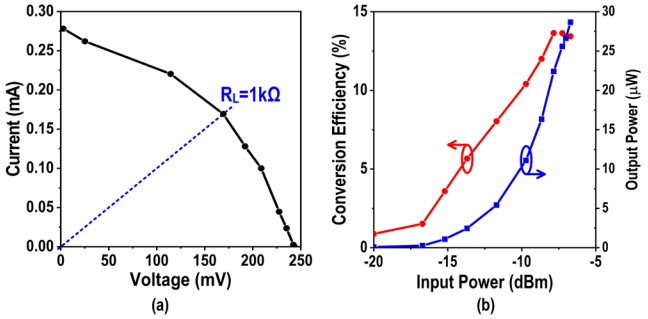
Pushing the wave frequency of far-field wireless power transfer (WPT) to the terahertz regime is essential for ultra-miniaturized, battery-less platforms, which currently can only be powered through light or ultra-sound. As an example, the mm²-size THz identification tag (THz-ID) in [1] relies on integrated photo-diodes, and THz WPT will allow embedding the tags into optically-opaque packages of small-size goods (e.g., semiconductor chips). In this work, a 263-GHz energy harvester using Intel's 22nm FinFET process is reported, increasing the highest frequency of CMOS harvester by ~3x. The antenna-integrated harvester is ultra-compact (~0.5mm²) and does not rely on any external component. In Fig.1, a self-biased N-FinFET is simulated with various ($v_{gs}$, $v_{ds}$) combinations while keeping input power equal to -8dBm. An $\eta$max of 25.8% is obtained, when phase difference $v_{ds}$-$v_{gs}$ is $\Delta\phi_{opt}$=45° and the amplitude ratio $|v_{ds}|/|v_{gs}|$ is $A_{opt}$=3.75. The schematic meeting these conditions is shown in Fig.2, where the additional phase tuning is provided by $TL_7$. Lastly, connecting the central AC ground nodes of the patch antennas together enables self-biasing of the transistor without interfering with antenna operations. The same connection is also used to extract DC output power. The measured load line performance of the harvester at 5cm distance, shown in Fig.3a, results in an optimum load of ~1kΩ. Fig. 3b shows that -8dBm input power the measured $\eta$max is 13.6% and 22μW of DC power is harvested.




▲Figure 1 Simulated rectification performance of a N-FinFET at various $v_{ds}$ and $v_{gs}$ ratio and phase difference.




▲Figure 2 Schematic of the THz energy harvester.




▲Figure 3 Measured (a) load line, and (b) conversion efficiency and output power with 1kΩ load.

## FURTHER READING


- M. I. W. Khan, M. I. Ibrahim, C. S. Juvekar, W. Jung, R. T. Yazicigil, A. P. Chandrakasan, and R. Han, "CMOS THz-ID: A 1.6-mm² Package-Less Identification Tag Using Asymmetric Cryptography and 260-GHz Far-Field Backscatter Communication," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 2, pp. 340–354, 2021.
- M. I. W. Khan, E. Lee, N. M. Monroe, A. Chandrakasan, and R. Han, "A Dual-Antenna, 263-GHz Energy Harvester in CMOS for Ultra-Miniaturized Platforms with 13.6% RF-to-DC Conversion Efficiency at -8dBm Input Power," to be presented at *2022 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, Denver, CO, 2022.

# Stability Improvement of CMOS Molecular Clocks Using an Auxiliary Loop Based on High-order Detection and Digital Integration

M. Kim, H.-S. Lee, R. Han
Sponsorship: JPL, NSF

An ultra-stable frequency reference is a key element for a wide variety of applications, ranging from sensing to navigation. Recently, chip-scaled molecular clocks (CSMC) have achieved high frequency stability with low power and compact size by using a rotational-mode transition of carbonyl sulfide (OCS) centered around 231.061 GHz as a frequency reference (f0). In the molecular clock, the probing signal generated from the transmitter is frequency-modulated at fm around the center frequency (fc). Since fc is locked to f0 in a feedback loop, the output frequency inherits the excellent stability of the OCS transition frequency. Due to its fully electronic implementation, CSMC provided a solution to significantly reduce the cost of high-stability miniaturized clocks. However, the frequency stability is still limited by a finite loop gain of the frequency locked loop and detection non-idealities coming from baseline variations that are susceptible to environmental disturbance even though an invariant physical constant is used as the frequency reference.

In this work, we propose a new dual loop CSMC architecture based on both fundamental and high-order transition probing as well as digital integration. While the fundamental harmonic detection forms the main loop, the higher-order probing is used in an auxiliary loop. The main loop enables the fast correction of the frequency, and the auxiliary loop responds against long-term frequency variation. As a result, the dual-loop architecture combines the advantages of both fundamental locking and high-order locking: high signal-to-noise ratio (SNR) and robustness against the environmental variations. The proposed CSMC was implemented in 65-nm complementary metal-oxide-semiconductor (CMOS) and achieved 20-ppt Allan Deviation at 104 s averaging time with 71-mW power consumption. This demonstrates the feasibility of miniaturization, as well as the low power and low cost of the clock.

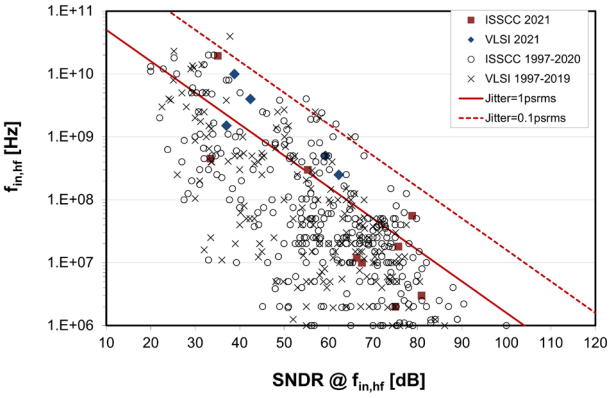
# A Sampling Jitter Tolerant Continuous-time Pipelined ADC in 16-nm FinFET

R. Mittal, H. Shibata, S. Patil, G. Manganaro, A. P. Chandrakasan, H.-S. Lee
Sponsorship: Analog Devices, Inc.

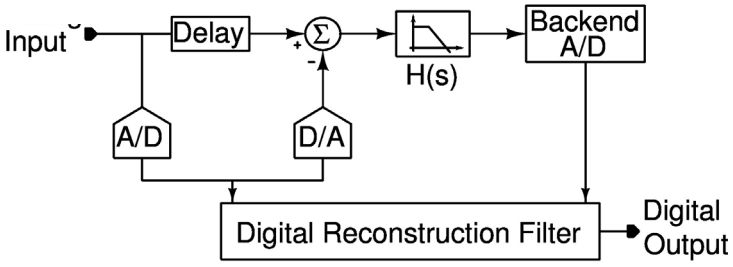
Almost all real-world signals are analog. Yet most of the data is stored and processed digitally due to advances in the integrated circuit technology. Therefore, analog-to-digital converters (ADCs) are an essential part of any electronic system. The advances in modern communication systems including 5G mobile networks and baseband processors require the ADCs to have a large dynamic range and bandwidth. 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 (Figure 1). 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. But since delta-sigma modulators rely on relatively high oversampling, they are unsuitable for high frequency applications. Therefore, ADCs with low oversampling ratio are desirable for high-speed data conversion.

In conventional Nyquist-rate ADCs, the input is sampled upfront. Any jitter in the sampling clock directly affects the sampled input and degrades the signal-to-noise ratio (SNR). It is well known that for a given root-mean-square (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. In a silicon-on-chip environment, it is difficult to reduce the RMS jitter below 100 fs. This limits the maximum SNR to just 44 dB 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.

We propose a continuous-time pipelined ADC having reduced sensitivity to sampling jitter (Figure 2). The analog input is processed in continuous time in the first stage. The residue is sampled by the backend ADC after amplification and low-pass filtering. This results in a much smaller derivative for the residue signal compared to the analog input. Since the error voltage due to clock jitter is proportional to the derivative of the sampled signal, the effect of sampling jitter is greatly reduced. 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: Performance survey for published ADCs (ISSCC 1997-2021 and VLSI 1997-2021).




▶ Figure 2: A schematic of the proposed continuous-time pipelined ADC tolerant to sampling jitter.

## FURTHER READING


- B. Murmann, "ADC Performance Survey 1997-2021," [Online]. Available: http://web.stanford.edu/~murmann/adcsurvey.html.
- H. Shibata, Hajime, V. Kozlov, Z. Ji, A. Ganesan, H. Zhu, D. Paterson, J. Zhao, S. Patil, and S. Pavan. "A 9-GS/s 1.125-GHz BW Oversampling Continuous-time Pipeline ADC Achieving– 164-dBFS/Hz NSD." *IEEE Journal of Solid-State Circuits* 52, no. 12 (2017): 3219-3234.
- J. D. Boles, E. Ng, J. H. Lang, and D. J. Perreault, "High-efficiency operating Modes for Isolated Piezoelectric-transformer-based DC-DC Converters," *Proc. IEEE Workshop on Control and Modeling for Power Electronics (COMPEL)*, Aalborg, Denmark, Nov. 2020.

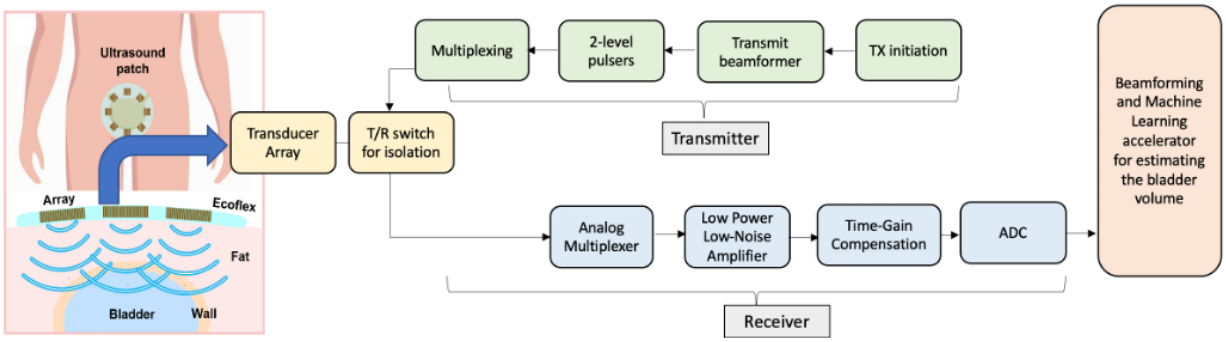
# Energy-efficient System for Bladder Volume Monitoring with Conformable Ultrasound Patches

V. Mittal, Z. Song, C. Marcus, L. Zhang, S. J. Schoen, V. Kumar, Y. Eldar, C. Dagdeviren, A. E. Samir, H.-S. Lee, A. P. Chandrakasan
Sponsorship: Texas Instruments

Continuous monitoring of bladder volume aids the management of many common conditions such as post-operative urinary retention and benign prostatic hyperplasia. Despite the success of ultrasound technology, there is a lack of wearable ultrasound probes capable of imaging curved body parts with high spatio-temporal resolution and making diagnostic decisions. Current systems are not sufficiently energy-efficient to permit continuous wearable device deployment for more than 1-2 days, as their power budget is several mW. We aim to develop a conformable, energy-efficient, battery-operated, wearable ultrasound patch capable of real-time organ monitoring. The wearable patch will be fully integrated with the transceiver electronics for energy-efficient processing of the ultrasonic signals and an efficient inference engine for bladder volume estimation. This system will incorporate several key innovations, including (1) deep neural network- (DNN) based segmentation algorithms employed to generate accurate bladder volume estimates; (2) low voltage ultrasound transceivers to enable low power, portable integrated system; and (3) signal processing algorithms capable of working with low signal-to-noise ratio (SNR) environments.

