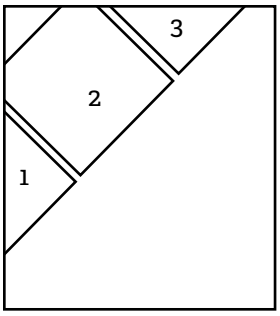
2020 Annual Research Report

MTL MICROSYSTEMS TECHNOLOGY LABORATORIES

MIT.nano

MASSACHUSETTS INSTITUTE OF TECHNOLOGY



Front Cover Credits

1. Pine (P. radiata) cells grown in liquid culture and marked with fluorescent probes to indicate live cells (green) and the cell walls of all cells (blue).
2. Using AI to Make Better AI: New approach brings faster, AI-optimized AI that runs efficiently on IoT devices
3. A monolithic array of SS 316L 3D-printed MEMS corona ionizers and close-up of a single tip; devices can be used as electrohydrodynamic gas pumps.

MTL Annual Research Report 2020

Hae-Seung Lee
Vladimir Bulovic
Shereece Beckford
Elizabeth Fox
Tina Gilman
Elizabeth Green
Stacy McDaid
Meghan Melvin
Jami Mitchell

Research Abstracts

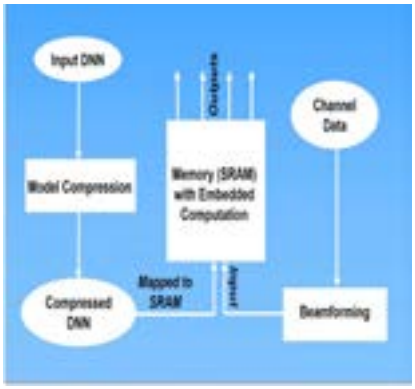
Conformable Ultrasound Patch with Energy-efficient In-memory Computation for Bladder Volume Monitoring

K. Brahma, L. Zhang, V. Kumar, A. P. Chandrakasan, C. Dagdeviren, A. E. Samir, Y. C. Eldar
Sponsorship: Texas Instruments

Continuous monitoring of urinary bladder volume aids management of common conditions such as post-operative urinary retention. Urinary retention is prevented by catheterization, an invasive procedure that greatly increases urinary tract infection. Ultrasound imaging has been used to estimate bladder volume as it is portable, non-ionizing, and low-cost. Despite this, ultrasound technology faces fundamental challenges limiting its usability for next generation wearable technologies. (1) Current ultrasound probes cannot cover curved human body parts or perform whole-organ imaging with high spatiotemporal resolution. (2) Current systems require skilled manual scanning with attendant measurement variability. (3) Current systems are insufficiently energy-efficient to permit ubiquitous wearable device deployment.

We are developing an energy-efficient body contour conformal ultrasound patch capable of real-time bladder volume monitoring. This system will incorporate (1) deep neural network- (DNN) based segmentation algorithms to generate spatiotemporally accurate bladder volume estimates and (2) energy-efficient static random-access memory (SRAM) with in-memory dot-product computation for low-power segmentation network implementation. We aim to develop platform technology embodiments deployable across a wide range of health-monitoring wearable device applications requiring accurate, real-time, and autonomous tissue monitoring.

We are training a low-precision (pruned and quantized weights) DNN for accurate bladder segmentation. DNNs are computation-intensive and require large amounts of storage due to high dimensionality data structures with millions of model parameters. This shifts the design emphasis towards data movement between memory and compute blocks. Matrix vector multiplications (MVM) are a dominant kernel in DNNs, and In-Memory computation can use the structural alignment of a 2D SRAM array and the data flow in matrix vector multiplications to reduce energy consumption and increase system throughput.



▲ Figure 1: The flowchart of an energy-efficient system implementing a compressed segmentation network using SRAM designed for in-memory dot product computation.

FURTHER READING

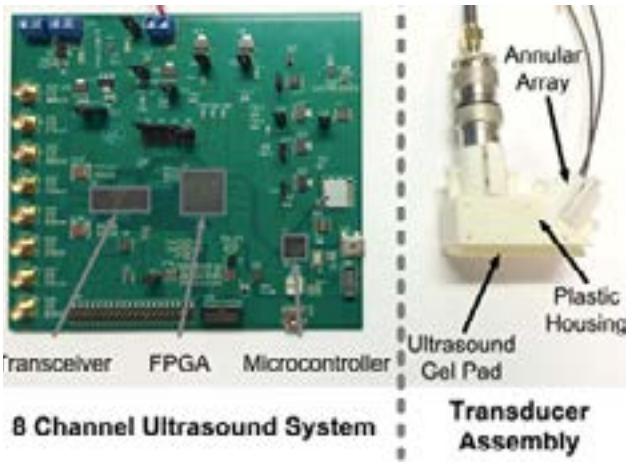
- C. M. W. Daft, "Conformable Transducers for Large-volume, Operator-independent Imaging," *2010 IEEE International Ultrasonics Symposium*, pp. 798-808, 2010.
- R. J. V. Sloun, R. Cohen, and Y. C. Eldar, "Deep Learning in Ultrasound Imaging," arXiv, vol. 1907, p. 02994, 2019.
- A. Biswas and A. P. Chandrakasan, "Conv-RAM: An Energy-efficient SRAM with Embedded Convolution Computation for Low-power CNN-based Machine Learning Applications," *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, pp. 488-490, 2018

Arterial Blood Pressure Estimation Using Ultrasound Technology

A. Chandrasekhar, C. G. Sodini, H.-S. Lee
Sponsorship: MEDRC-Philips, MIT J-Clinic, CICS

Hypertension, or high blood pressure (BP), is a major risk factor for cardiovascular diseases. Doctors prefer monitoring BP waveforms of ICU patients as the morphology and absolute values of these signals help to assert the cardiovascular fitness of the patient. At present, doctors use invasive radial catheters to record these waveforms. Invasive transducers are inconvenient and can be painful and risky to the patient. Hence, we are developing an algorithm to estimate BP waveforms using non-invasive ultrasound measurements at the brachial and carotid arteries.

Ultrasound probes are a commonly used sensing modality for non-invasive cardiovascular imaging. For instance, doctors use a linear array transducer to image superficial blood vessels like the brachial or the carotid artery. These multifunctional probes can record the lumen area waveform of these arteries and measure the velocity of the blood. In this project, we will record the aforementioned signals with a commercial ultrasound probe and a custom-designed probe (see Figure 1) and use the physics of the arterial pulse wave transmission to estimate the shape and absolute values of the pressure waveform. The pressure waves originating from the heart traverse the arterial wall with a velocity commonly referred to as pulse wave velocity (PWV). According to the physics of the arterial pulse wave transmission, we can calculate PWV from the ultrasound signals. Compliance and pulse pressure of the pressure waves in the artery may be obtained using the Bramwell-Hill equation. Finally, absolute values of the pressure will be derived using a combination of a transmission line model of the artery and machine learning algorithms.



▲Figure 1: Final design of the ultrasound based probe.

▲Figure 2: Ultrasound transducer placed above the carotid artery.

FURTHER READING

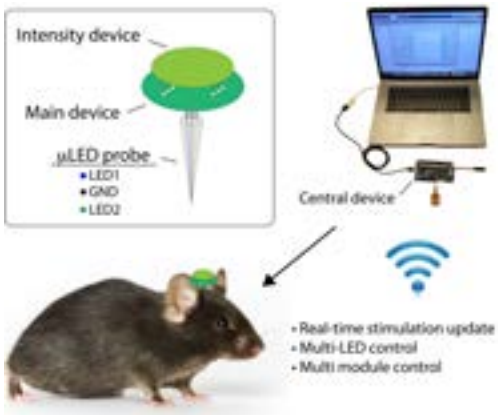
- J. Seo, "Noninvasive Arterial Blood Pressure Waveform Monitoring using Two-element Ultrasound System," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 62, no. 4, pp. 776-784, 2015.

Modular Optoelectronic System for Wireless, Programmable Neuromodulation

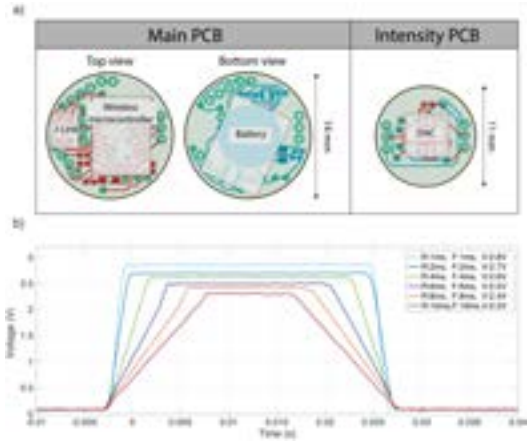
S. Orguc*, J. Sands*, A. Sahasrabudhe, P. Anikeeva, A. P. Chandrakasan
Sponsorship: Delta Electronics

Optogenetics is a technique that uses visible light stimulation to activate or inhibit neurons genetically modified to express light-sensitive proteins from the microbial rhodopsin family. It offers light-sensitive opsin proteins to the region of interest and provide advantages such as cell type specificity, millisecond temporal precision, and rapid reversibility. Furthermore, compared to the electrical stimulation, it causes negligible electrical perturbation to the environment, which enables simultaneous electrical recording while stimulating a region of interest. The stimulation of the targeted neurons can be achieved using lasers, light-emitting diode (LED)-coupled optical fibers, or wireless μLEDs.

This work presents a modular, light-weight head-borne neuromodulation platform that achieves low-power wireless neuromodulation and allows real-time programmability of the stimulation parameters such as the frequency, duty cycle, and intensity. This platform is composed of two parts: the main device and the optional intensity module (Figure 1). The main device is functional independently; however, the intensity control module can be introduced on demand (Figure 2). The stimulation is achieved through the use of LEDs directly integrated in the custom-drawn fiber-based probes. Our platform can control up to 4 devices simultaneously, and each device can control multiple LEDs in a given subject. Our hardware uses off-the-shelf components and has a plug-and-play structure, which allows for fast turnover time and eliminates the need for complex surgeries. The rechargeable, battery-powered wireless platform uses Bluetooth Low Energy (BLE) and is capable of providing stable power and communication regardless of orientation. This platform presents a potential advantage over the battery-free, fully implantable systems that rely on wireless power transfer, which is typically direction-dependent, requires sophisticated implantation surgeries, and demands complex experimental apparatuses. Although the battery life is limited to several hours, this is sufficient to complete the majority of behavioral neuroscience experiments. Our platform consumes 0.5 mW and has a battery life of 12 hours.



▲ Figure 1: System overview. The platform communicates with the central device for stimulation updates.



▲ Figure 2: a) The PCB design of the main device and the intensity device. b) Programmable intensity control demonstration.

FURTHER READING

- O. Yizhar, L. E. Fenno, T. J. Davidson, M. Mogri, and K. Deisseroth, "Optogenetics in Neural Systems," *Neuron*, vol. 71, no. 1, pp. 9–34, 2011.
- Y. Jia, W. Khan, B. Lee, B. Fan, F. Madi, A. Weber, W. Li, and M. Ghovanloo, "Wireless Opto-electro Neural Interface for Experiments with Small Freely Behaving Animals," *Journal of Neural Engineering*, vol. 15, no. 4, p. 046032, 2018.
- P. Gutruf, V. Krishnamurthi, A. Vazquez-Guardado, Z. Xie, A. Banks, C.-J. Su, Y. Xu, C. R. Haney, E. A. Waters, I. Kandela, "Fully Implantable Optoelectronic Systems for Battery-free, Multimodal Operation in Neuroscience Research," *Nature Electronics*, vol. 1, no. 12, pp. 652–660, 2018.

DC-DC Converter Implementations Based on Piezoelectric Resonators

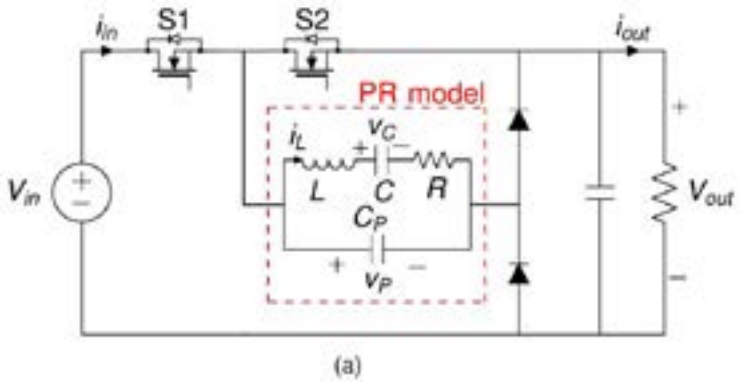
J. D. Boles, J. J. Piel, D. J. Perreault
Sponsorship: Texas Instruments, NSF Graduate Research Fellowship, MASDAR

Power electronics play a vital role in the technological advancement of transportation, energy systems, manufacturing, healthcare, information technology, and many other major industries. Demand for power electronics with smaller volume, lighter weight, and lower cost often motivates designs that better utilize a converter's energy storage components, particularly magnetics. However, the achievable power densities of magnetic components inherently reduce as volume decreases, so further progress in converter miniaturization will eventually require new energy storage mechanisms with fundamentally higher energy density and efficiency capabilities.