We aim to integrate the transducers with the analog front-end and DNN accelerator while ensuring that heat dissipation is within FDA specified limits. The power-efficient patch will operate at low voltage, thus posing the challenge of working with a low SNR signal. The transmitter consists of energy-efficient pulsers, appropriately beam-formed and multiplexed for different sub-apertures of the transducer array. On the receiver end, low-voltage, energy-efficient techniques are used to optimize the active power of the analog front end.

An on-chip DNN will extract the segmented mask from the beamformed image. The network is trained on bladder ultrasound images from MGH. The network is mostly binarized with the remaining operations quantized to minimize the memory requirement and eliminate the need for on-chip floating-point operation support. A DNN accelerator is designed for optimal binary DNN performance but also supports low bit-width computation. Lastly, the bladder volume is extracted from the segmented images and given as the system output using the double area method.




▲ Figure 1: System-level diagram of an ultrasound system.

## FURTHER READING

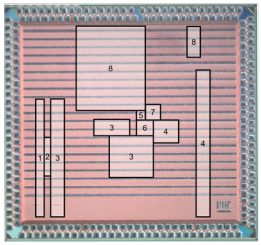

- N. K. Kristiansen, J. C. Djurhuus, H. Nygaard, "Design and Evaluation of an Ultrasound-based Bladder Volume Monitor," *Medical and Biological Engineering and Computing*, vol. 42, pp. 762-769, 2004.
- C. M. W. Daft, "Conformable Transducers for Large-volume, Operator-independent Imaging," *2010 IEEE International Ultrasonics Symposium*, pp. 798-808. 2010.
- M. Dicuio, G. Pomara, F.M. Fabris, V., Ales, C. Dahlstrand, and G. Morelli, "Measurements of Urinary Bladder Volume: Comparison of Five Ultrasound Calculation Methods in Volunteers," *Archivio Italiano Di Urologia e Andrologia*, vol. 77, no. 1, pp. 60–62, 2005.

# Hardware Design for Efficient Video Understanding on the Edge

M. Wang, Y. Lin, Z. Zhang, J. Lin, S. Han, A. P. Chandrakasan
Sponsorship: Qualcomm Incorporated

With the rise of various applications including autonomous driving, object tracking for unmanned aerial vehicles, etc., there is an increasing need for accurate and energy-efficient video understanding on the edge. Although there are many deep learning chips designed for images, little work has been done for videos. Video understanding on the edge has three major challenges. First, video understanding requires temporal modeling. For example, it identifies the difference between opening and closing a box, which is distinguishable only with temporal information considered. Second, many applications are delay-critical, such as self-driving cars. Third, high energy efficiency is important for edge devices with a tight power budget. Due to temporal continuity, consecutive frames might share a lot of common information, providing the potential to improve processing efficiency. However, an image-based processing system cannot utilize that since each frame is processed individually.

In this project, we co-design algorithms and hardware for energy-efficient video processing on delay-critical applications. We design architecture to natively support temporal shift module on the backbone of 2D convolutional neural networks for temporal modeling. Moreover, we propose a Real-Time DiffFrame method to utilize temporal redundancy and reduce on-chip energy and dynamic random-access memory (DRAM) traffic for delay critical applications. Compared to an ordinary convolution baseline, our method achieves around 2x reduction in both DRAM and static RAM (SRAM) accesses and 2x improvement in throughput with temporal modeling capability and no accuracy loss. The system has been fabricated in TSMC 28-nm complementary metal-oxide-semiconductor (CMOS) process. Figure 1 shows the chip photograph and specifications. We are evaluating our proposed system and measuring the performance of the chip.




| Chip Specifications | |
|---|---|
| Technology | TSMC 28nm HPC+ 1P10M |
| Die Size | 3000um × 2800um = 8.4mm$^2$ |
| Core Area | 2572um × 2372um = 6.1mm$^2$ |
| Package | Chip on board |
| Logic Gates | 576.5K (NAND2 equiv.) |
| On-Chip SRAM | 809.984K bits |
| Number of Multipliers | 64 |
| Arithmetic Precision | 8-bit WGT & ACT,<br>22-bit PSUM |
| Operating voltage | Core: 0.55V~ 0.9V; I/O: 1.8V |
| Core clock freq. | 29.4MHz ~ 90.9MHz |

1. Weight buffer (16kB)
2. 8x8 MAC array
3. IA buffer (32kB)
4. OA & RefFrame unit (44kB)
5. DiffFrame generator
6. DiffFrame pruning
7. Kernel map buffer
8. Kernel map generator & Point coordinate buffer

▲ Figure 1: Chip micrograph and specifications.

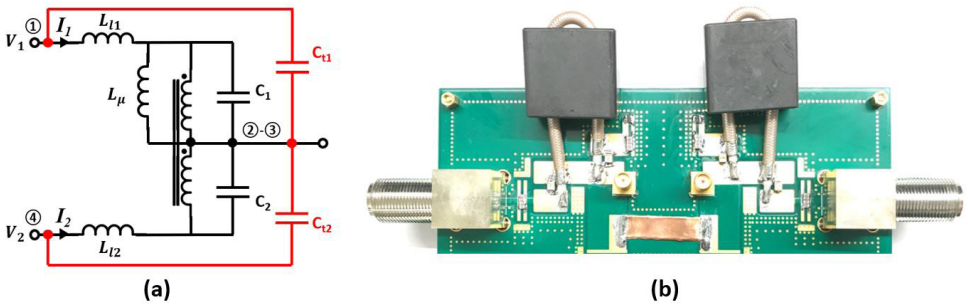
## FURTHER READING


- J. D. Boles, J. J. Piel, and D. J. Perreault, "Enumeration and Analysis of DC-DC Converter Implementations Based on Piezoelectric Resonators," *IEEE Transactions on Power Electronics*, vol. 36, no. 1, pp. 129-145, 2021.
- E. Ng, J. D. Boles, J. H. Lang, and D. J. Perreault, "Non-isolated DC-DC Converter Implementations Based on Piezoelectric Transformers," Proc. *IEEE Energy Conversion Congress and Exposition (ECCE)*, Vancouver, Canada, Oct. 2021.
- J. D. Boles, E. Ng, J. H. Lang, and D. J. Perreault, "High-efficiency Operating Modes for Isolated Piezoelectric-transformer-based DC-DC Converters," *Proc. IEEE Workshop on Control and Modeling for Power Electronics (COMPEL)*, Aalborg, Denmark, Nov. 2020.

# Modeling and Design of High-power RF Power Combiners Based on Transmission-lines

H. Zhang, G. Cassidy, A. Jurkov, K. Luu, A. Radomski, D. J. Perreault
Sponsorship: MKS Instruments, Inc.

Industrial plasma generation for semiconductor processing applications are usually characterized by high power levels (e.g., kWs), wide power ranges (e.g., 30dB dynamic range), narrow-frequency-band operations (e.g., 13.56MHz ± <5%), and the need to combine power from multiple sources. Power combiners based on transmission lines are attractive due to their small form factor and high efficiency. However, most existing literature focuses on frequency response, with little consideration regarding losses or co-design with magnetic components. Here we introduce a lumped-element circuit model better suited for this application space and further propose a tuning technique that, by adding two capacitors, minimizes impedance distortion while preserving high efficiency. A 13.56-MHz, 1-kW prototype is designed and built, validating the model and tuning technique with both small-signal measurements and high-power tests. The study would help in realizing radio frequency power generation systems that maintain high efficiency over a very wide power range.




▲ Figure 1: (a) Proposed lumped-element circuit model with tuning technique (achieved by adding two capacitors, Ct1 and Ct2), and (b) high-power testbench with 2-combiners connected in a back-to-back fashion.

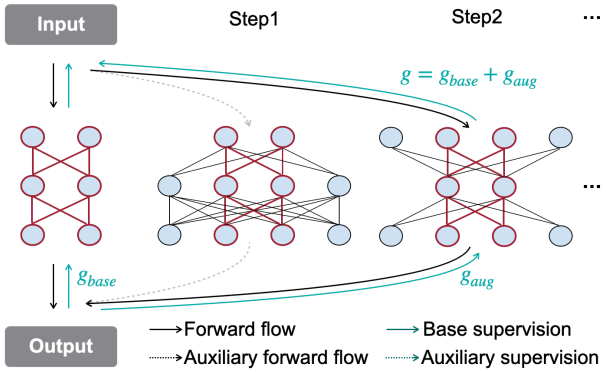
## FURTHER READING


- H. Zhang et. al., "Modeling and Design of High-Power Non-Isolating RF Power Combiners based on Transmission Lines," *2022 IEEE Applied Power Electronics Conference and Exposition (APEC)*, to be published.
- H. Zhang et. al., "Multi-Inverter Discrete Backoff: A High-Efficiency, Wide-Range RF Power Generation Architecture," *2020 IEEE 21st Workshop on Control and Modeling for Power Electronics (COMPEL)*, pp. 1-8, 2022, doi: 10.1109/COMPEL49091.2020.9265702.
- A. Al Bastami et. al., "Comparison of Radio-Frequency Power Architectures for Plasma Generation," *2020 IEEE 21st Workshop on Control and Modeling for Power Electronics (COMPEL)*, pp. 1-8, 2020, doi: 10.1109/COMPEL49091.2020.9265700.