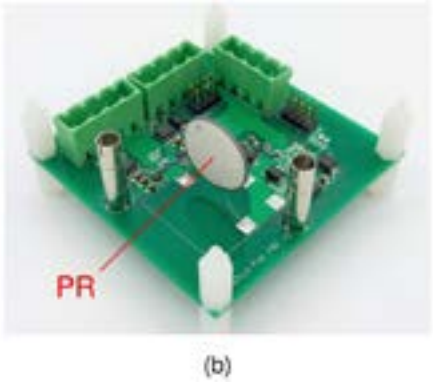
This prompts investigation into piezoelectric energy storage for power conversion; piezoelectrics have comparatively superior volume scaling properties. While piezoelectrics have been used extensively for sensing, actuation, transduction, and energy harvesting applications, their adoption in power conversion has been more limited. Converter designs based on single-port piezoelectric resonators (PRs) report limited power and/or performance capability, but without investigation into the full realm of possible converter implementations.

In this work, we conduct a systematic enumeration and downselection of practical dc-dc converter switching sequences and topologies that best leverage PRs as their only energy storage components. In particular, we focus on switching sequences that facilitate high-efficiency behaviors (e.g., low-loss resonant charging/discharging of the PR's input capacitance and all-positive instantaneous power transfer) with voltage regulation capability. To analyze and compare implementations, we demonstrate methods for mapping PR state trajectories across a switching cycle, imposing practical constraints on PR behavior, evaluating PR utilization, and estimating PR efficiency.

Effective use of the PR's resonant cycle enables these converter implementations to achieve strong experimental performance with peak efficiencies >99%, even with presently commercially-available PRs. This suggests that these PR-based converters are promising alternatives to those based on traditional energy storage. With further development, PR-based converters may pave the way for high-performance converter miniaturization in applications spanning consumer electronics, biomedical implants, and flight.



▲ Figure 1: Schematic of one PR-based converter topology, which relies on only the PR for energy storage and has only two active switches.



▲ Figure 2: Photograph of a converter prototype corresponding to this schematic.

FURTHER READING

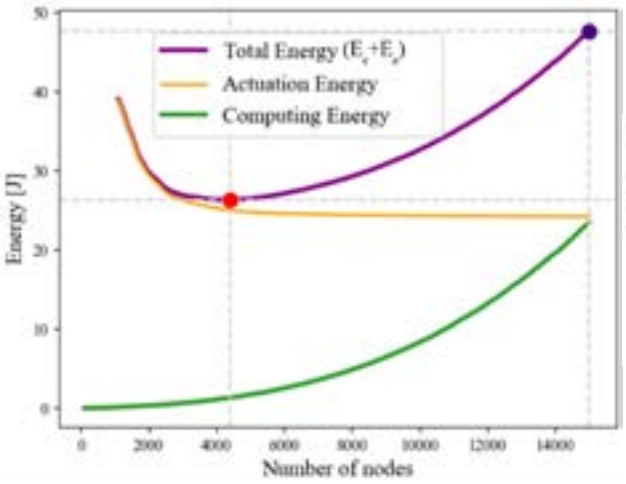
- J. D. Boles, J. J. Piel, and D. J. Perreault, "Enumeration and Analysis of Dc-dc Converter Implementations Based on Piezoelectric Resonators," in Proc. *IEEE Workshop on Control and Modeling for Power Electronics*, Toronto, Canada, Jun. 2019, pp. 1–8.
- J. D. Boles, J. J. Piel, and D. J. Perreault, "Analysis of High-efficiency Operating Modes for Piezoelectric Resonator-based Dc-dc Converters," in Proc. *IEEE Applied Power Electronics Conference and Exposition*, New Orleans, LA, USA, Mar. 2020, pp. 1–8.
- C.-S. Kim, S. R. Bishop, and H. L. Tuller, "Electro-chemo-mechanical Studies of Perovskite-structured Mixed Ionic-electronic Conducting $SrSn_{1-x}Fe_xO_{3-x/2+\delta}$ Part I: Defect Chemistry," *J. Electroceram.*, vol. 38, pp. 74-80, 2017.

Balancing Actuation and Computing Energy in Low-power Motion Planning

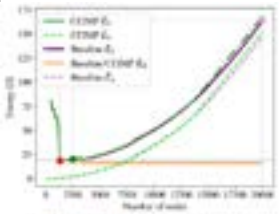
S. Sudhakar, V. Sze, S. Karaman
Sponsorship: NSF, Cyber-Physical Systems (CPS) Program

We study a novel class of motion planning problems, inspired by emerging low-power robotic vehicles, such as insect-size flyers, high-endurance autonomous blimps, and chip-size satellites for which the energy consumed by computing hardware while planning a path can be as large as the energy consumed by actuation hardware during the execution of the same path. For these new applications, we must consider the total energy of executing and computing a candidate solution to evaluate a motion plan. Figure 1 shows average actuation energy and computing energy curves for a selected robotic platform and computing platform. Here, minimizing only the actuation energy does not minimize the total energy. Instead, stopping computing earlier and accepting a higher actuation energy cost for a lower computing energy cost minimizes the total energy.

We propose a new algorithm, called Computing Energy Included Motion Planning (CEIMP). CEIMP operates similarly to other anytime planning algorithms, except it stops when it estimates further computing will require more computing energy than potential savings in actuation energy. The algorithm relies on Bayesian inference to estimate future energy savings to evaluate the trade-off between the computing energy required to continue sampling and the potential future actuation energy savings after such computation. CEIMP outperforms the average baseline of using maximum computing resources in realistic computational experiments involving 10 MIT building floor plans. On the ARM Cortex-A15, for a simulated vehicle that uses 1 Watt to travel 1 m/s, CEIMP saves 2.1-8.9x the total energy on average across floor plans compared to the baseline, translating to missions that can last 2.1-8.9x longer on the same battery. Figure 2 shows CEIMP in action; while the path returned by CEIMP is longer than the path returned by the baseline, CEIMP's total energy is much closer to the true minimum of total energy than the baseline.



▲Figure 1: Average computing energy, actuation energy, and total energy (computing + actuation) curves vs. nodes in PRM*, a sampling-based motion planner.



(a) Energy curves vs. n

(b) Path returned by CEIMP

(c) Path returned by baseline

▲Figure 2: Energy curves vs. nodes in PRM* for a single trial and paths returned by CEIMP and the baseline. True minimum of total energy curve (red marker), baseline total energy (purple marker), and CEIMP total energy (green marker) are marked.

FURTHER READING

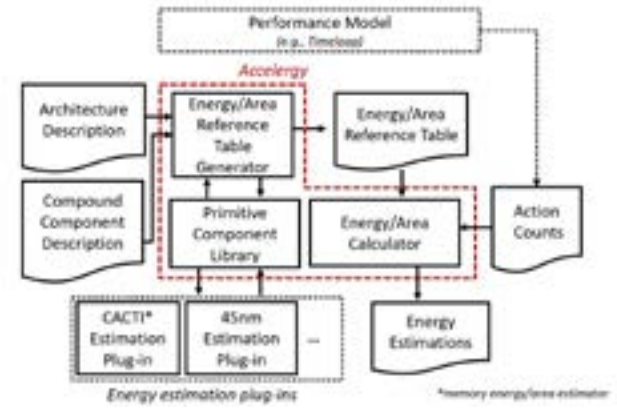
- S. Sudhakar, S. Karaman, and V. Sze, "Balancing Actuation and Computing Energy in Motion Planning," presented at *International Conference on Robotics and Automation (ICRA)*, Paris, France, 2020.

Architecture-level Energy Estimation of Accelerator Designs

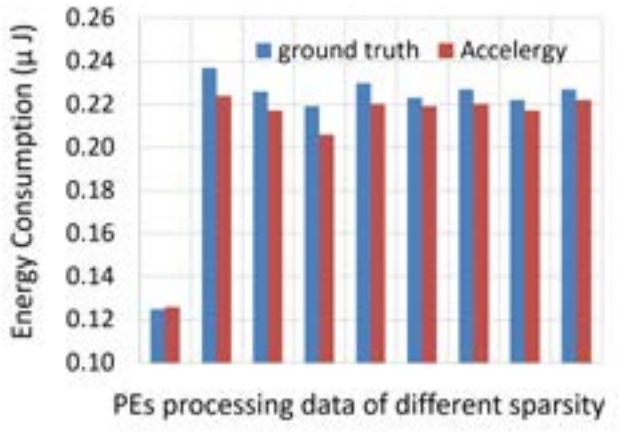
Y. N. Wu, J. S. Emer, V. Sze
Sponsorship: DARPA, MIT Presidential Fellowship, Facebook Faculty Award

With Moore's law slowing down and Dennard scaling ending, energy-efficient domain-specific accelerators have become a promising direction for hardware designers to continue bringing energy efficiency improvements to data and computation intensive applications. To ensure fast exploration of accelerator design space, architecture-level energy estimators, which perform energy estimations without requiring complete hardware description of the designs, are critical to designers. However, it is hard to use existing architecture-level energy estimators to obtain accurate estimates for accelerator designs, as accelerator designs are diverse and sensitive to data patterns.

To solve this problem, we present Accelergy (Figure 1), an architecture-level energy estimation methodology. Accelergy allows the users to define their own components in their designs to allow descriptions of the diverse design space. At the same time, to reflect the energy differences brought by special data patterns, e.g., sparsity in data, Accelergy also allows the users to define special actions types related to the components. To enhance flexibility, Accelergy defines an interface to communicate with other estimators that focus on energy estimations of specific types of components in the designs (e.g., memory storage components). To illustrate the usage of Accelergy methodology, we implemented an example framework for energy estimations of deep neural network (DNN) accelerator designs. We further integrate Accelergy with Timeloop, a DNN mapping space exploration tool, to enable accurate estimation of processing-in-memory (PIM) based DNN accelerator designs. We validated the Accelergy framework on a conventional digital design Eyeriss as well as a PIM-based design, both achieving a total energy estimation accuracy of 95% and accurate energy breakdowns of various components in the designs (Figure 2).



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



▲Figure 2: Energy estimation comparison on the energy breakdown across the PEs, each of which processes data of a different sparsity, in Eyeriss PE array (only selected PE are shown).

FURTHER READING

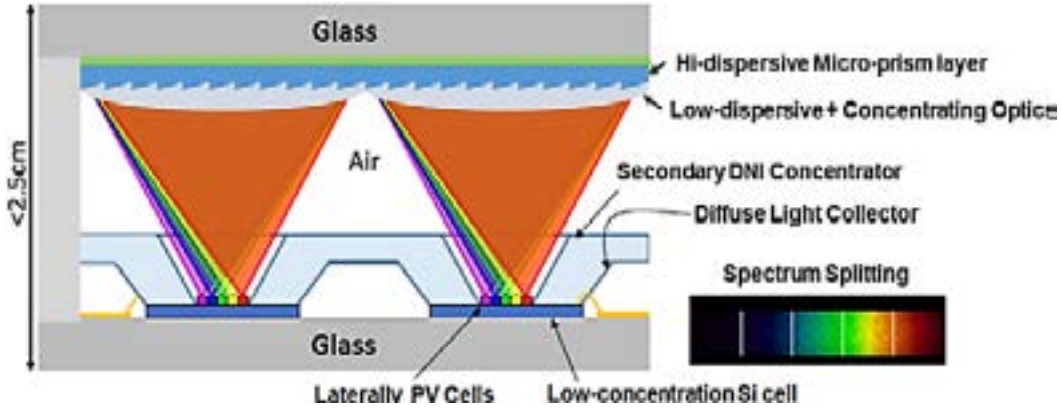
- Y. N. Wu, J. S. Emer, and V. Sze, "Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs," *ICCAD*, 2019.
- Y. N. Wu, V. Sze, and J. S. Emer, "An Architecture-Level Energy and Area Estimator for Processing-In-Memory Accelerator Designs," *ISPASS*, 2020.
- A. Parashar, P. Raina, Y. S. Shao, Y.-H. Chen, V. A. Ying, A. Mukkara, R. Venkatesan, B. Khailany, S. W. Keckler, and J. Emer, "Timeloop: A Systematic Approach to DNN Accelerator Evaluation," *ISPASS*, 2019.