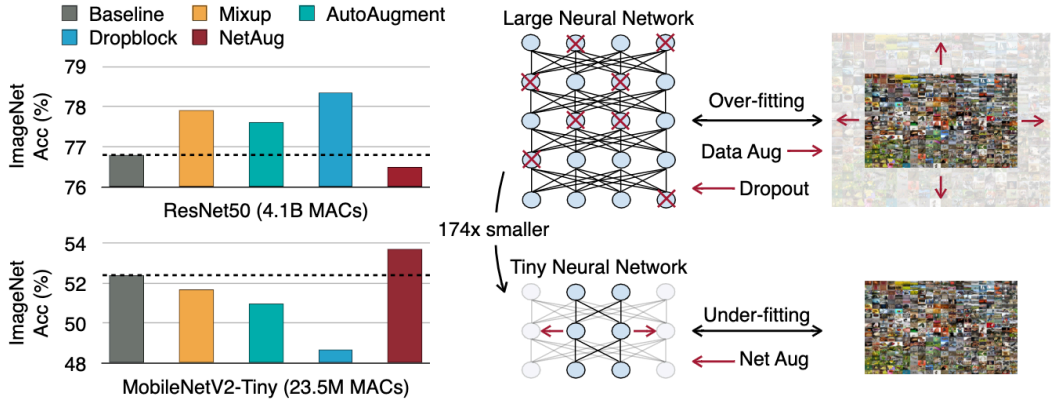
# Network Augmentation for Tiny Deep Learning

H. Cai, C. Gan, J. Lin, S. Han
Sponsorship: MIT-IBM Watson AI Lab, NSF, Hyundai, Ford, Intel, Amazon

We introduce Network Augmentation (NetAug), a new training method for improving the performance of tiny neural networks. Existing regularization techniques (e.g., data augmentation, dropout) have shown much success on large neural networks by adding noise to overcome over-fitting. However, we found that these techniques hurt the performance of tiny neural networks. We argue that training tiny models differ from large models: rather than augmenting the data, we should augment the model, since tiny models tend to suffer from under-fitting rather than over-fitting due to limited capacity. To alleviate this issue, NetAug augments the network (reverse dropout) instead of inserting noise into the dataset or the network. NetAug puts the tiny model into larger models and encourages it to work as a sub-model of larger models to get extra supervision, in addition to functioning as an independent model. At test time, only the tiny model is used for inference, incurring zero inference overhead. We demonstrate the effectiveness of NetAug on image classification and object detection. NetAug consistently improves the performance of tiny models, achieving up to 2.2% accuracy improvement on ImageNet. On object detection, achieving the same level of performance, NetAug requires 41% fewer MACs on Pascal VOC and 38% fewer MACs on COCO than the baseline.




◀ Figure 1: NetAug encourages the target tiny model to work as a sub-model of larger models to get extra supervision.




▲ Figure 2: NetAug improves the accuracy of the tiny model while regularization methods hurt its accuracy.

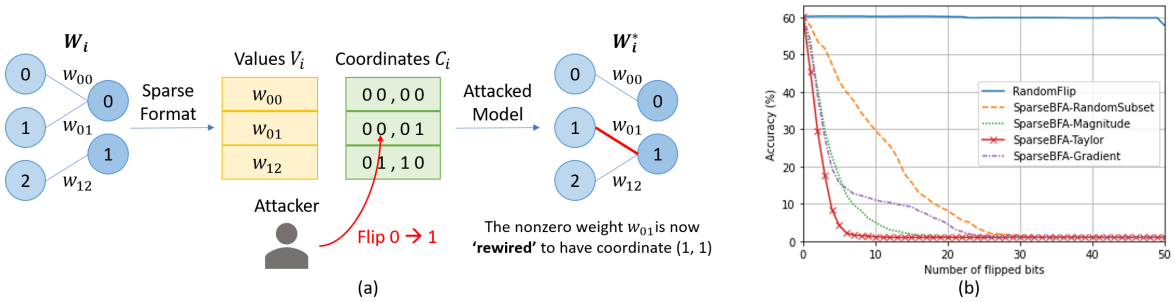
## FURTHER READING


- H. Cai, C. Gan, J., Lin, and S. Han, "Network Augmentation for Tiny Deep Learning," *ICLR*, 2022.
- H. Cai, et al. "TinyTL: Reduce Activations, Not Trainable Parameters for Efficient On-Device Learning," *Advances in Neural Information Processing Systems*, vol. 33, p. ?, 2020.
- H. Cai, et al. "Once-for-all: Train One Network and Specialize it for Efficient Deployment," *ICLR*, 2020.

# SparseBFA: Attacking Sparse Deep Neural Networks with the Worst-case Bit Flips on Coordinates

K. Lee, A. P. Chandrakasan
Sponsorship: Facebook, Korea Foundation for Advanced Studies

Deep neural networks (DNNs) are shown to be vulnerable to a few carefully chosen bit flips in their parameters, and bit flip attacks (BFAs) exploit such vulnerability to degrade the performance of DNNs. In this work, we show that DNNs with high sparsity that typically result from weight pruning have a unique source of vulnerability to bit flips when their coordinates of nonzero weights are attacked. We propose SparseBFA, an algorithm that searches for a small number of bits among the coordinates of nonzero weights when the parameters of DNNs are stored using sparse matrix formats. Using SparseBFA, we find that the performance of DNNs drops to the random-guess level by flipping less than 0.00005% (1 in 2 million) of the total bits.




▲ Figure 1: (a) When an attacker flips a bit in the coordinates representing the location of nonzero weights, the connection between neurons is rewired. (b) Accuracy of the ResNet50 model as bits in the coordinate list are flipped using SparseBFA.

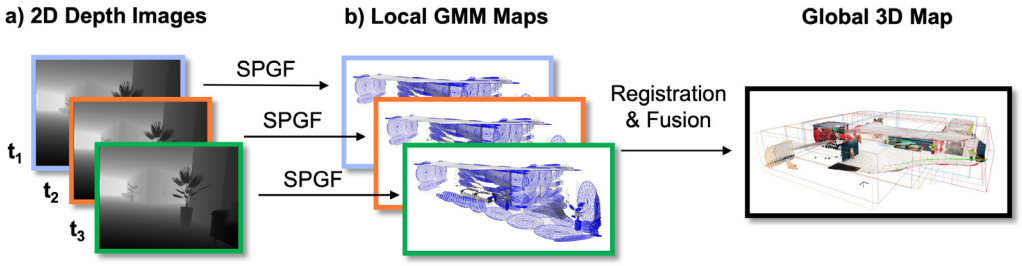
# Memory-efficient Gaussian Fitting for Depth Images in Real Time

P. Z. X. Li, S. Karaman, V. Sze
Sponsorship: NSF RTML 1937501, NSF CPS 1837212

Energy-constrained microrobots, such as insect-sized flapping wing robots and palm-sized drones, are expected to be deployed for search and rescue missions in dangerous and unknown environments. These robots have very limited battery capacity, which limits the energy available for computation. Since the energy cost of memory access can be significant, algorithms designed for these robots should reduce memory overhead so that most data and variables used during computation can be efficiently stored in and accessed from lower-level caches (KBs in storage) instead of a larger off-chip dynamic random-access memory (DRAM).

Constructing a compact representation for 3D environments is essential for enabling autonomy for tasks such as navigation, localization, and exploration. From a sequence of depth images, many existing algorithms convert each image into a compact Gaussian mixture model (GMM) where each Gaussian models a surface in the environment. Then, GMMs across all images are fused together into a coherent global 3D map (Figure 1). While existing algorithms focus on reducing the size of each GMM, they require significant memory overhead due to the storage of the entire depth image or its intermediate representation in memory for multi-pass processing.

In this work, we present the Single-Pass Gaussian Fitting (SPGF) algorithm that incrementally constructs a GMM one pixel at a time in a single pass through a depth image. Since only one pixel is stored in memory at any time, SPGF achieves orders-of-magnitude lower memory overhead than prior approaches. By processing each depth image row-by-row, SPGF can efficiently and accurately infer surface geometries, which leads to higher precision than prior multi-pass approaches while maintaining the same compactness of the GMM. Using a low-power ARM Cortex-A57 CPU, SPGF operates at 32 fps, requires 43 KB of memory overhead, and consumes only 0.11 J per image. Thus, SPGF enables real-time mapping of large 3D environments on energy-constrained robots.




▲ Figure 1: (a) A depth image from a depth camera, and (b) a GMM (blue) generated using the proposed SPGF algorithm with a root-mean-square error of 9 cm, a memory overhead of 43 KB, a throughput of 32 fps, and an energy consumption of 0.11 J per frame using the low-power ARM Cortex-A57 CPU.

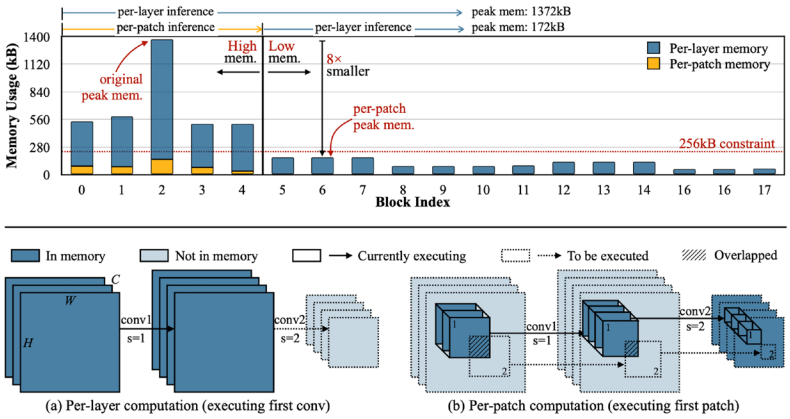
## FURTHER READING


- P. Z. X. Li, S. Karaman, V. Sze, "Memory-Efficient Gaussian Fitting for Depth Images in Real Time," *IEEE International Conference on Robotics and Automation (ICRA)*, May 2022.