A CMOS-based Energy Harvesting Approach for Laterally-arrayed Multi-bandgap Concentrated Photovoltaic Systems

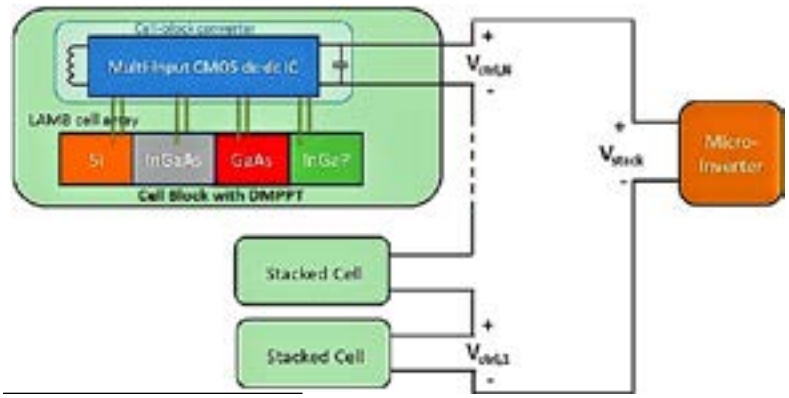
H. Zhang, K. Martynov, D. J. Perreault
Sponsorship: ARPA-E

When high solar conversion efficiency is desired, people often adopt concentrated photovoltaic systems with multi-junction cells. However, traditional tandem structures widely used in such systems can suffer from current-mismatch effects with spectrum variations, whereas the Laterally-Arrayed Multi-Bandgap (LAMB) cell structure is a potentially higher-efficiency and lower-cost alternative.

Here we show an energy harvesting approach designed to take full advantage of the LAMB cell structure. Individual cells within a sub-module block are connected for approximate voltage-matching, and a Multi-Input Single-Output (MISO) buck converter combines the energy and performs Maximum Power Point Tracking locally. A miniaturized MISO dc-dc converter prototype is developed in a 130nm CMOS process. For 45-160mW power levels, >95% peak efficiency is achieved in a small form factor designed to fit within available space in a LAMB cell block. The results demonstrate the potential of the LAMB CPV system for enhanced solar energy capture.



▲ Figure 1: Structure of a LAMB cell unit. An optical layer spectrally-splits and focuses direct sunlight onto multi-bandgap III-V cells, and a Si cell collects diffuse light.



◀ Figure 2: Proposed power management structure. Each cell block comprises several LAMB cell units, and a dc-dc converter that tracks local maximum power point and combines energy generated from multiple cells into a single output. The power from individual converters can then be combined, e.g. by stacking in series.

FURTHER READING

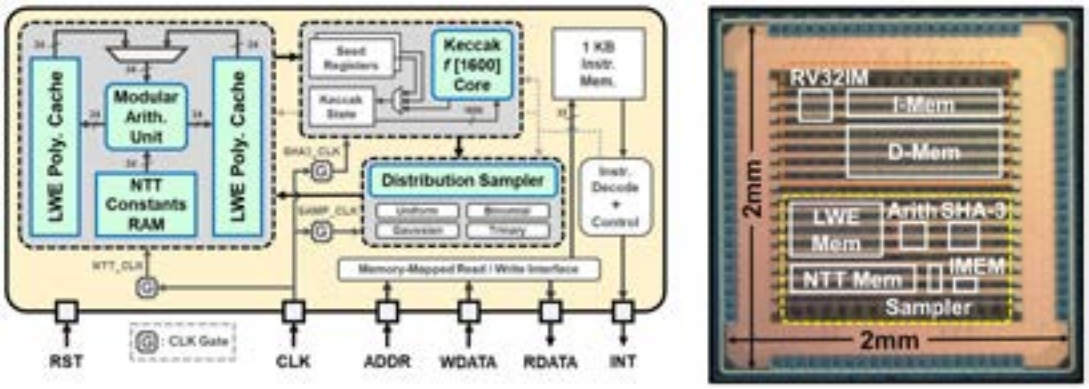
- H. Zhang, K. Martynov, and D. J. Perreault, "A CMOS-Based Energy Harvesting Approach for Laterally-Arrayed Multi-Bandgap Concentrated Photovoltaic Systems," in *IEEE Transactions on Power Electronics*.
- H. Zhang, K. Martynov, D. Li, R. Wen, J. Michel, and D. J. Perreault, "A Power Management Approach for Laterally-Arrayed Multi-Bandgap Concentrated Photovoltaic Systems," *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*, Chicago, IL, USA, 2019, pp. 0702-0707.
- D. Li, et al., "Spectrum Splitting Micro-Concentrator Assembly for Laterally-arrayed Multi-Junction Photovoltaic Module," *2018 Conference on Lasers and Electro-Optics (CLEO)*, San Jose, CA, 2018, pp. 1-2.

An Energy-efficient Configurable Accelerator for Post-quantum Lattice-based Cryptography

U. Banerjee, T. S. Ukyab, A. P. Chandrakasan
Sponsorship: Texas Instruments

Public key cryptography protocols, such as RSA and elliptic curve cryptography, will be rendered insecure by Shor's algorithm when large-scale quantum computers are built. Cryptographers are working on quantum-resistant algorithms, and lattice-based cryptography has emerged as a prime candidate. However, the high computational complexity of these algorithms makes it challenging to implement lattice-based protocols on low-power embedded devices. To address this challenge, we present an energy-efficient lattice cryptography processor with configurable parameters. Efficient sampling, with a SHA-3-based PRNG, provides two orders-of-magnitude energy savings; a single-port RAM-based number theoretic transform memory architecture is proposed, which provides 124k-gate area savings, while a low-power modular arithmetic unit accelerates polynomial computations. This is the first ASIC implementation to demonstrate multiple lattice-based protocols proposed for post-quantum standardization by NIST.

Figure 1 shows the architecture of our lattice cryptography processor along with the chip micrograph. Our test chip was fabricated in TSMC 40-nm low-power CMOS process and supports voltage scaling from 1.1V down to 0.68V. The cryptographic core occupies 0.28 mm2 area consisting of 106k logic gates and 40.25 KB SRAM. It can be programmed with custom instructions for polynomial arithmetic and sampling and it coupled with a low-power RISC-V micro-processor to demonstrate NIST Round 2 lattice-based key encapsulation and digital signature protocols Frodo, NewHope, qTESLA, CRYSTALS-Kyber and CRYSTALS-Dilithium, achieving up to an order-of-magnitude improvement in performance and energy-efficiency compared to state-of-the-art hardware implementations. All key building blocks are constant-time and secure against timing and simple power analysis side-channel attacks. The cryptographic core can also be programmed to implement masking-based differential power analysis side-channel countermeasures, with additional computation cost, with no change to the hardware.



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

FURTHER READING

- U. Banerjee, T. S. Ukyab, and A. P. Chandrakasan, "Sapphire: A Configurable Crypto-Processor for Post-Quantum Lattice-based Protocols," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, vol. 2019, no. 4, pp. 17-61, Aug. 2019.
- U. Banerjee, A. Pathak, and A. P. Chandrakasan, "An Energy-efficient Configurable Lattice Cryptography Processor for the Quantum-secure Internet of Things," *IEEE International Solid-State Circuits Conference*, pp. 46-48, 2019.
- U. Banerjee, A. Wright, C. Juvekar, M. Waller, Arvind, and A. P. Chandrakasan, "An Energy-efficient Reconfigurable DTLS Cryptographic Engine for Securing Internet-of-Things Applications," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 8, pp. 2339-2352, Aug. 2019.

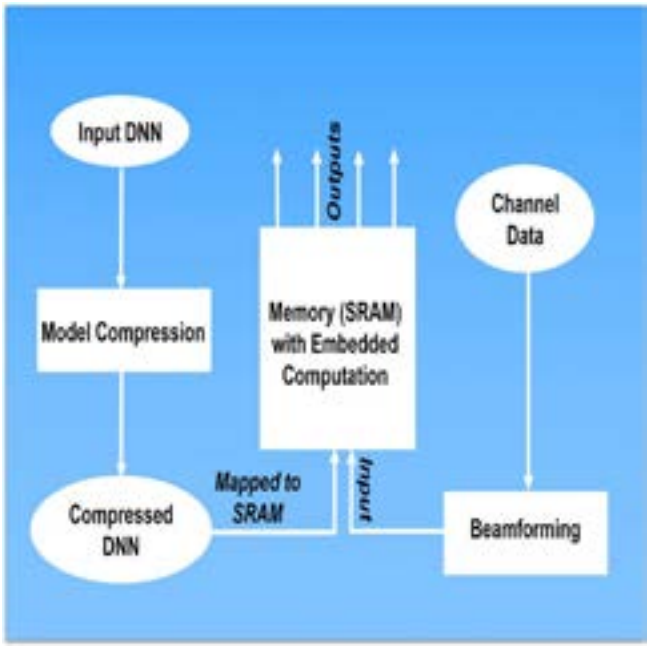
Conformable Ultrasound Patch with Energy-efficient In-memory Computation for Bladder Volume Monitoring

K. Brahma, L. Zhang, V. Kumar, A. P. Chandrakasan, C. Dagdeviren, A. E. Samir, Y. C. Eldar
Sponsorship: Texas Instruments

Continuous monitoring of urinary bladder volume aids management of common conditions such as post-operative urinary retention. Urinary retention is prevented by catheterization, an invasive procedure that greatly increases urinary tract infection. Ultrasound imaging has been used to estimate bladder volume as it is portable, non-ionizing, and low-cost. Despite this, ultrasound technology faces fundamental challenges limiting its usability for next generation wearable technologies. (1) Current ultrasound probes cannot cover curved human body parts or perform whole-organ imaging with high spatiotemporal resolution. (2) Current systems require skilled manual scanning with attendant measurement variability. (3) Current systems are insufficiently energy-efficient to permit ubiquitous wearable device deployment.

We are developing an energy-efficient body contour conformal ultrasound patch capable of real-time bladder volume monitoring. This system will incorporate (1) deep neural network (DNN) based segmentation algorithms to generate spatiotemporally accurate bladder volume estimates and (2) energy-efficient static random-access memory (SRAM) with in-memory dot-product computation for low-power segmentation network implementation. We aim to develop platform technology embodiments deployable across a wide range of health-monitoring wearable device applications requiring accurate, real-time, and autonomous tissue monitoring.

We are training a low-precision (pruned and quantized weights) DNN for accurate bladder segmentation. DNNs are computation-intensive and require large amounts of storage due to high dimensionality data structures with millions of model parameters. This shifts the design emphasis towards data movement between memory and compute blocks. Matrix vector multiplications (MVM) are a dominant kernel in DNNs, and In-Memory computation can use the structural alignment of a 2D SRAM array and the data flow in matrix vector multiplications to reduce energy consumption and increase system throughput.



◀ Figure 1: The flowchart of an energy-efficient system implementing a compressed segmentation network using SRAM designed for in-memory dot product computation.

FURTHER READING

- C. M. W. Daft, "Conformable Transducers for Large-volume, Operator-independent Imaging," *2010 IEEE International Ultrasonics Symposium*, pp. 798-808, 2010.
- R. J. V. Sloun, R. Cohen, and Y. C. Eldar, "Deep Learning in Ultrasound Imaging," arXiv, vol. 1907, p. 02994, 2019.
- A. Biswas and A. P. Chandrakasan, "Conv-RAM: An Energy-efficient SRAM with Embedded Convolution Computation for Low-power CNN-based Machine Learning Applications," *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, pp. 488-490, 2018.