# MCUNetV2: Memory-efficient Patch-based Inference for Tiny Deep Learning

J. Lin, W. Chen, H. Cai, C. Gan, S. Han
Sponsorship: MIT-IBM Watson AI Lab, Samsung, Woodside Energy, NSF CAREER Award #1943349

Tiny deep learning on microcontroller units (MCUs) is challenging due to the limited memory size. We find that the memory bottleneck is due to the imbalanced memory distribution in convolutional neural network (CNN) designs: the first several blocks have an order-of-magnitude larger memory usage than the rest of the network. To alleviate this issue, we propose a generic patch-by-patch inference scheduling, which operates only on a small spatial region of the feature map and significantly cuts down the peak memory. However, naive implementation brings overlapping patches and computation overhead. We further propose network redistribution to shift the receptive field and floating-point operations (FLOPs) to the later stage and reduce the computation overhead. Manually redistributing the receptive field is difficult. We automate the process with neural architecture search to jointly optimize the neural architecture and inference scheduling, leading to MCUNetV2. Patch-based inference effectively reduces the peak memory usage of existing networks by 4-8x. Co-designed with neural networks, MCUNetV2 sets a record ImageNet accuracy on MCU (71.8%), and achieves >90% accuracy on the visual wake words dataset under only 32kB static random access memory (SRAM). MCUNetV2 also unblocks object detection on tiny devices, achieving 16.9% higher mean Average Precision (mAP) on Pascal VOC compared to the state-of-the-art result. Our study largely addresses the memory bottleneck in tinyML and paves the way for various vision applications beyond image classification.




▲ Figure 1: MobileNetV2 has a very imbalanced memory usage distribution: the peak memory is determined by the first 5 blocks with high peak memory, while the later blocks all share a small memory usage. By using per-patch inference, we are able to significantly reduce the peak memory by 8x, fitting MCUs with a 256 kB memory budget.

## FURTHER READING

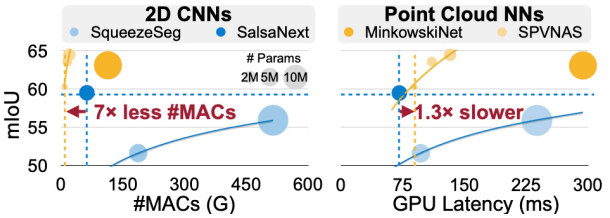

- J. Lin, W. M. Chen, Y. Lin, C. Gan, and S. Han, "MCUNet: Tiny Deep Learning on Iot Devices," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11711-11722, 2020.

# PointAcc: Efficient Point Cloud Deep Learning Accelerator

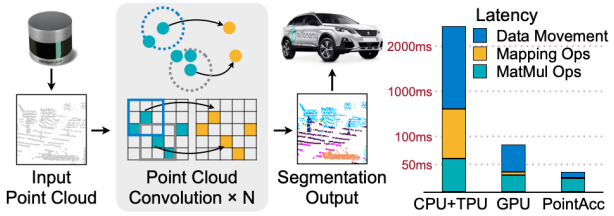
Y. Lin, Z. Zhang, H. Tang, H. Wang, S. Han
Sponsorship: NSF, Hyundai, Qualcomm, MIT-IBM Watson AI Lab

Deep learning on point clouds plays a vital role in a wide range of applications such as autonomous driving and augmented reality (AR) and virtual reality (VR). These applications interact with people in real time on edge devices and thus require low latency and low energy. Compared to projecting the point cloud to 2D space, directly processing 3D point cloud yields higher accuracy and lower number of multiply-accumulations (#MACs). However, the extremely sparse nature of point cloud poses challenges to hardware acceleration. For example, we need to explicitly determine the nonzero outputs and search for the nonzero neighbors (mapping operation), which is unsupported in existing accelerators. Furthermore, explicit gathering and scattering of sparse features are required, resulting in large data movement overhead.

In this work, we comprehensively analyze the performance bottleneck of modern point cloud networks on central processing, graphics processing, and tensor processing units (CPU/GPU/TPU). To address the challenges, we then present PointAcc, a novel point cloud deep learning accelerator. PointAcc maps diverse mapping operations onto one versatile ranking-based kernel, streams the sparse computation with configurable caching, and temporally fuses consecutive dense layers to reduce the memory footprint. Evaluated on 8 point cloud models across 4 applications, PointAcc achieves 3.7× speedup and 22× energy savings over RTX 2080Ti GPU. Co-designed with light-weight neural networks, PointAcc rivals the prior accelerator Mesorasi by 100× speedup with 9.1% higher accuracy running segmentation on the S3DIS dataset. PointAcc paves the way for efficient point cloud recognition.




▲Figure 1: Compared to 2D CNNs, point cloud networks have higher accuracy and lower #MACs, but higher GPU latency due to low utilization brought by sparsity and irregularity.




▲Figure 2: Point cloud deep learning is crucial for real-time AI applications. PointAcc accelerates point cloud computations by resolving sparsity and data movement bottlenecks.

## FURTHER READING


- Y. Lin, Z. Zhang, H. Tang, H. Wang, and S. Han, "PointAcc: Efficient Point Cloud Accelerator," *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 449-461, Oct. 2021.
- Y. Feng, B. Tian, T. Xu, P. Whatmough, and Y. Zhu, "Mesorasi: Architecture Support for Point Cloud Analytics via Delayed-aggregation," *MICRO-53: 53rd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 1037-1050, Oct. 2020.

# Algorithm-system Co-design for Efficient Calorimetry Clustering

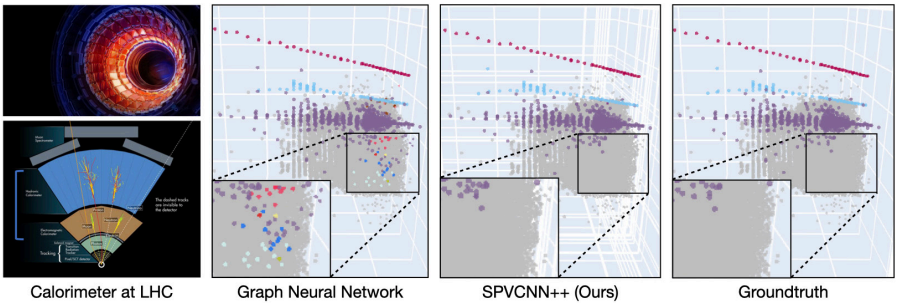
Z. Liu, X. Yang, S. Han
Collaborators: A. Schuy (UW), S-C. Hsu (UW), J. Krupa (MIT), P. Harris (MIT)
Sponsorship: NSF

The content management system (CMS) detector at the Large Hadron Collider (LHC) reconstructs high-energy proton-proton collisions to understand physics beyond the standard model. A key part of the CMS detector is the calorimeter, which reconstructs particle energies by clustering 3D energy deposits from particle showers. The LHC observes ~1 billion collisions per second and must decide within ~1 millisecond which collisions to keep; this imposes a strict throughput/latency requirement. Furthermore, the LHC data flow will increase tenfold by 2027. The corresponding increase in computing requirements using traditional algorithms is beyond our capabilities. Therefore, there is an urgent need to develop accurate algorithms capable of scaling under resource and latency constraints.

3D point cloud neural networks are very suitable for calorimetry clustering. However, they are ten times more computationally expensive than 2D convoluted neural networks (CNNs). Moreover, the sparse and irregular nature of the point cloud makes them less favored by general-purpose hardware (such as CPU, GPU, and TPU). We approach these challenges with algorithm-system co-design.

From the algorithm side, we have developed SPVCNN++, which brings together the best from point-based and voxel-based models. SPVCNN++ is composed of a fine-grained point-based branch that keeps the 3D data in high resolution without large memory footprints and a coarse-grained voxel-based branch that aggregates the neighboring features without many random memory accesses. Compared with the GNN-based approach, our SPVCNN++ achieves a 4% higher panoptic quality on the particle physics benchmark.

From the system side, we have developed TorchSparse, a specialized high-performance GPU computing library for 3D sparse computations. TorchSparse directly optimizes the two bottlenecks of sparse convolution: irregular computation and data movement. As a result, our TorchSparse achieves more than 1.5x measured end-to-end speedup over the state of the art.




▲ Figure 1: Results of our algorithm-system co-design solution for efficient calorimetry clustering. Compared with conventional GNN-based approach, our SPVCNN++ provides much more accurate clustering results.

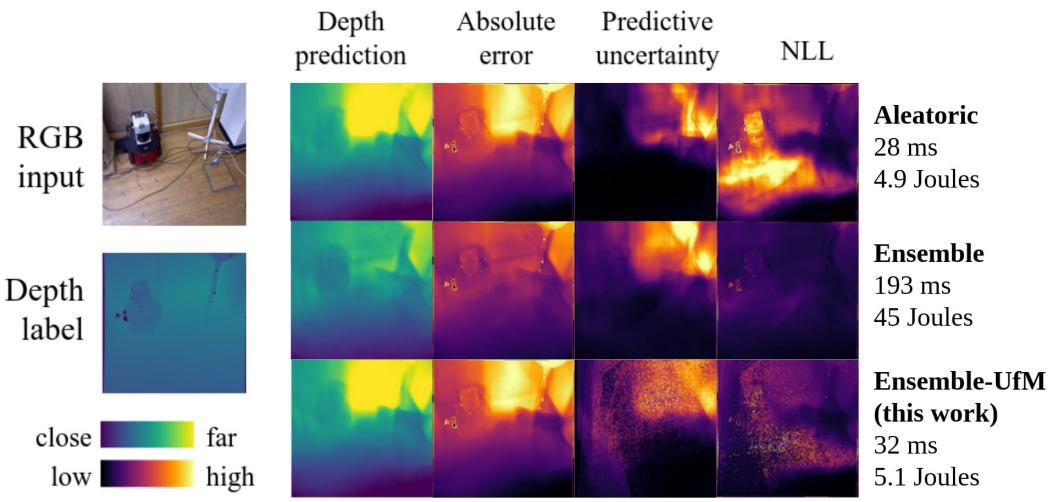
## FURTHER READING


- H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution," *European Conference on Computer Vision (ECCV)*, Aug. 2020.
- H. Tang, Z. Liu, X. Li, Y. Lin, and S. Han, "TorchSparse: Efficient Point Cloud Inference Engine," to be presented at *Conference on Machine Learning and Systems (MLSys)*, August 2022.

# Uncertainty from Motion for DNN Monocular Depth Estimation

S. Sudhakar, V. Sze, S. Karaman
Sponsorship: NSF Cyber-Physical Systems Program Grant no. 1837212, NSF Real-Time Machine Learning Program Grant no. 1937501, MIT-Accenture Fellowship

Deployment of deep neural networks (DNNs) for monocular depth estimation in safety-critical scenarios on resource-constrained platforms requires well-calibrated and efficient uncertainty estimates. However, uncertainty estimates from state-of-the-art ensembles are computationally expensive, requiring multiple inferences per input. We propose a new algorithm, called Uncertainty from Motion (UfM), that runs only one inference per input by exploiting the temporal redundancy in video inputs to merge incrementally the per-pixel depth prediction and per-pixel uncertainty over a sequence of frames. In a set of experiments using a DenseNet-based autoencoder on a single GPU, UfM offers near ensemble uncertainty quality while consuming on average 5.1 Joules with a latency of 32 ms per frame, which is 8.8x less energy and 6.4x faster than the ensemble. In Figure 1, we compare the results of a DNN that predicts only its data (aleatoric) uncertainty, an ensemble that predicts its overall uncertainty, and a DNN with UfM. We see that UfM retains the uncertainty quality of ensembles at a fraction of the energy and latency, enabling uncertainty estimation for resource-constrained, real-time scenarios.




▲ Figure 1: Uncertainty estimation comparison for an aleatoric network, ensemble, and UfM applied to ensembles on an out-of-distribution example from the TUM RGBD dataset. Lower negative log-likelihood (NLL) indicates better uncertainty quality.

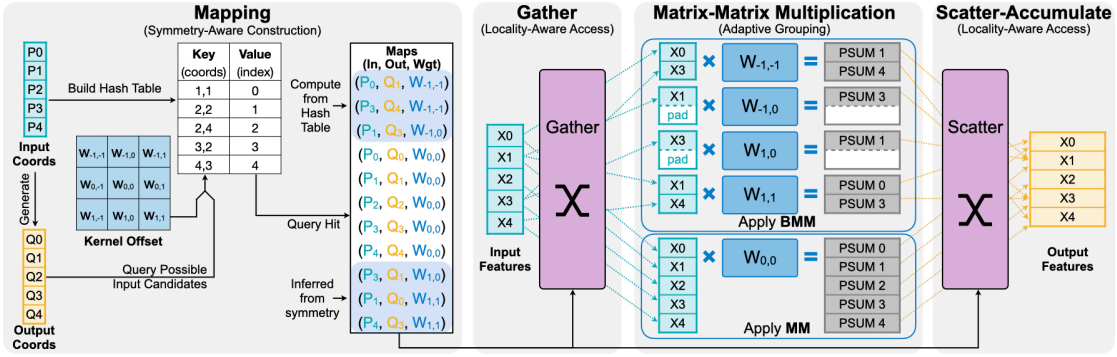
## FURTHER READING


- S. Sudhakar, S. Karaman, and V. Sze, "Uncertainty from Motion for DNN Monocular Depth Estimation," to be presented at *2022 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2022.

# TorchSparse: Efficient Point Cloud Inference Engine

H. Tang, Z. Liu, X. Li, Y. Lin, S. Han
Sponsorship: NSF CAREER Award, Ford, Hyundai, Qualcomm Innovation Fellowship

Deep learning on point clouds has received increased attention thanks to its wide applications in augmented and virtual reality and autonomous driving. These applications require low latency and high accuracy to provide real-time user experience and ensure user safety. Unlike conventional dense workloads, the sparse and irregular nature of point clouds poses severe challenges to running sparse convoluted neural networks efficiently on general-purpose hardware. Furthermore, existing sparse acceleration techniques for 2D images do not translate to 3D point clouds. In this paper, we introduce TorchSparse, a high-performance point cloud inference engine that accelerates sparse convolution computation on graphics processing units. TorchSparse directly optimizes the two bottlenecks of sparse convolution: irregular computation and data movement. It applies adaptive matrix multiplication grouping to trade computation for better regularity, achieving 1.4-1.5x speedup for matrix multiplication. It also optimizes the data movement by adopting vectorized, quantized, and fused locality-aware memory access, reducing the memory movement cost by 2.7x. Evaluated on seven representative models across three benchmark datasets, TorchSparse achieves 1.6x and 1.5x measured end-to-end speedup over the state-of-the-art MinkowskiEngine and SpConv, respectively.




▲ Figure 1: TorchSparse aims at accelerating sparse convolution, which consists of four stages: mapping, gathering, matrix multiplication. and scatter-accumulation. We follow two general principles: (1) memory footprint should be reduced, and (2) computation regularity should be increased to optimize these four components with quantized, vectorized, row-major scatter/gather (Principle 1); adaptively batched MM (Principle 2); and mapping kernel fusion (Principle 1).

## FURTHER READING


- H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution," *European Conference on Computer Vision (ECCV)*, Aug. 2020.
- H. Tang, Z. Liu, X. Li, Y. Lin, and S. Han, "TorchSparse: Efficient Point Cloud Inference Engine," *Conference on Machine Learning and Systems (MLSys)*, Aug. 2022.

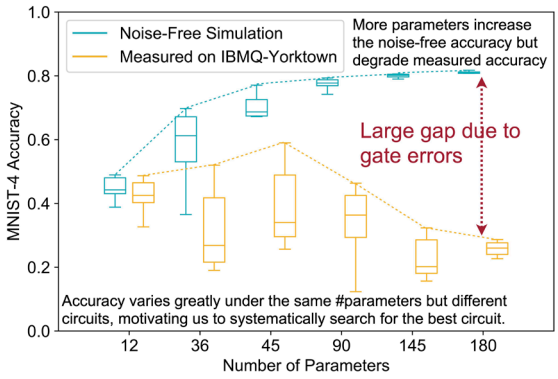
# QuantumNAS: Noise-Adaptive Search for Robust Quantum Circuits

H. Wang, Y. Ding, J. Gu, Z. Li, Y. Lin, D. Z. Pan, F. T. Chong, S. Han
Sponsorship: MIT-IBM Watson AI Lab, NSF CAREER Award, Qualcomm Innovation Fellowship

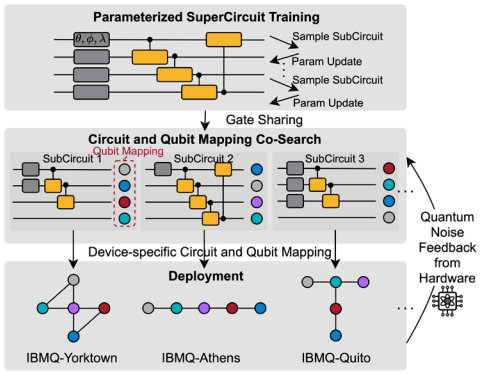
Quantum noise is the key challenge in noisy intermediate-scale quantum (NISQ) computers. Previous work on mitigating noise has primarily focused on gate-level or pulse-level noise-adaptive compilation. However, few research efforts have explored a *higher level of optimization* by making the quantum circuits themselves resilient to noise.

We propose QuantumNAS, a comprehensive framework for noise-adaptive co-search of the variational circuit and qubit mapping. Variational quantum circuits are a promising approach for performing quantum machine learning (QML) and simulation. However, finding the best variational circuit and its optimal parameters is challenging due to the large design space and parameter training cost. We propose to decouple the circuit search and parameter training by introducing a novel *SuperCircuit*. The SuperCircuit is constructed with multiple layers of pre-defined parameterized gates and trained by iteratively sampling and updating the parameter subsets (SubCircuits) of it. It provides an accurate estimation of SubCircuits performance trained from scratch. Then we perform an evolutionary co-search of SubCircuit and its qubit mapping. The SubCircuit performance is estimated with parameters inherited from SuperCircuit and simulated with real device noise models. Finally, we perform iterative gate pruning and finetuning to remove redundant gates.

Extensively evaluated with 12 QML and Variational Quantum Eigensolver (VQE) benchmarks on 14 quantum computers, QuantumNAS significantly outperforms baselines. For QML, QuantumNAS is the first to demonstrate over 95% 2-class, 85% 4-class, and 32% 10-class classification accuracy on real quantum machines. It also achieves the lowest eigenvalue for VQE tasks on $H_2$, $H_2O$, LiH, $CH_4$, and $BeH_2$ compared with UCCSD. We also open-source TorchQuantum (https://github.com/mit-han-lab/torchquantum) for fast training of parameterized quantum circuits to facilitate future research.




▲Figure 1: MNIST-4 on noise-free simulator / real QC. More parameters increase the noise-free accuracy but degrade measured accuracy due to larger gate errors. Accuracy gap is large.




▲Figure 2: Noise-adaptive circuit and qubit mapping co-search improves the robustness on real machines.

## FURTHER READING

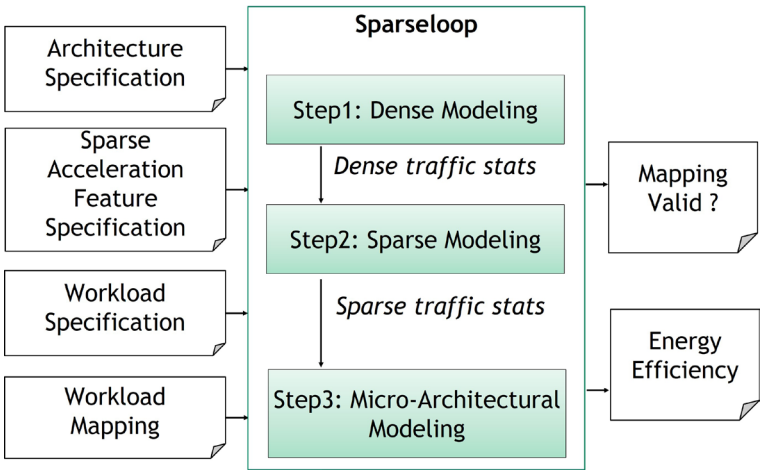

- H. Wang, Y. Ding, J. Gu, Z. Li, Y. Lin, D. Z. Pan, F. T., Chong, and S. Han, "QuantumNAS: Noise-Adaptive Search for Robust Quantum Circuits," *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2022, pp. 692-708, doi: 10.1109/HPCA53966.2022.00057..
- H. Wang, J. Gu, Y. Ding, Z. Li, F. T. Chong, D. Z. Pan, and S. Han, "QuantumNAT: Quantum Noise-Aware Training with Noise Injection, Quantization and Normalization," *2022 Design Automation Conference*, 2022.
- H. Wang, Z. Li, J. Gu, Y. Ding, D. Z. Pan, & S. Han, "QOC: Quantum On-Chip Training with Parameter Shift and Gradient Pruning," *2022 Design Automation Conference*, 2022.

# Sparseloop: An Analytical Approach to Sparse Tensor Accelerator Modeling

Y. N. Wu, P.-A. Tsai, A. Parashar, V. Sze, J. S. Emer
Sponsorship: DARPA (HR0011-18-3-0007), Ericsson

In recent years, a myriad of accelerators has been proposed to efficiently process sparse tensor algebra applications (e.g., neural networks), leading to a large and diverse design space. However, the lack of systematic description and modeling support for these sparse tensor accelerators prevents hardware designers from efficient design space exploration.