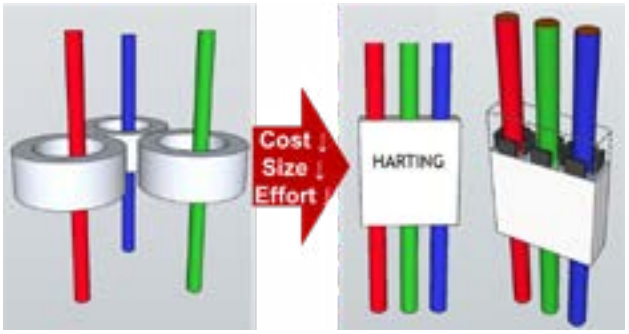
Bandwidth Scalable Current Sensing with Integrated Fluxgate Magnetometers

P. Garcha, V. Schaffer, B. Haroun, S. Ramaswamy, J. Wieser, J. Lang, A. P. Chandrakasan
Sponsorship: Texas Instruments

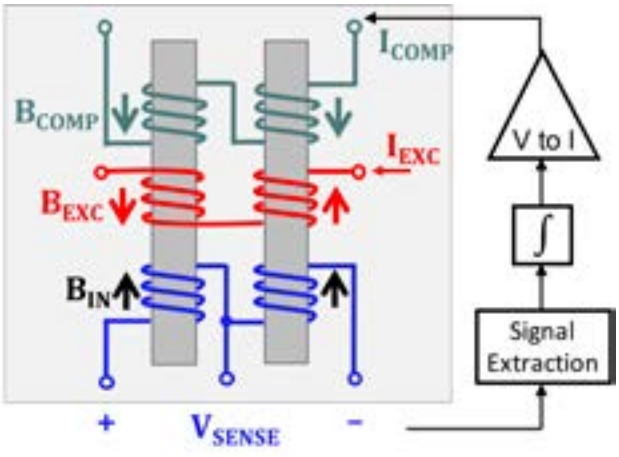
Contactless current sensing finds use in many industrial applications including power line monitoring, motor control, and electric vehicle battery management, as it provides inherent galvanic isolation over direct shunt-sensing. Magnetometers indirectly sense current through a wire by measuring the magnetic fields around it. For stray magnetic field rejection, magnetic sensors need to be placed in the air gap of a magnetic core around each wire. This solution is costly, bulky, and inconvenient to install. [1] proposes a plug-in solution with an array of integrated fluxgate (IFG) magnetometers for contactless current sensing in industrial internet of things applications.

IFG offers a better alternative than Hall sensors in terms of dynamic range (~10^5), sensitivity (200 V/T), linearity (0.1%), and low temperature drift. IFG sensors work by driving magnetic cores in and out of saturation and sensing the resulting voltage difference. They achieve high linearity by balancing external magnetic fields within the core using compensation current, which can be quite power hungry, requiring up to 1W power for a three-phase measurement. Previous IFG sensors are designed for continuous operation at high sampling rates and cannot be duty cycled efficiently due to the long convergence time needed per measurement.

The primary goal of this work is to reduce the energy needs of IFG sensors so they can be used in an array in energy constrained environments. Secondary goals are to increase the bandwidth to >100 kHz for fault detection and increase the measurement range to +/- 60 A at 0.5 cm away from the wire for a compact solution. We achieve these goals through a mixed signal front-end design to enable energy-efficient duty cycling in a bandwidth scalable fluxgate magnetic-to-digital converter. This work achieves higher measurement range, > 100 kHz bandwidth, and considerable energy savings with duty cycling from >100 kHz bandwidths for machine health monitoring to <1 kHz for power quality management.



▲ Figure 1: Smart connectors for contactless current sensing use an array of integrated fluxgate sensors to measure multi-phase currents. The plug-in solution offers lower cost, area, and installation effort over Hall sensing with field concentrators, but the fluxgate magnetometers are power hungry



▲ Figure 2: Fluxgate sensor with two magnetic cores and three sets of coils: excitation, sense, and compensation. When excited, one core saturates before the other, producing voltage VSENSE as a function of (BIN-BCOMP). Compensation provides feedback to improve linearity.

FURTHER READING

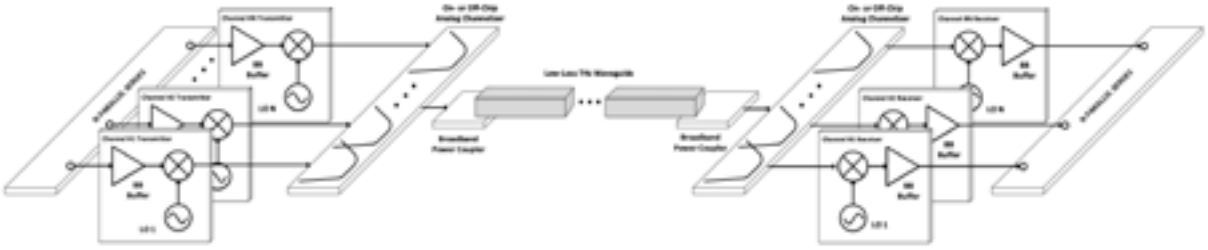
- A. Casallas, "Contactless Voltage and Current Estimation Using Signal Processing and Machine Learning," *M.Eng. thesis, Elec. Eng. & Comp. Sc.*, MIT, Cambridge, 2019.
- M. F. Snoeij, V. Schaffer, S. Udayashankar, and M. V. Ivanov, "Integrated Fluxgate Magnetometer for Use in Isolated Current Sensing," *IEEE J. Solid-State Circuits*, vol. 51, no. 7, pp. 1684–1694, Jul. 2016.

Wideband Sub-THz Components for Ultra-efficient Meter-class Interconnect

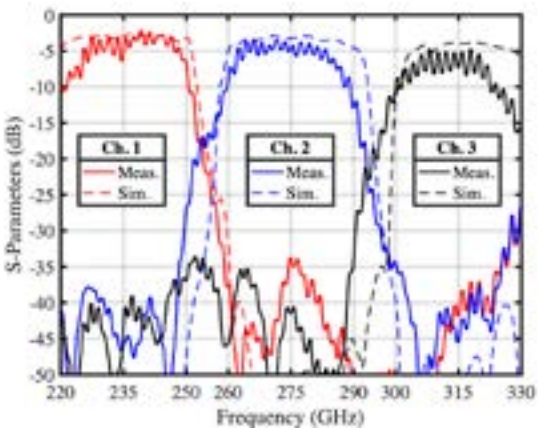
J. Holloway, G. Dogiamis, S. Shin, R. Han
Sponsorship: Intel, Naval Research Laboratory

With the growing interest in millimeter wave and terahertz (THz) electronics, there has been an associated interest in the various components that are required to realize these systems. In one such application, guided and modulated sub-THz (approximately 220-330 GHz) waves are used to transport high-rate data over backplane-scale distances. Such a scheme is attractive for a number of reasons, including broad available fractional bandwidth, compact system size (driven by smaller wavelengths compared to lower-frequency operations), relative robustness to misalignment during assembly versus optical systems, and lower transmission losses than those exhibited by copper lines for high-speed data transmission. One of the challenges associated with the development of the above link system is the realization of compact, low-loss channelizers over the wide operating bandwidth afforded by these types of lines. While waveguide-based channelizers have been demonstrated at lower bands and waveguide components are available at higher operating frequencies, they are relatively large and require more expensive packaging and interface schemes. This type of scheme would require a planar integration approach to be economically feasible.

We have demonstrated the best-in-class channelization performance on a new Intel organic packaging process over 40% fractional bandwidth and occupying up to 200x less area than competing approaches. The design makes use of a very fast circuit-EM co-design technique to overcomes computational hurdles associated with large-scale, full-wave dimensional optimization to rapidly optimize the design. The work utilizes a ridged-SIW resonator design, enabled by the Intel packaging technology, provides superior performance, enables the wide operating band, and reduces the device size by 40%. This design methodology, the selected channelizer topology, and the packaging technology provide a feasible path toward ubiquitous, highly-integrated, and low-cost THz-communication systems-in-package at the board/back-plane level.



▲ Figure 1: High-speed, energy-efficient inter-chip transmission using guided THz wave, requiring wideband multiplexers to enable carrier aggregation



▲ Figure 2: (left) 3D X-ray of the fabricated multiplexer structure; (middle) full-wave simulation of the structure's E-field intensity; (right) measured device response across the 220-330 GHz band.

FURTHER READING

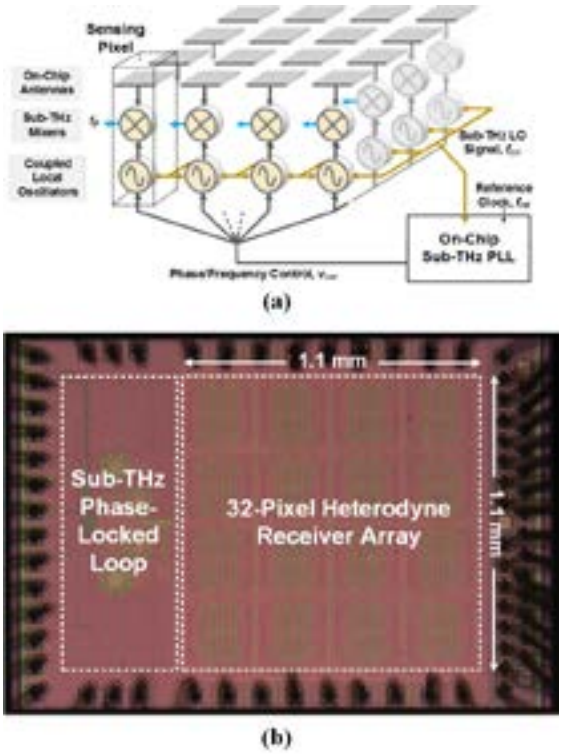
- J. W. Holloway, G. C. Dogiamis, S. Shin, and R. Han, "220-to-330 GHz Manifold Triplexer with Wide Stopband Utilizing Ridged Substrate Integrated Waveguides," *IEEE Transactions on Microwave Theory and Techniques*, to be published, 2020.
- J. W. Holloway, G. C. Dogiamis, and R. Han, "Innovations in Terahertz Interconnects: High-Speed Data Transport Over Fully Electrical Terahertz Waveguide Links," *IEEE Microwave Magazine*, 2020.

A CMOS-based Dense 240-GHz Scalable Heterodyne Receiving Array with Globally-accessible Phase-locked Local Oscillation Signals

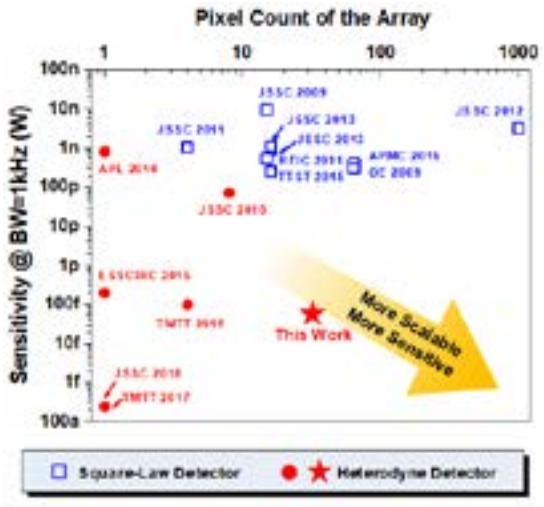
Z. Hu, C. Wang, R. Han
Sponsorship: NSF, MIT-SMART

Driven by the thrust of sensor miniaturization, there is a growing interest in forming steerable beams on the chip scale, which calls for pushing the operation frequency of beam-steering systems towards the terahertz (THz) range. However, this requires disruptive changes to traditional THz receiver architectures, e.g. square-law direct detector arrays (low sensitivity and no phase information preserved) and small heterodyne mixer arrays (bulky and not scalable). The major issue that prevents the latter case from being scalable is the need of large-scale power distribution network for local oscillation signals (LO), which can be very lossy at such high frequency. Here, we report a highly scalable 240-GHz 4×8 heterodyne array achieved by replacing the LO power distributor with a network that couples LOs generated locally at each unit. Now the major challenge for this specific architecture is that each unit should fit into a tight λ/2×λ/2 area to suppress side lobes from beam forming - it makes integrating mixer, local oscillator, and antenna into a unit difficult. Our design addresses this challenge well: the highly compact units ultimately enable the integration of two interleaved 4×4 phase-locked sub-arrays in 1.2-mm2.

The architecture of the entire array is shown in Figure 1(a). Its core component is a self-oscillating harmonic mixer (SOHM), which can simultaneously (1) generate high-power LO signal and (2) down-mix the radio frequency (RF) signal. Since coupling is designed to be global, LOs generated in all units are all locked to an external reference signal by phase-locking two units only. The die (Figure 1(b)) photo shows the placement of the array and the PLL. The measured sensitivity (required incident RF power to achieve SNR=1 at baseband) over 1-kHz detection bandwidth is 58fW, which is more than 4000× improvement over state-of-the-art large-scale square-law detector arrays. Figure 2 shows that this work has pushed the boundary of THz receiver arrays in terms of scale and sensitivity.



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



▲ Figure 2: Comparison with previous works.

FURTHER READING

- Y. Zhang, J. B. Chou, J. Li, H. Li, Q. Du, J. Hu, et al., "Extreme Broadband Transparent Optical Phase Change Materials for High-performance Nonvolatile Photonics," [online], arXiv preprint arXiv:1811.00526, 2018.
- Y. Zhang, J. Liang, M. Shalaginov, S. Deckoff-Jones, C. Ríos, J. Chou, C. Roberts, S. An, et al., "Electrically Reconfigurable Nonvolatile Metasurface using Optical Phase Change Materials," *Conference on Lasers and Electro-Optics, OSA Technical Digest (Optical Society of America, 2019)*, p. JTh5B.3, 2019.

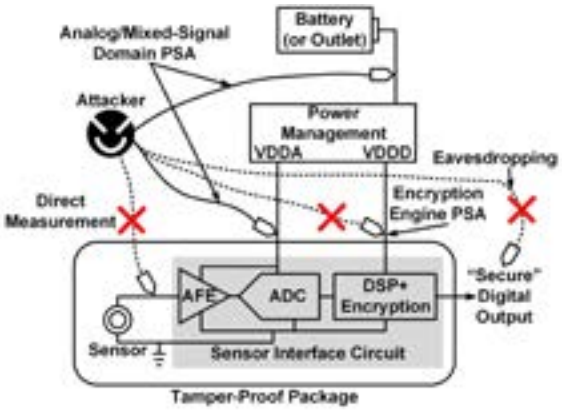
Method and Countermeasure for SAR ADC Power Side-channel Attack

T. Jeong, A. P. Chandrakasan, H.-S. Lee
Sponsorship: Analog Devices Inc., Korea Foundation for Advanced Studies (KFAS), MIT Center for Integrated Circuits and Systems (CICS)

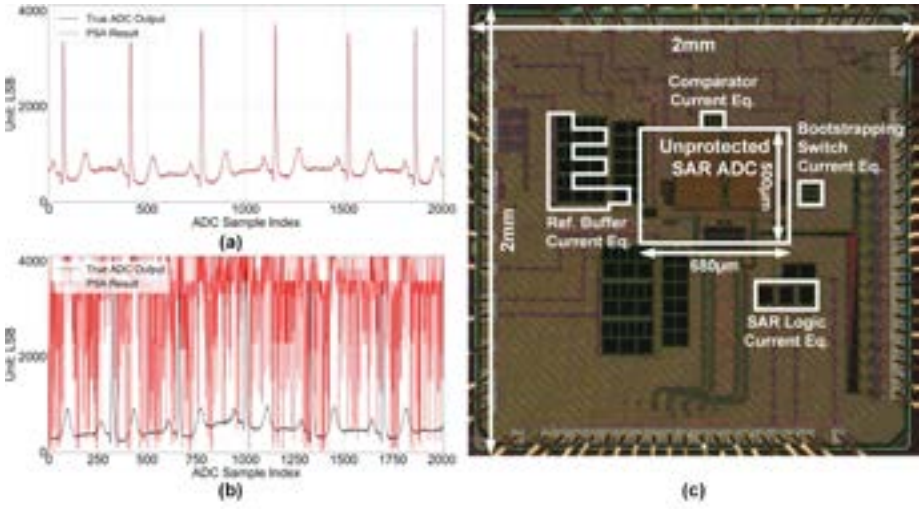
Analog-to-digital converters (ADCs) are essential building blocks of most electronic systems as they convert analog signals into digital bits. Since the demand for digital signal processing keeps growing, researchers have focused on enhancing the ADC performance to keep up with the demand of digital processors. However, recent studies have raised a hardware security concern regarding the ADC-related security loophole, warning that private signal information can be leaked through power supply current waveforms of an ADC.

Figure 1 illustrates an example ADC power side-channel attack (PSA) scenario in sensing hardware that is acquiring a private signal (e.g., healthcare, smart home devices, industrial monitoring). By employing an encryption engine equipped with a PSA-countermeasure, an attacker is prevented from performing eavesdropping and extracting the secret key of the encryption algorithm by tapping into the power supply of the encryption engine. Also, a tamper-proof package can be used to prevent an attacker from directly tapping into the sensor output signal. However, for practical reasons such as a provision for battery replacement and a limitation on physical dimensions, the tamper-proof package may not extend to the ADC power supplies, allowing an attacker to tap into the power supply waveforms of the ADC. Due to the strong correlation between the ADC power supply current waveforms and the ADC digital outputs, an attacker can perform an ADC PSA to obtain the private signal data of the sensing hardware.

This work explores both aspects of ADC PSA: method and countermeasure with an emphasis on SAR ADCs. In this work, neural networks are used as a mapping function that converts a SAR ADC power supply current waveform into the corresponding ADC digital output. To protect a SAR ADC from the proposed PSA method, switched-capacitor circuits called current equalizers are used to decorrelate the on-chip ADC activity and the ADC power supply current waveforms. Figure 2 shows the experimental PSA results on a custom-designed SAR ADC (Figure 2c) that demonstrate the effectiveness of the proposed SAR ADC PSA method and countermeasure schemes.



◀ Figure 1: Example ADC PSA scenario in sensing hardware.



▼ Figure 2: Experimental PSA results on a custom-designed 12-bit, 1.25MS/s SAR ADC (a) PSA-unprotected mode (b) PSA-protected mode (c) Die photo of the custom-designed SAR ADC.

FURTHER READING

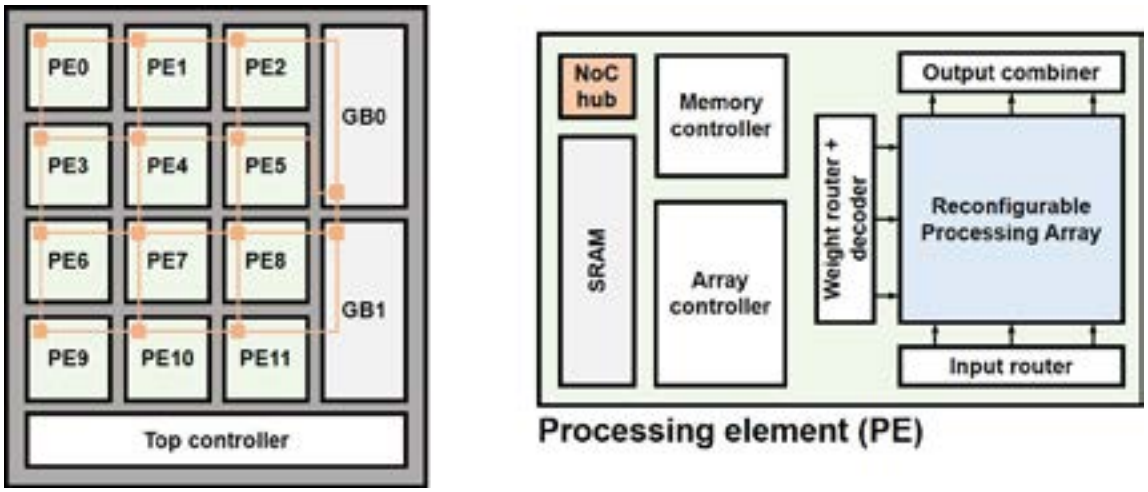
- T. Miki et al., "A Random Interrupt Dithering SAR Technique for Secure ADC Against Reference-Charge Side-Channel Attack," *IEEE Trans. Circuits Syst. II* (Early Access), Feb. 2019.
- T. Jeong, A. P. Chandrakasan, and H.-S. Lee, "S2ADC: A 12-bit, 1.25MS/s Secure SAR ADC with Power Side-Channel Attack Resistance," in Proc. *IEEE Custom Integrated Circuits Conf. (CICC)*, Mar. 2020.

Reconfigurable CNN Processor for Compressed Networks

A. Ji, W. Jung, J. Woo, K. Sethi, S.-L. Luc, A. P. Chandrakasan
Sponsorship: TSMC

Convolutional neural networks (CNNs) have become the standard for performing complex tasks such as image classification due to their high accuracy. However, they typically involve substantial computation (~109 multiplies and adds) to process a single image and require a large amount of storage (~10 to 100 MB) for the fixed weight parameters and intermediate output activations. This makes it challenging to process CNNs locally on edge devices with low power and low latency. To address this, we need custom hardware accelerators to exploit the high parallelism present in the computations. At the same time, they should be flexible enough to support various networks, especially as new and better networks are continuously being developed. Because of the memory constraints on edge devices, we focus on networks compressed by techniques such as Deep Compression and Trained Ternary Quantization, which quantize the weights to a small number of unique values (usually 16 or fewer).

We propose a scalable architecture for efficiently processing compressed networks by reordering the multiplications and additions. Instead of performing each multiply-and-add separately, we accumulate all the activations multiplied by the same weight together and perform the multiplication at the end. With a small number of unique weights, the number of multiplications is greatly reduced, and consequently decreasing the average energy per operation. To enable the tradeoff between accuracy and efficiency, we added reconfigurability for different weight and activation bit widths. This allows us to use shorter bit widths in applications where energy must be minimized and a drop in accuracy can be tolerated. With added support for residual connections and depthwise convolutions, our accelerator can run modern networks such as ResNet and MobileNet, enabling CNN processing for a wide range of applications on energy-constrained devices including cell phones and IoT nodes.



▲Figure 1: System block diagram (left) consisting of processing elements (PE), global bluffers (GB), network-on-chip (NoC), and controller; processing element block diagram (right).

FURTHER READING

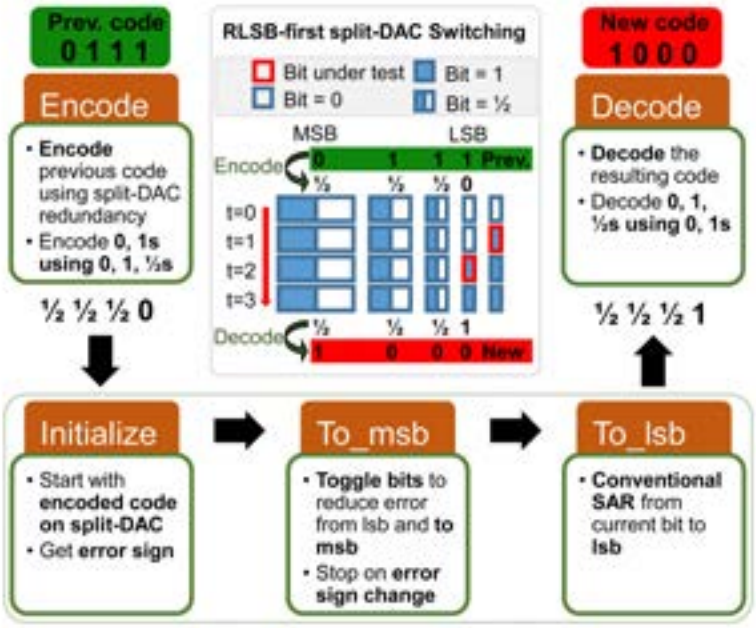
- S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," *ICLR*, 2016.
- Y.-H. Chen, J. Emer, and V. Sze, "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolution Neural Networks," *ISCA*, 2016.
- J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU: An Energy-Efficient Deep Neural Network Accelerator With Fully Variable Weight Bit Precision," *JSSC*, 2019.

Energy-efficient SAR ADC with Background Calibration and Resolution Enhancement

H. S. Khurana, A. P. Chandrakasan, H.-S. Lee
Sponsorship: CICS

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

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



▲ Figure 1: Algorithm for RLSB-first.

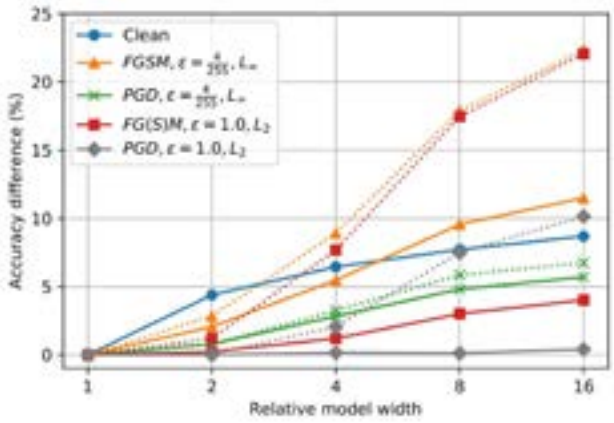
FURTHER READING

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

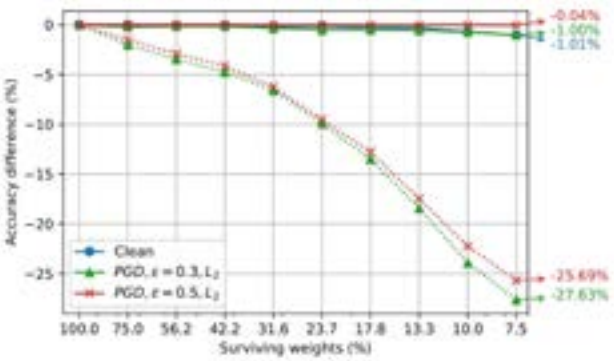
Rethinking Empirical Evaluation of Adversarial Robustness Using First-order Attack Methods

K. Lee, A. P. Chandrakasan
Sponsorship: NXP, Siebel Scholars Foundation, Korea Foundation for Advanced Studies, DARPA

Deep neural networks (DNNs) are known to be vulnerable to adversarial perturbations, which are often imperceptible to humans but can alter predictions of machine learning systems; robustness against those perturbations is becoming an important design factor. A practical approach to measuring adversarial robustness of DNNs is to use the accuracy of those models on examples generated by adversarial attack methods as a proxy for adversarial robustness. However, the failure of those attack methods to find adversarial perturbations cannot be equated with being robust. In our work, we identify three phenomena that inflate accuracy against popular bounded first-order attack methods: 1) a loss function numerically becoming zero when using standard floating point representation, resulting in non-useful gradients; 2) innate non-differentiable functions in DNNs, such as ReLU activation and Max Pooling, incurring "gradient masking"; and 3) certain regularization methods used during training to induce the models to be less amenable to first-order approximation. For each case, we propose compensation methods to improve the evaluation metric for adversarial robustness. The impact of these three sources of overestimated adversarial robustness can be significant when comparing different model capacities for adversarial robustness. For example, Figure 1 shows the adversarial robustness of deep models with the same architecture but different number of neurons per layer. Compensating for these three phenomena can change the relative benefit of using larger models in terms of adversarial accuracy. Similarly, Figure 2 shows adversarial robustness when we iteratively prune weights of an over-parameterized deep model. Adversarial accuracy against the baseline attack method significantly drops as we prune the model; however, actually there is little difference between the original dense model and the sparser models in their adversarial robustness when we properly compensate for these phenomena. Therefore, it is important to rethink the metric we use before we draw conclusions on model capacities or other design factors for their adversarial robustness.