To solve the problem, we present Sparseloop, the first fast, accurate, and flexible analytical modeling framework for sparse tensor accelerators. Figure 1 shows Sparseloop's high-level framework. Based on a unified taxonomy to describe the diverse designs, Sparseloop comprehends a wide set of architecture specifications and calculates designs' performance based on stochastic tensor density models. Across a representative set of accelerators and workloads, Sparseloop achieves >600x faster modeling speed than cycle-level simulations, <1% error compared to a custom accelerator model with statistical data modeling, and <8% error compared to simulations with real data.




▲ Figure 1: MNIST-4 on noise-free simulator / real QC. More parameters increase the noise-free accuracy but degrade measured accuracy due to larger gate errors. Accuracy gap is large.

## FURTHER READING


- Y. N. Wu, P.-A. Tsai, A. Parashar, V. Sze, and J. S. Emer, "Sparseloop: An Analytical, Energy-Focused Design Space Exploration Methodology for Sparse Tensor Accelerators," presented at *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, Mar., 2021.
- Y. N. Wu, J. S. Emer, and V. Sze, "Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs," presented at *2019 International Conference on Computer Aided Design (ICCAD)*, Nov. 2019.
- A. Parashar et al., "Timeloop: A Systematic Approach to DNN Accelerator Evaluation," presented at *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, Mar., 2019.

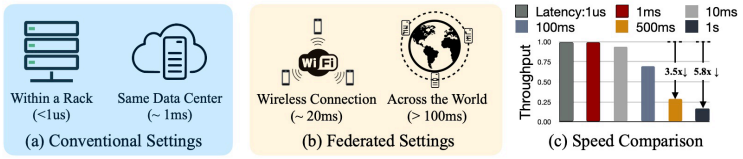
# Latency-tolerant On-device Learning

L. Zhu, S. Han
Sponsorship: MIT-IBM Watson AI Lab, Samsung, Woodside Energy, NSF, Amazon

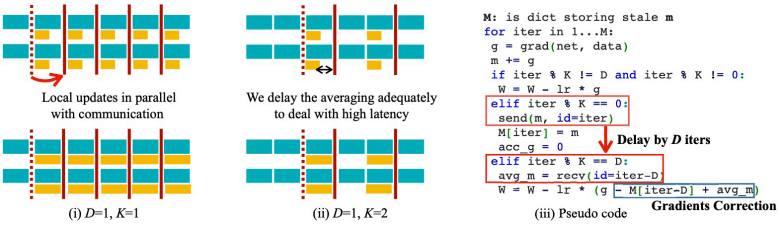
Much new and sensitive data are generated and collected by intelligent edge devices with rich sensors every day. On-device federated learning is an emerging direction that enables jointly training a model without sharing the data. Since the data is distributed across many edge devices through wireless / long-distance connections, federated learning suffers from inevitable high communication latency. However, the latency issues are undermined in the current literature and existing approaches such as FedAvg become less efficient when the latency increases.

To overcome the problem, we propose delayed gradient averaging (DGA) to address the latency bottleneck. The key idea is to delay the gradient averaging to a future iteration; thus the communication can be pipelined with computation (as shown in Figure 2). By accepting stale average gradients for model updates, DGA allows the communication to execute in parallel with the computation and become scalable even under extreme latency.

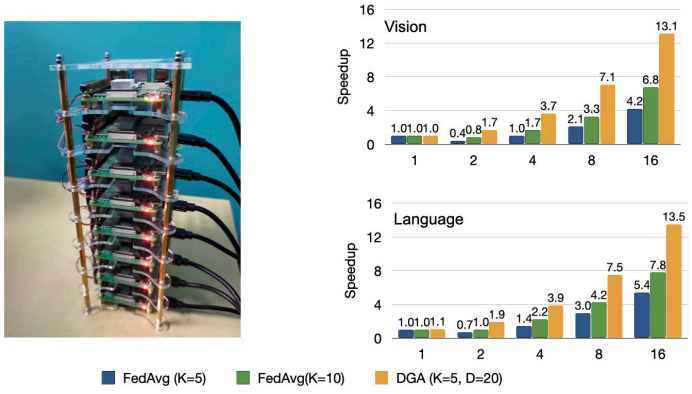
We theoretically prove that DGA attains a similar convergence rate as FedAvg and empirically show that our algorithm can tolerate high network latency without compromising accuracy. Specifically, we benchmark the training speed on various vision (CIFAR, ImageNet) and language tasks (Shakespeare), with both independent and identically distributed (IID) and non-IID partitions, and show that DGA can bring 2.55× to 4.07× speedup. Moreover, we built a 16-node Raspberry Pi cluster and show that DGA can consistently speed up real-world federated learning applications




▲ Figure 1: Training settings of conventional distributed training and federated learning differ greatly. High latency cost greatly degrades FedAvg's performance, posing a severe challenge to scale up the training.




▲ Figure 2: Overview of our proposed DGA.




▲ Figure 3: Our benchmark Pi Farm setup and speedup comparison.

## FURTHER READING

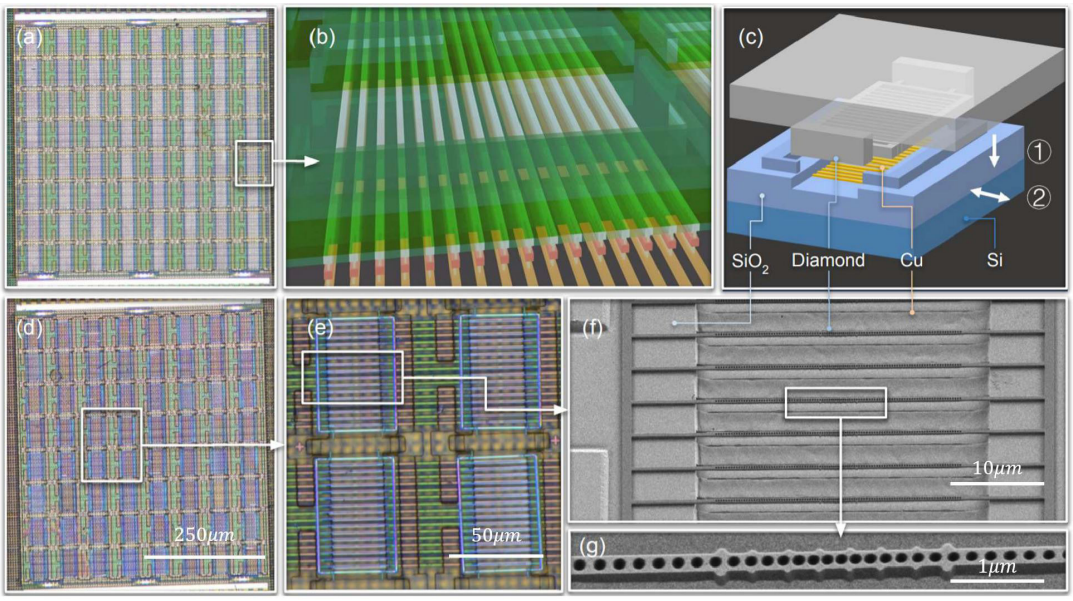

- H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *arXiv*, 2016.
- L. Zhu, H. Lin, Y. Lu, Y. Lin, and S. Han, "Delayed Gradient Averaging: Tolerate the Communication Latency for Federated Learning," *NeurIPS* 2021.

# Scalable Quantum Information Processing Architecture Using a Programmable Array of Spin-photon Interfaces

L. Li, L. D. Santis, I. Harris, K. C. Chen, Y. Song, I. Christen, M. Trusheim, C. E. Herranz, R. Han, D. R. Englund
Sponsorship: MITRE, Center for Integrated Quantum Materials, NSF

A central challenge in quantum information processing is to generate a large-scale entanglement of quantum systems. A leading hardware platform consists of qubits in the form of spin states of color centers in diamond. However, it is estimated that for general-purpose quantum information processors, millions of qubits will be required, motivating the need for hardware architectures that are highly scalable using modern semiconductor integration systems.

Here, we demonstrate a scalable quantum information processing architecture in a proof of concept consisting of a 2D array of tin-vacancy centers, addressable and tunable across thousands of diamond cavities hybrid integrated on a control chip based on the foundry process. We demonstrate core capabilities including tuning of the color center emission wavelength, spin initialization, and single-shot spin readout. The above works together are a proof of concept for a freely scalable architecture capable of hosting thousands toward millions of qubits.




▲Figure 1: (a) The complementary metal-oxide-semiconductor (CMOS) chip after surface post-processing. (b) CMOS chip region marked by the white box in (a) with chiplet locking structure and metal routing. (c) The scalable hybrid integration illustration. (d) The 1024 diamond cavities hybrid integrated on CMOS control chip. (e) The zoom-in optical microscope image. (f) The diamond chiplet scanning electron microscopy (SEM) image. (g) The SEM image of the perturbative cavity design optimized for free space collection.

## FURTHER READING:


- L. Li, L. D. Santis, I. Harris, K. C. Chen, Y. Song, I. Christen, M. Trusheim, C. E. Herranz, R. Han, and D. Englund. "Scalable Quantum Information Processing Architecture Using a Programmable Array of Spin-photon Interfaces," *CLEO: QELS_Fundamental Science*. Optical Society of America, 2022.

# Center for Integrated Circuits and Systems

Professor Hae-Seung Lee, Director

The Center for Integrated Circuits and Systems (CICS) at MIT, established in 1998, is an industrial consortium created to promote new research initiatives in circuits and systems design, as well as to promote a tighter technical relationship between MIT's research and relevant industry. Eight faculty members participate in the CICS: Director Hae-Seung (Harry) Lee, Anantha Chandrakasan, Ruonan Han, Song Han, David Perreault, Negar Reiskarimian, Charles Sodini, and Vivienne Sze.

CICS investigates a wide range of circuits and systems, including wireless and wireline communication, high-speed, THz, and RF circuits, microsensor/actuator systems, imagers, digital and analog signal processing circuits, biomedical circuits, deep learning systems, hardware security, emerging technologies, and power conversion circuits, among others.

We strongly believe in the synergistic relationship between industry and academia, especially in practical research areas of integrated circuits and systems. CICS is designed to be the conduit for such synergy.

CICS's research portfolio includes all research projects that the eight participating faculty members conduct, regardless of source(s) of funding, with a few exceptions.