▲Figure 1: Relative change in accuracy against attack methods among models with different numbers of neurons per layer. Dashed and solid lines represent accuracy against baseline attack methods and compensated attack methods, respectively.



▲Figure 2: Accuracy change (both standard accuracy and adversarial accuracy) throughout iterative weight pruning. Dashed and solid lines represent accuracy against baseline attack methods and compensated attack methods, respectively.

FURTHER READING

- A. Athalye, N. Carlini, and D. Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," *Proc. 35th International Conference on Machine Learning, (PMLR)* vol. 80, pp. 274-283, 2018.
- B. Biggio and F. Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," *Pattern Recognition Letters*, vol. 84, pp. 317-331, Dec. 2018.

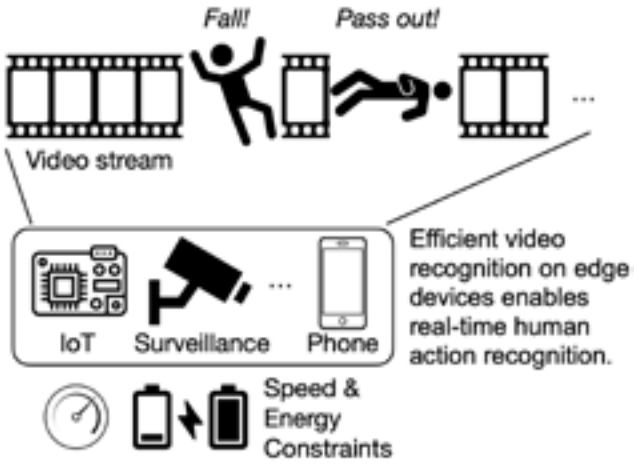
Efficient Video Understanding with Temporal Shift Module

J. Lin, C. Gan, S. Han
Sponsor: MIT-IBM Watson AI Lab, Oak Ridge National Lab

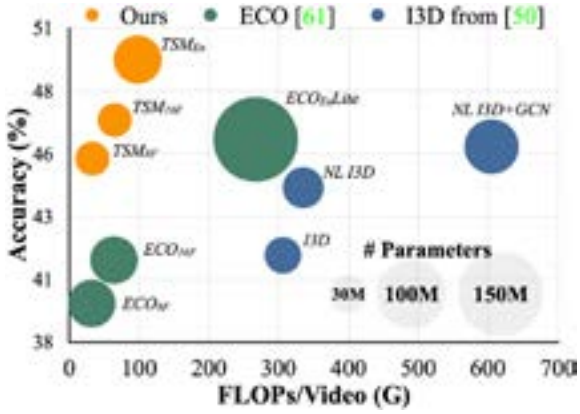
Hardware-efficient video understanding is an important step towards real-world deployment, both in the cloud and on the edge. For example, there are over 10^5 hours of videos uploaded to YouTube every day to be processed for recommendation and ad ranking; similarly, terabytes of sensitive videos in hospitals need to be processed locally on edge devices to protect privacy. All these industry applications require both accurate and efficient video understanding.

Traditionally, a 2D convolutional neural network (CNN) is more efficient but cannot model temporal information; 3D CNN can perform spatial-temporal feature learning, but at the cost of high computation. In this paper, we propose a novel temporal shift module (TSM), which achieves 3D CNN performance at 2D cost. By shifting some of the channels bi-directionally along the temporal dimension, we can facilitate temporal reasoning in 2D CNN at the cost of zero FLOPs and zero parameters. We also propose a uni-directional TSM for online video understanding, supporting online classification and detection from a streaming video.

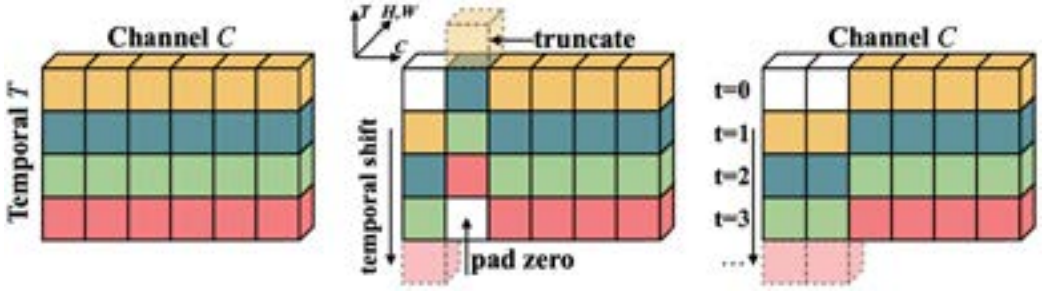
TSM is efficient and accurate: on temporal related datasets, we can improve the performance by double digits at almost no overhead compared to a 2D network. TSM ranks first on Something-Something leaderboard upon submission. TSM is highly scalable: it can be scaled up to 1,536 GPUs and finish the training on Kinetics in 15 minutes; it can also be scaled down to edge deployment, achieving 77 FPS on Jetson Nano and 29 FPS on Galaxy Note 8.



▲ Figure 1: TSM enables efficient online video recognition on edge devices.



▲ Figure 2: TSM achieves high efficiency and high performance.



▲ Figure 3: TSM module design.

FURTHER READING

- J. Lin, C. Gan, and S. Han, "TSM: Temporal Shift Module for Efficient Video Understanding," in Proceedings of the *IEEE International Conference on Computer Vision* (pp. 7083-7093), 2019
- J. Lin, C. Gan, and S. Han, "Training Kinetics in 15 Minutes: Large-scale Distributed Training on Videos," arXiv preprint arXiv:1910.00932, 2019

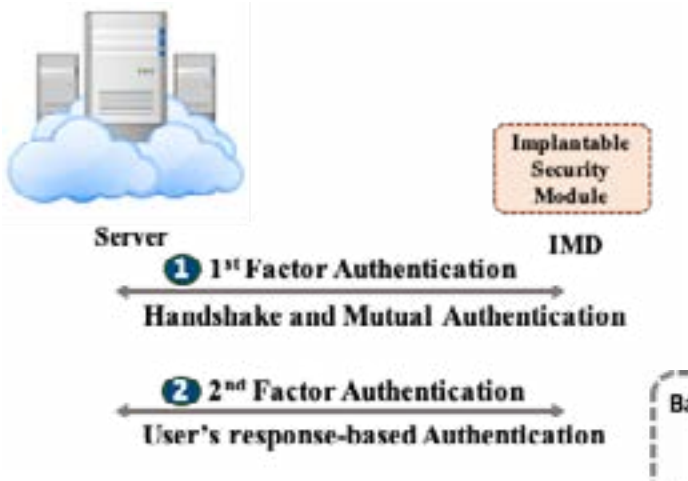
Secure System for Implantable Drug Delivery

S. Maji, U. Banerjee, R. T. Yazicigil, S. H. Fuller, A. P. Chandrakasan
Sponsorship: Analog Devices Inc.

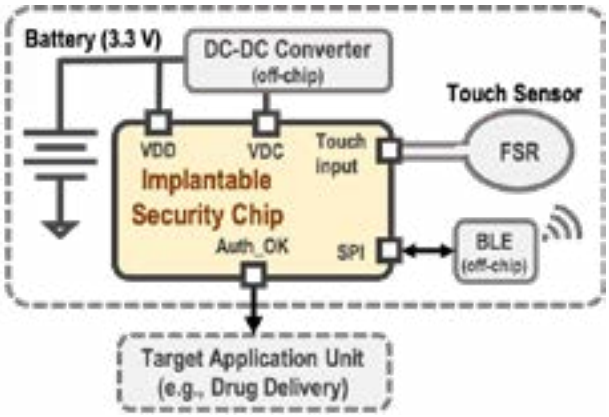
Recent advances in microelectronics and medical technology have enabled Internet-connected IMDs that the patients/users can control through external handheld or wearable devices. However, several proof-of-concept attacks have been demonstrated on such devices by exploiting weaknesses in authentication protocols or their implementations. While such connected implantable devices have the potential to enable many emerging medical applications such as on-command implantable drug delivery, security concerns pose a threat to their widespread deployment. To address this challenge, we present a secure low-power integrated circuit (IC) with sub-nW sleep-state power, energy-efficient cryptographic acceleration, and a novel dual-factor authentication mechanism that ensures that the ultimate security of the IMD lies in the hands of the user.

As a solution, we propose a dual-factor authentication scheme in which cryptographic authentication is supplemented with a voluntary response from the user. The voluntary response serves as a guarded action from the user; that is, it represents consent from the user for executing the desired action without causing them much inconvenience. In our protocol, we have selected a touch-based voluntary response where the user taps on their skin near the IMD. Since most implants are subcutaneous, they can easily detect the tap-pattern and authenticate using this second-factor response. Clearly, for an adversary to provide correct second-factor response to the IMD without alerting the user is difficult, which provides higher security guarantees. In addition to second-factor authentication, the human voluntary factor (human touch) is also used for waking up the system. This provides dual benefits of achieving extremely low-power wake-up and protecting against energy-drainage attacks.

Through circuit-level optimizations, energy-efficient architecture and a novel dual-factor authentication mechanism, this work demonstrates a low-power IC for securing connected biomedical devices of the near future.



◀ Figure 1: A generic diagram of the proposed dual-factor authentication-based protocol for enhanced security of IMD.



▶ Figure 2: Components of the proposed secure, implantable drug-delivery system.

FURTHER READING

- S. Maji, et al., "A Low-Power Dual-Factor Authentication Unit for Secure Implantable Devices," *IEEE CICC*, March 2020.
- S. Maji, "Energy-Efficient Protocol and Hardware for Security of Implantable Devices," *Master's thesis*, Massachusetts Institute of Technology, Cambridge, 2019.
- U. Banerjee, et al., "An Energy-Efficient Reconfigurable DTLS Cryptographic Engine for Securing Internet-of-Things Applications," *IEEE J. Solid-State Circuits*, vol. 54, no. 8, pp. 2339-2352, Aug. 2019.