Technical interaction between industry and MIT researchers occurs on both a broad and individual level. Since its inception, CICS recognized the importance of holding technical meetings to facilitate communication among MIT faculty, students, and industry. We hold two informal technical meetings per year open to CICS faculty, students, and representatives from participating companies. Throughout each full-day meeting, faculty and students present their research, often presenting early concepts, designs, and results that have not been published yet. The participants then offer valuable technical feedback, as well as suggestions for future research. The meeting also serves as a valuable networking event for both participants and students. Closer technical interaction between MIT researchers and industry takes place during work on projects of particular interest to participating companies. Companies may invite students to give on-site presentations, or they may offer students summer employment. Additionally, companies may send visiting scholars to MIT or enter into a separate research contract for more focused research for their particular interest. The result is truly synergistic, and it will have a lasting impact on the field of integrated circuits and systems.

## Anantha P. Chandrakasan

Dean of Engineering, Vannevar Bush Professor of Electrical Engineering & Computer Science
Department of Electrical Engineering and Computer Science


*Design of digital integrated circuits and systems. Energy efficient implementation of signal processing, communication and medical systems. Circuit design with emerging technologies.*


**Rm. 38-107 | 617-258-7619 | anantha @ mit . edu**


### POSTDOCTORAL ASSOCIATE

Yeseul Jeon, RLE

### GRADUATE STUDENTS

Aya Amer, EECS
Maitreyi Ashok, EECS
Ruicong Chen, EECS (co-supervised with H. Lee)
Adam Gierlach, EECS (co-supervised with G. Traverso)
Alex Ji, EECS
Jaeyoung Jung, EECS
Dimple Kochar, EECS
Eunseok Lee, EECS (co-supervised with R. Han)
Kyungmi Lee, EECS
Saurav Maji, EECS
Vipasha Mittal, EECS (co-supervised with H-S. Lee)
Rishabh Mittal, EECS (co-supervised with H-S. Lee)
Zoey Song, EECS
Miaorong Wang, EECS
Jongchan Woo, EECS (co-supervised with Rabia T. Yazicigil)
So-Yoon Yang, EECS (co-supervised with G. Traverso)
Deniz Yildirim, EECS

### VISITING SCHOLARS

Rabia Tugce Yazicigil, Boston University

### SUPPORT STAFF

Jessie-Leigh Thomas, Administrative Assistant

### SELECTED PUBLICATIONS


A. Maitreyi, M. J. Turner, R. L Walsworth, E. V. Levine, and A. P. Chandrakasan, "Hardware Trojan Detection Using Unsupervised Deep Learning on Quantum Diamond Microscope Magnetic Field Images," *ACM J. on Emerging Technologies in Computing Systems*, Apr. 2022.

A. Maitreyi, E. Levine, and A. P. Chandrakasan, "Randomized Switching SAR (RS-SAR) ADC Pro-tections for Power and Electromagnetic Side Channel Security," I*EEE Custom Integrated Circuits Conference (CICC)*, Apr. 2022.

G. Preet, V. Schaffer, B. Haroun, S. Ramaswamy, J. Wieser, J. Lang, and A. P. Chandrraksan, "A Duty-Cycled Integrated-Fluxgate Magnetometer for Current Sensing," *IEEE J. of Solid-State Circuits*, Mar. 2022.

S. Maji S., U. Benerjee, S. H. Fuller, and A. P. Chandrakasan, "A Threshold-Implementation-Based Neural-Network Accelerator Securing Model Parameters and Inputs Against Power Side-Channel Attacks," *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2022.

M. I. W. Khan, J. Woo, X. Yi, M. I. Ibrahim, R. T. Yazicigil, A. P. Chandrakasan, and R. Han, "A 0.31-THz Orbital-Angular-Momentum (OAM) Wave Transceiver in CMOS With Bits-to-OAM Mode Mapping," *IEEE J. of Solid-State Circuits*, pp.1-1, Jan. 2022.

K. Lee, and A. P. Chandrakasan, "Understanding the Energy vs. Adversarial Robustness Trade-Off in Deep Neural Networks," *IEEE Workshop on Signal Processing Systems (SiPS)*, Oct. 2021.

M. I. W. Khan, J. Woo, X. Yi, M. I. Ibrahim, R. T. Yazicigil, A. P. Chandrakasan, and R. Han, "A 0.31THz CMOS Uniform Circular Antenna Array Enabling Generation/Detection of Waves with Orbital-Angular Momentum," *IEEE Radio-Frequency Integrated Circuit Symposium (RFIC)*, Jun. 2021.

X. Yi, C. Wang, Z. Hu, J. W. Holloway, M. I. W. Khan, M. I. Ibrahim, M. Kim, G. C. Do-giamis, B. Perkins, M. Kaynak, R. T. Yazicigil, A. P. Chandrakasan, R. Han, "Emerging Terahertz Integrated Systems in Silicon," *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp.1-14, Jun. 2021.

T. Jeong, A. P. Chandrakasan, and H.-S. Lee, "S2ADC: A 12-bit, 1.25MS/s Secure SAR ADC with Power Side-Channel Attack Resistance," *IEEE J. of Solid-State Circuits*, vol. 56, no. 3, pp.844-854, Mar. 2021.

## Song Han

Associate Professor
Department of Electrical Engineering & Computer Science

*Machine learning, artificial intelligence, model compression, hardware accelerator, domain-specific architecture.*

**Rm. 38-344 | 617-253-0086 | songhan @ mit . edu**


### POSTDOCTORAL ASSOCIATES

Wei-Ming Chen, EECS
Wei-Chen Wang, EECS

### GRADUATE STUDENTS

Han Cai, EECS
Ji Lin, EECS
Yujun Lin, EECS
Zhijian Liu, EECS
Haotian Tang, EECS
Hanrui Wang, EECS
Zhekai Zhang, EECS
Ligeng Zhu, EECS

### UNDERGRADUATE STUDENTS

Kevin Shao, EECS

### SUPPORT STAFF

Jami L. Mitchell, Administrative Assistant

### SELECTED PUBLICATIONS


J. Lin, W. Chen, H. Cai, C. Gan, and S. Han, "MCUNet-v2: Memory-Efficient Inference for Tiny Deep Learning," *Neural Information Processing System(NeurIPS)*, 2021.

L. Zhu, H. Lin, Y. Lu, Y. Lin, and S. Han, "Delayed Gradient Averaging: Tolerate the Communication Latency for Federated Learning," *Neural Information Processing System(NeurIPS)*, 2021.

Y. Lin, Z. Zhang, H. Tang, H. Wang, and S. Han, "PointAcc: Efficient Point Cloud Accelerator," *International Symposium on Microarchitecture (MICRO)*, 2021.

Y. Lin, M. Yang, and S. Han, "NAAS: Neural Accelerator Architecture Search," *Design Automation Conference (DAC)*, 2021.

Y. Ding, L. Zhu, Z. Jia, G. Pekhimenko, and S. Han, "IOS: Inter-Operator Scheduler For CNN Acceleration,"*Conference on Machine Learning and Systems (MLSys)*, 2021.

J. Lin, W.-M. Chen, Y. Lin, J. Cohn, C. Gan, and S. Han, "MCUNet: Tiny Deep Learning on IoT Devices," *Neural Information Processing System(NeurIPS)*, spotlight presentation, 2020.

H. Cai, C. Gan, L. Zhu, and S. Han, "Tiny Transfer Learning: Reduce Activations, not Trainable Parameters for Efficient On-Device Learning," *Neural Information Processing System (NeurIPS)*, 2020.

S. Zhao, Z. Liu, J. Lin, J. Zhu, and S. Han, "Differentiable Augmentation for Data-Efficient GAN Training," *Neural Information Processing System (NeurIPS)*, 2020.

H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once For All: Train One Network and Specialize It for Efficient Deployment," *International Conference on Learning Representations (ICLR)*, 2020.

H. Wang, K. Wang, J. Yang, L. Shen, N. Sun, H.-S. Lee, and S. Han, "Transferable Transistor Sizing with Graph Neural Networks and Reinforcement Learning," *Design Automation Conference (DAC)*, 2020.

Z. Zhang, H. Wang, S. Han, and B. Dally, "SpArch: Efficient Architecture for Sparse Matrix Multiplication," *International Symposium on High-Performance Computer Architecture (HPCA)*, 2020.

Z. Liu, H. Tang, Y. Lin, and S. Han, "Point Voxel CNN for Efficient 3D Deep Learning," *Neural Information Processing System (NeurIPS)*, Spotlight presentation.2019.

L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," *Neural Information Processing System (NeurIPS)*, 2019.

J. Lin, C. Gan, and S. Han, "TSM: Temporal Shift Module for Efficient Video Understanding," *International Conference on Computer Vision (ICCV)*, 2019.

H. Cai, L. Zhu, and S. Han, "ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware," *International Conference on Learning Representations (ICLR)*, 2019.

## Hae-Seung Lee

Director of Microsystems Technology Laboratories
Director of Center for Integrated Circuits and Systems
ATSC Professor of Electrical Engineering & Computer Science
Department of Electrical Engineering & Computer Science

*Analog and Mixed-signal Integrated Circuits, with a Particular Emphasis in Data Conversion Circuits in scaled CMOS.*

**Rm. 39-521 | 617-253-5174 | hslee @ mtl . mit . edu**


## POSTDOCTORAL ASSOCIATE

Anand Chandrasekhar, EECS

## GRADUATE STUDENTS

Ruicong Chen, EECS
Rebecca Ho, EECS
Mina Kim, EECS
Jaehwan Kim, EECS
Rishabh Mittal, EECS
Vipasha Mittal, EECS

## SUPPORT STAFF

Elizabeth Green, Sr. Administrative Assistant
Elizabeth Kubicki, Administrative Assistant

## PUBLICATIONS


J. Seo, H.-S. Lee, and C. Sodini, "Non-invasive Evaluation of a Carotid Arterial Pressure Waveform Using Motion-Tolerant Ultrsound Measurements During Valsalva Maneuver," *IEEE J. of Biomedical and Health Informatics*, vol. 25, pp. 163-174, Jan. 2021

T. Jeong, A. Chandrakasan, and H.-S. Lee, "S2ADC: A 12-bit, 1.25MS/s Secure SAR ADC with Power Side-Channel Attack Resistance," *IEEE J. Solid-State Circuits*, vol. SC-53, pp. 844-854 Mar. 2021.

M. Kim, C. Wang, L. Yi, H.-S. Lee, and R. Han, "A Sub-THz CMOS Molecular Clock with 20 ppt Stability at 10,000 s Based on A Dual-Loop Spectroscopic Detection and Digital Frequency Error Integration," to appear in *2022 IEEE RFIC*, Jun. 19-21, 2022, Denver, CO.

R. Chen, A. Chandrakasan, and H.-S. Lee, "RaM-SAR: A Low Energy and Area Overhead, 11.3fJ/conv.-step. 12b 25MS/s Secure Random-mapping SAR ADC with Power and EM Side-channel Attack Protection," to appear in *2022 IEEE Symposium on VLIS Circuits*, Jun. 12-17, 2022, Hawaii, HI.

H.-S. Lee and D. Daly, "Push-pull dynamic amplifier circuits," U.S. Patent 10,951,184, Mar. 16, 2021

H.-S. Lee, "Analog current memory with droop compensation," U.S. Patent 11,031,090, Jun. 8, 2021

H.-S. Lee, "Photo receiver circuits", U.S. Patent 11,067,439, Jul. 20, 2021

H.-S. Lee, "Constant level-shift buffer amplifier circuits," U.S. Patent 11,114,986, Sep. 7, 2021

# Negar Reiskarimian

X-Window Consortium Career Development Assistant Professor
Department of Electrical Engineering & Computer Science

*Integrated circuits and systems and applied electromagnetics with a focus on analog, RF, millimeter-Wave (mm-Wave) and optical integrated circuits, metamaterials and systems for a variety of applications.*

**Rm. 39-427a | 617-253-0726 | negarr @ mit . edu**


## GRADUATE STUDENTS

Soroush Araei, EECS
Shahabeddin Mohin, EECS

## SUPPORT STAFF

Jami L. Mitchell, Administrative Assistant

## SELECTED PUBLICATIONS


N. Reiskarimian, "A Review of Nonmagnetic Nonreciprocal Electronic Devices: Recent advances in nonmagnetic nonreciprocal components," *IEEE Solid-State Circuits Magazine* vol. 13, no. 4, pp. 112-121, Fall 2021.

N. Reiskarimian, M. Khorshidian, and H. Krishnaswamy, "Inductorless, Widely Tunable N-Path Shekel Circulators Based on Harmonic Engineering," *IEEE J. of Solid-State Circuits (JSSC)* (invited), vol. 56, no. 4, Apr. 2021.

A. Nagulu, N. Reiskarimian and H. Krishnaswamy, "Non-reciprocal Electronics Based on Temporal Modulation," *Nature Electronics*, May 2020.

N. Reiskarimian, M. Khorshidian and H. Krishnaswamy, "Inductorless, Widely-Tunable N-Path Shekel Circulators Based on Harmonic Engineering," in *IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, pp 39-42, Jun. 2020.

M. Khorshidian, N. Reiskarimian and H. Krishnaswamy, "A Compact Reconfigurable N-Path Low-Pass Filter Based on Negative Trans-Resistance with <1dB Loss and >21dB Out-of-Band Rejection," in *IEEE International Microwave Symposium (IMS)*, pp. 799-802, Jun. 2020.

M. Khorshidian, N. Reiskarimian and H. Krishnaswamy, "High-Performance Isolators and Notch Filters based on N-Path Negative Trans-Resistance," in *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2020.

M. Baraani Dastjerdi, S. Jain, N. Reiskarimian, A. Natarajan and H. Krishnaswamy, "Analysis and Design of Full-Duplex 2x2 MIMO Circulator-Receiver with High TX power handling Exploiting MIMO RF and Shared-delay Baseband Self-Interference Cancellation," *IEEE J. of Solid State Circuits (JSSC)* (invited), vol. 54, no. 12, pp. 3525-3540, Dec. 2019.

N. Reiskarimian, M. Tymchenko, A. Alu and H. Krishnaswamy, "Breaking Time-Reversal Symmetry Within Infinitesimal Dimensions Through Staggered Switched Networks," in *Metamaterials*, Sep. 2019.

N. Reiskarimian, A. Nagulu, T. Dinc and H. Krishnaswamy, "Non-Reciprocal Devices: A Hypothesis Turned into Reality," *IEEE Microwave Magazine* (invited), vol. 20, no. 4, pp. 94-111, Apr. 2019.

N. Reiskarimian, T. Dinc, J. Zhou, T. Chen, M. Baraani Dastjerdi, J. Diakonikolas, G. Zussman, and H. Krishnaswamy, "A One-Way Ramp to a Two-Way Highway: Integrated Magnetic-Free Non-Reciprocal Antenna Interfaces for Full Duplex Wireless," in *IEEE Microwave Magazine* (invited), vol. 20, no. 2, pp. 56-75, Feb. 2019.

M. B. Dastjerdi, S. Jain, N. Reiskarimian, A. Natarajan, and H. Krishnaswamy, "Full-Duplex 2x2 MIMO Circulator-Receiver with High TX Power Handling Exploiting MIMO RF and Shared-Delay Baseband Self-Interference Cancellation," *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2019.

## Charles G. Sodini

LeBel Professor
Department of Electrical Engineering & Computer Science


*Electronics and integrated circuit design and technology. Specifically, his research involves technology intensive integrated circuit and systems design, with application toward medical electronic devices for personal monitoring of clinically relevant physiological signals.*


**Rm. 39-527b | 617-253-4938 | sodini @ mtl . mit . edu**


### COLLABORATORS

Sam Fuller, Analog Devices, Inc
Thomas O'Dwyer, MTL Research Affiliate
Joohyun Seo, Analog Devices, Inc.

### POSTDOCTORAL ASSOCIATES

Anand Chandraksekhar, MTL

### GRADUATE STUDENT

Jeanne Harabedian, MTL

### SUPPORT STAFF

Kathleen Brody, Administrative Assistant

### SELECTED PUBLICATIONS


S. M. Imaduddin, C. G. Sodini, and T. Heldt, "Deconvolution-based partial volume correction for volumetric blood flow measurement," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control,* Accepted for publication, Apr. 2022

J. Seo, H.-S. Lee, and C. G. Sodini, "Non-Invasive Evaluation of a Carotid Arterial Pressure Waveform Using Motion-Tolerant Ultrasound Measurements During the Valsalva Maneuver," *IEEE J. of Biomedical and Health Informatics*, vol. 25, no. 1, Jan. 2021.

H. Lai, G. Saavedra-Peña, C. G. Sodini, V. Sze and T. Heldt, "Measuring Saccade Latency Using Smartphone Cameras," *IEEE J. of Biomedical and Health Informatics*, vol. 24, no. 3, pp. 885-897, Mar. 2020.

J. Liu, C. G. Sodini, Y. Ou, B. Yan, Y. T. Zhang, N. Zhao, "Feasibility of Fingertip Oscillometric Blood Pressure Measurement: Model-based Analysis and Experimental Validation," *IEEE J. of Biomedical and Health Informatics (JBHI)*, May 2019.

M. Delano, and C. Sodini, "Evaluating Calf Bioimpedance Measurements for Fluid Overload Management in a Controlled Environment," *Physiological Measurement*, vol. 39, no. 12, p. 125009, 2018.

## Vivienne Sze

Associate Professor of Electrical Engineering & Computer Science
Department of Electrical Engineering & Computer Science


*Joint design of signal processing algorithms, architectures, VLSI and systems for energy-efficient implementations. Applications include computer vision, machine learning, autonomous navigation, image processing and video coding.*


**Rm. 38-260 | 617-324-7352 | sze @ mit . edu**


## GRADUATE STUDENTS

Tanner Andrulis (co-advised with Joel Emer)
Keshav Gupta (co-advised with Sertac Karaman)
Jamie Koerner, EECS (co-advised with Thomas Heldt)
Peter Li, EECS (co-advised with Sertac Karaman)
Soumya Sudhakar, AeroAstro (co-advised with Sertac Karaman)
Yannan Nellie Wu, EECS (co-advised with Joel Emer)

## UNDERGRADUATE STUDENTS

Michael Gilbert, EECS
Ari Grayzel, AeroAstro

## SUPPORT STAFF

Janice L. Balzer, Administrative Assistant

## SELECTED PUBLICATIONS


P. Z. X. Li, S. Karaman, and V. Sze, "Memory-Efficient Gaussian Fitting for Depth Images in Real Time," *IEEE International Conference on Robotics and Automation (ICRA)*, May 2022.

S. Sudhakar, V. Sze, and S. Karaman, "Uncertainty from Motion for DNN Monocular Depth Estimation," *IEEE International Conference on Robotics and Automation (ICRA)*, May 2022.

H.-Y. Lai, G. Saavedra-Pena, C. G. Sodini, T. Heldt, and V. Sze, "App-based saccade latency and directional error determination across the adult age spectrum," *IEEE Transactions on Biomedical Engineering (TBME)*, Vol. 69, No. 2, pp. 1029 – 1039, Feb. 2022.

Y.-L. Liao, S. Karaman, V. Sze, "Searching for Efficient Multi-Stage Vision Transformers," *ICCV 2021 Workshop on Neural Architectures: Past, Present and Future*, Oct. 2021.

K. Gupta, P. Z. X. Li, S. Karaman, V. Sze, "Efficient Computation of Map-scale Continuous Mutual Information on Chip in Real Time," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept. 2021.

T.-J. Yang, Y.-L. Liao, and V. Sze, "NetAdapt v2: Efficient Neural Architecture Search with Fast Super-Network Training and Architecture Optimization," *Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021.

F. Wang, Y. Wu, M. Woicik, V. Sze, and J. S. Emer, "Architecture-Level Energy Estimation for Heterogeneous Computing Systems," *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, Mar. 2021.

Y. Wu, P. Tsai, A. Parashar, V. Sze, and J. Emer, "Sparseloop: An Analytical, Energy-Focused Design Space Exploration Methodology for Sparse Tensor Accelerators," *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, Mar. 2021.

J. Ray, A. Brahmakshatriya, R. Wang, S. Kamil, A. Reuther, V. Sze, and S. Amarasinghe, "Domain-Specific Language Abstractions for Compression," *Data Compression Conference (DCC)*, Mar. 2021.

L. Bernstein, A. Sludds, R. Hamerly, V. Sze, J. Emer, and D. Englund, "Freely-scalable, reconfigurable optical hardware for deep learning," *Scientific Reports*, vol. 11, no. 3144, Feb. 2021.

V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks," *Synthesis Lectures on Computer Architecture* - Morgan & Claypool Publishers, 2020.

## IN APPRECIATION OF OUR MICROSYSTEMS INDUSTRIAL GROUP MEMBER COMPANIES:

Analog Devices, Inc.
Applied Materials
Draper
Edwards
HARTING
Hitachi High-Tech Corporation
IBM
Lam Research
NEC
TSMC
Texas Instruments

## AND MIT.NANO CONSORTIUM MEMBER COMPANIES:

Agilent Technologies
Analog Devices, Inc.
Dow
Draper
DSM
Edwards
Fujikura
IBM
Lam Research
NCSOFT
NEC
Oxford Instruments Asylum Research
Raith
Waters