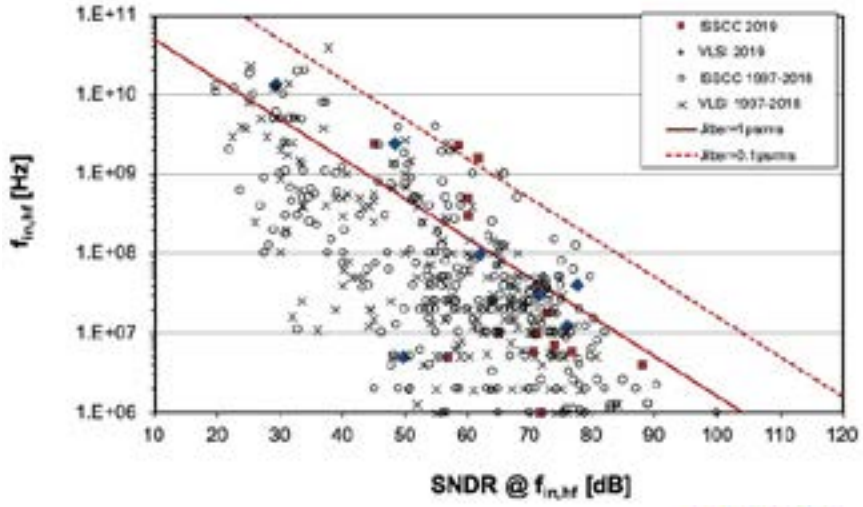
A Sampling Jitter Tolerant Continuous-time Pipelined ADC in 16-nm FinFET

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

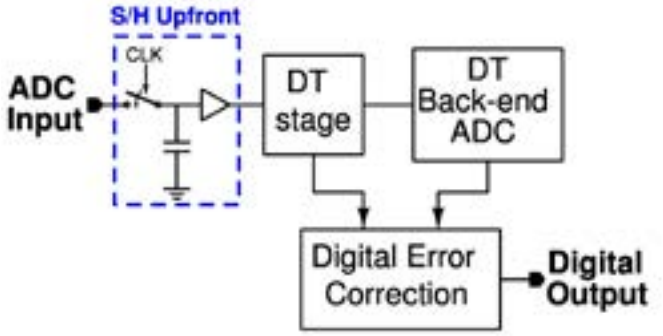
Almost all real-world signals are analog. Yet most data is stored and processed digitally due to advances in the integrated circuit technology. Therefore, analog-to-digital converters (ADCs) are an essential part of any electronic system. The advances in modern communication systems including 5G mobile networks and baseband processors require the ADCs to have 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 sampling clock jitter limits the speed-resolution product (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 Nyuistrate ADCs, the input is sampled upfront (Figure 2). 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 known rms sampling jitter σt the maximum achievable SNR is limited to $1/(2\pi f_{in}\sigma t,)$ where fin is the input signal frequency. In an SoC 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.

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



◀ Figure 1: Performance survey for published ADCs (ISSCC 1997-2019 and VLSI 1997-2019).



▶ Figure 2: A conventional discrete-time pipelined ADC with a sample-and-hold upfront.

FURTHER READING

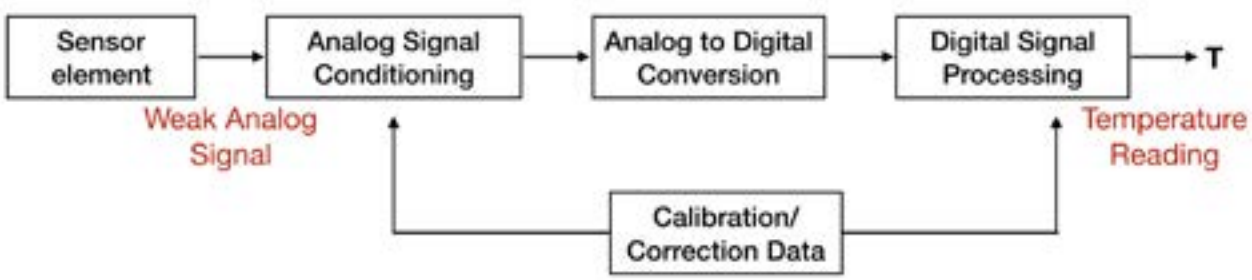
- B. Murmann, "ADC Performance Survey 1997-2019," [Online]. Available: http://web.stanford.edu/~murmann/adcsurvey.html.
- R. van Veldhoven, "A Tri-mode Continuous-time/spl Sigma//spl Delta/modulator with Switched-capacitor Feedback DAC for a GSM-EDGE/CDMA2000/UMTS Receiver," in *Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC. 2003 IEEE International*, pp. 60-477, IEEE, 2003.

Bandgap-less Temperature Sensors for High Untrimmed Accuracy

V. Mittal, A. P. Chandrakasan, H.-S. Lee
Sponsorship: Analog Devices Inc.

Temperature sensors are extensively used in measurement, instrumentation, and control systems. A sensor that integrates the sensing element, analog-to-digital converter, and other interface electronics on the same chip is referred to as a smart sensor. CMOS-based smart temperature sensors offer the benefits of low cost and direct digital outputs over conventional sensors. However, they are limited in their absolute accuracy due to the non-ideal behavior of the devices used to design them. Therefore, these sensors require either calibration or gain/offset adjustments in the analog domain to achieve desired accuracies (Figure 1). The latter process, also called trimming, needs additional expensive test equipment and valuable production time and is a major contributor to the cost of the sensors. To enable high volume production of CMOS-based temperature sensors at low cost, it is imperative to achieve high accuracies without trimming.

This work proposes the design of a CMOS temperature sensor that uses fundamental physical quantities resilient to process variations, package stress, and manufacturing tolerances, in order to achieve high accuracies without trimming. Simulation results prove that 3σ inaccuracy of less than 1° C can be obtained with the proposed method.



▲ Figure 1: System level diagram of a Smart Temperature Sensor

FURTHER READING

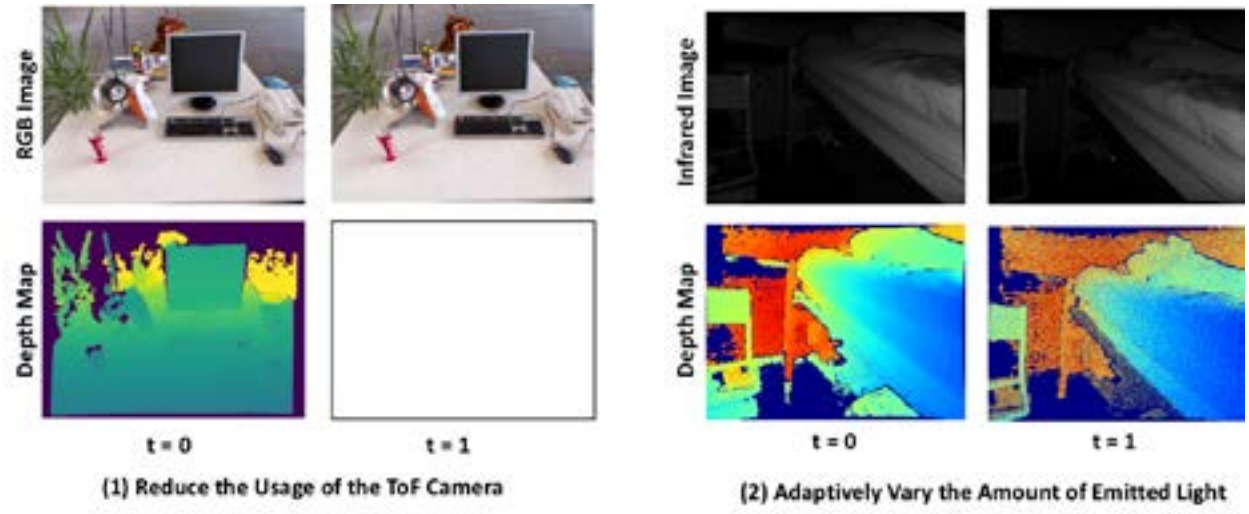
- G. Meijer, M. Pertijs, and K. Makinwa, "Smart Sensor Systems: Emerging Technologies and Applications," *John Wiley & Sons*, 2014.
- Y. Li, H. Lakdawala, A. Raychowdhury, G. Taylor, and K. Soumyanath, "A 1.05V 1.6mW 0.450C 3σ-resolution ΔΣ-based Temperature Sensor with Parasitic-resistance Compensation in 32nm CMOS," in *Solid-State Circuits Conference, 2009. Digest of Technical Papers. 2009 IEEE International*, pp. 340-341, Feb., 2009.

Low Power Time-of-flight Imaging for Dynamic Scenes

J. Noraky, V. Sze
Sponsorship: Analog Devices, Inc.

Depth sensing is useful for many emerging applications, which include mobile augmented reality and robotics. Time-of-flight (ToF) cameras are appealing depth sensors that obtain dense depth measurements, or depth maps, with minimal latency. However, because these sensors obtain depth by emitting light, they can be power-hungry and limit the battery-life of mobile devices. To address this limitation, we present two approaches, shown in Figure 1, that reduce the power for depth sensing by leveraging the other available data: (1) when RGB images are concurrently collected, our technique reduces the usage of the ToF camera and estimates new depth maps using a previous depth map and the consecutive images; (2) when only the data from a ToF camera is available, we adaptively vary the amount of light that the ToF camera emits to infrequently obtain high-power depth maps and to use them to denoise subsequent low power ones. In the second scenario, the ToF camera is always on, but we reduce the overall amount of emitted light while still obtaining accurate depth maps.

In contrast to our previous approaches that dealt with rigid environments, our techniques here can be used for applications that operate in dynamic environments, where the ToF camera and objects in the scenes can have independent, rigid, and non-rigid motions. For dynamic scenes, we show two benefits: (1) when RGB images are concurrently collected, our algorithm can reduce the usage of the ToF camera by over 90%, while still estimating new depth maps with a mean relative error (MRE) of 2.5% when compared to depth maps obtained using a ToF camera; and (2) when only the data from a ToF camera is available, our algorithm can reduce the overall amount of emitted light by up to 81% and the MRE of the low power depth maps by up to 64%. For these techniques, our algorithms use sparse operations and linear least squares to efficiently estimate or denoise depth maps at up to real-time (e.g., 30 fps) using the CPUs of a standard laptop computer and an embedded processor. Our work taken together enables energy-efficient, low latency, and accurate depth sensing for a variety of emerging applications.



▲ Figure 1: We depict the different approaches we take to reduce the power of ToF imaging: (1) we use the previous depth map (t=0) along with the consecutive and concurrently collected RGB images to estimate a new depth map (t=1) without using the ToF camera, and (2) we use the consecutive infrared images (that a ToF camera collects in the process of estimating depth) to combine the data from the high power depth map (t=0) with that from the low power depth map (t=1) to increase its accuracy.

FURTHER READING

- J. Noraky, C. Mathy, A. Cheng, and V. Sze, "Low Power Adaptive Time-of-Flight Imaging for Multiple Rigid Objects," *IEEE International Conference on Image Processing*, 2019.
- J. Noraky and V. Sze, "Low Power Depth Estimation of Rigid Objects for Time-of-Flight Imaging," to appear in *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2020.
- J. Noraky and V. Sze, "Depth Map Estimation of Dynamic Scenes Using Prior Depth Information," arXiv, 2020. Available: https://arxiv.org/abs/2002.00297

CMOS Molecular Clock Using High-order Rotational Transition Probing and Slot-array Couplers

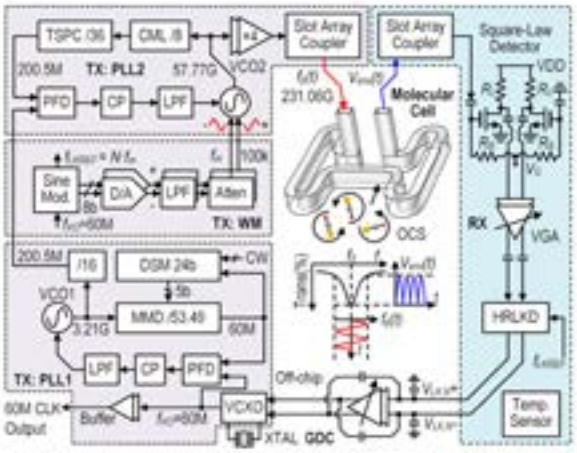
C. Wang, X. Yi, M. Kim, R. Han
Sponsorship: NSF, MIT Lincoln Laboratory, Texas Instruments, Kwanjeong Scholarship

Recently, chip-scale molecular clock (CSMC) referenced to sub-THz transitions of carbonyl sulfide (OCS) gas has emerged as a low-cost solution to achieve high stability with a small size. However, the long-term stability of the first CSMC is limited by the non-flat transmission baseline, which is susceptible to environmental disturbance.

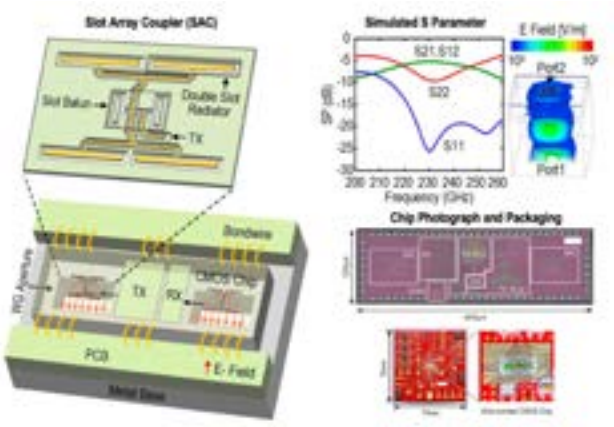
In order to enhance the long-term stability, we presented a CSMC chip that enables high-order dispersion curve locking. Since Nth-order dispersion curve can be comprehended as Nth-order derivative of the OCS line profile, the baseline tilting becomes negligible with high-order dispersion curve. Also, our chip adopts a pair of slot array couplers (SAC) for low loss chip-to-waveguide connection.

Figure 1 shows the clock architecture which consists of a spectroscopic transmitter (TX), referenced to a 60 MHz voltage-controlled crystal oscillator (VCXO), and a spectroscopic receiver (RX). In order to generate the TX probing signal which is wave-length-modulated at a rate of fm=100kHz, high-accuracy, differential sine signal at fm is generated by a pair of 8bit DACs and then fed to varactors in the 57.77 GHz VCO in TX PLL2. The harmonic-rejection lock-in detector (HRLKD) is referenced to fLKREF=3fm, since the 3rd-order dispersion curve is used in this work. Figure 2 shows the structure and simulated S parameter of the SAC. The simulated loss and 3dB fractional bandwidth of the SAC are 5.2dB and 22%, respectively.

The chip was fabricated in a 65nm bulk CMOS process and its DC power consumption was 70mW. The measured Allan Deviation are 3.2×10^{-10}@τ=1s and 4.3×10^{-11}@$\tau=10^3$s, respectively, and the measured magnetic sensitivity of the unshielded clock is $\pm2.9\times10^{-12}$/Gauss. With an on-chip temperature sensor and a 2nd-order polynomial compensation, the frequency drift over temperature range of 27~65°C is $\pm3.0\times10^{-9}$. This work based on very compact size and low cost demonstrates stability performance that is comparable with chip-scale atomic clocks. Its applications include 5G cellular basestations, portable navigation systems, communication and sensing under GPS-denied conditions.



▲ Figure 1: Two PLLs and a quadrupler form the TX, and a square-law detector, a variable-gain amplifier (VGA) and a harmonic-rejection lock-in detector (HRLKD) form RX. TX signal is coupled to a waveguide to probe the 231.06GHz transition line.



▲ Figure 2: The 2×2 double-slot radiators of the SAC collimate the beam and radiate downward into the waveguide aperture through silicon substrate of the chip. This design does not require any external component, and allows for robust clock operation against environmental variations.

FURTHER READING

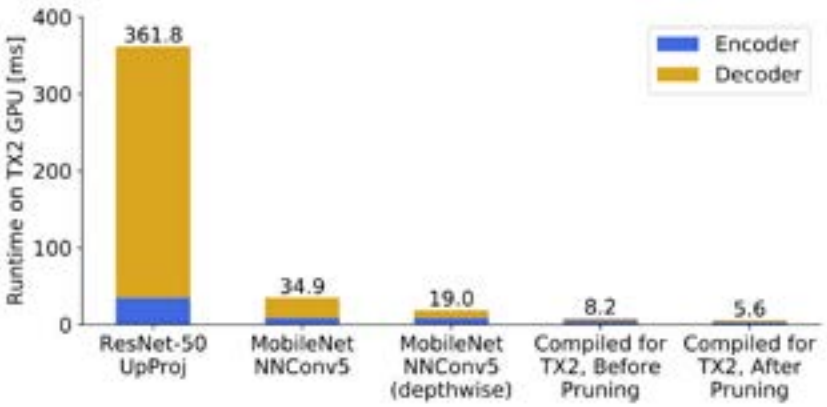
- C. Wang, et al., "An On-chip Fully-electronic Molecular Clock Based on Sub-terahertz Rotational Spectroscopy," *Nature Electronics*, Vol. 1, No. 7, pp. 1-7, Jul. 2018.
- C. Wang, et al., "Chip-scale Molecular Clock," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 914–926, 2018.
- C. Wang, et al., "29.5 Sub-THz CMOS Molecular Clock with 43ppt Long-Term Stability Using High-Order Rotational Transition Probing and Slot-Array Couplers," *ISSCC Dig. Tech. Papers*, pp. 448-449, Feb. 2020.

FastDepth: Fast Monocular Depth Estimation on Embedded Systems

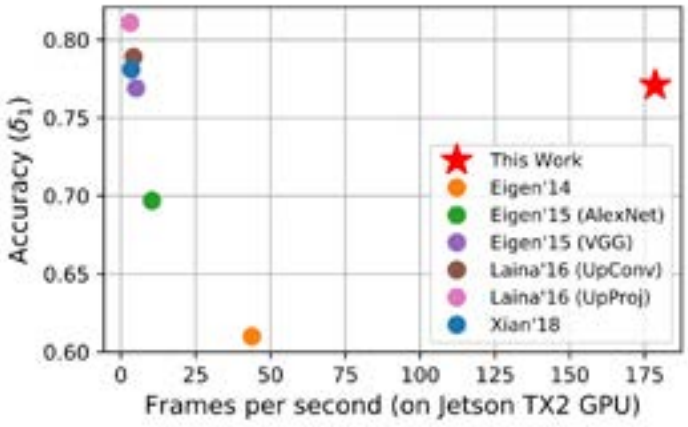
D. Wofk, F. Ma, T.-J. Yang, S. Karaman, V. Sze
Sponsorship: Analog Devices, Intel

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

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



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



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

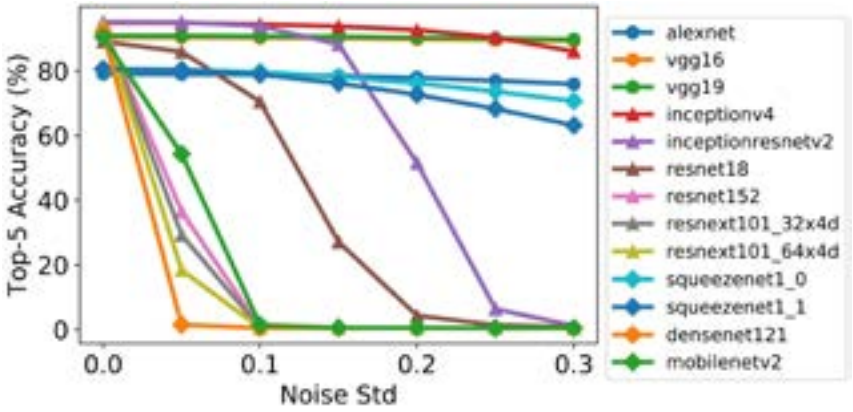
FURTHER READING

- D. Wofk, F. Ma, T.-J. Yang, S. Karaman, V. Sze, "FastDepth: Fast Monocular Depth Estimation on Embedded Systems," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
- T.-J. Yang, A. Howard, B. Chen, X. Zhang, A. Go, M. Sadler, V. Sze, and H. Adam, "NetAdapt: PlatformAware Neural Network Adaptation for Mobile Applications," in *European Conference on Computer Vision (ECCV)*, 2018.

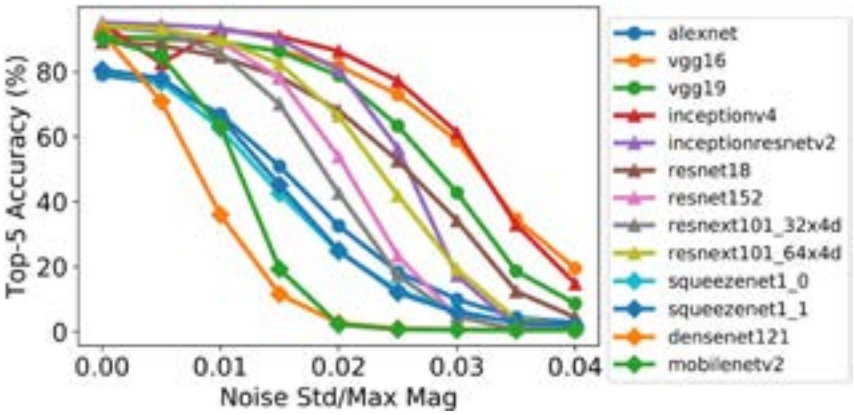
Design Considerations for Efficient Deep Neural Networks in Processing-in-memory Accelerators

T.-J. Yang, V. Sze
Sponsorship: MIT, MIT-IBM Watson AI Lab, MIT Quest for Intelligence, NSF E2CDA 1639921

Deep neural networks (DNNs) deliver state-of-the-art accuracy on a wide range of artificial intelligence tasks at the cost of high computational complexity. Since data movement tends to dominate energy consumption and can limit throughput for memory-bound workloads, processing in memory (PIM) has emerged as a promising way for processing DNNs. Unfortunately, the design of efficient DNNs specifically for PIM accelerators has not been widely explored. In this work, we highlight the key differences between PIM and digital accelerators and summarize how these differences need to be accounted for when designing DNNs for PIM accelerators. The key design considerations include (1) resilience to circuit and device non-idealities, which affect accuracy; (2) data movement of feature map activations, which affects energy consumption and latency; and (3) utilization of the memory array, which affects energy consumption and latency. We examine the use of PIM accelerators on 18 DNNs published since 2012 for image classification on the ImageNet dataset to highlight the importance of the various design considerations. Our experiment results show that the common principles used to design efficient DNNs for digital accelerators (e.g., making a DNN deeper with smaller layers) may not suit PIM accelerators. Therefore, we need to rethink how to design efficient DNNs for PIM accelerators.



◀ Figure 1: The impact of fixed noise in output activations on the accuracy of representative DNNs. The rank order of accuracy may change with the standard deviation of the noise.



▶ Figure 2: The impact of rescaled noise in output activations on the accuracy of representative DNNs. The rank order of accuracy may change with the ratio of the standard deviation of the noise to the maximum magnitude of the activations.

FURTHER READING

- T.-J. Yang and V. Sze,"Design Considerations for Efficient Deep Neural Networks on Processing-in-Memory Accelerators,"presented at *IEEE International Electron Devices Meeting (IEDM)*, 2019.
- T. Gokmen and Y. Vlasov, "Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations," *Frontiers in Neuroscience*, vol. 10, p. 333, 2016.

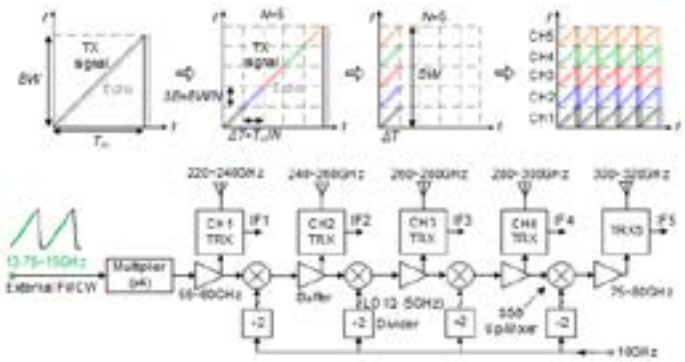
A Terahertz FMCW Comb Radar in 65-nm CMOS with 100GHz Bandwidth

X. Yi, C. Wang, M. Lu, J. Wang, J. Grajal, R. Han
Sponsorship: NSF, TSMC

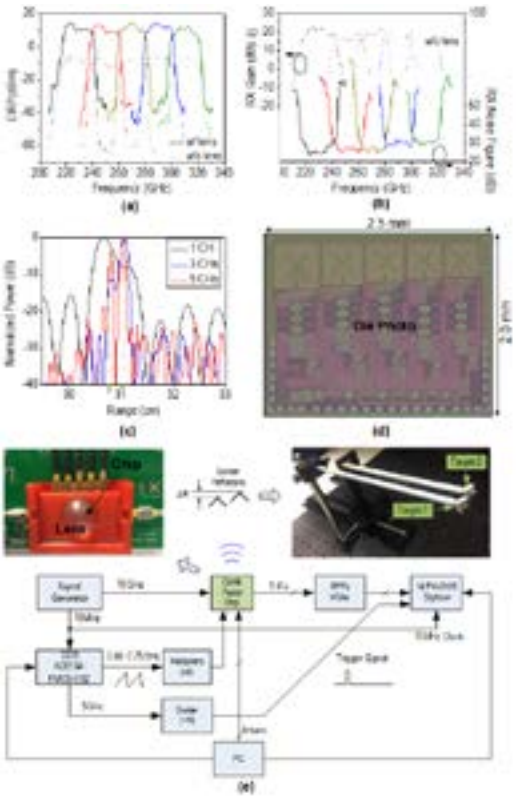
The increasing demands for low-cost, compact, and high-resolution radar systems have driven the operation frequency to terahertz due to the shorter wavelength and larger bandwidth. However, conventional single-transceiver frequency-modulated continuous-wave (FMCW) radar chips provide only limited signal bandwidth, especially when implemented using Complementary metal–oxide–semiconductor (CMOS) technologies with low fT and fmax. Therefore, prior THz integrated radars are based on compound semiconductors and have severely degraded performance near the band edges. That not only limits their applications in high-accuracy scenarios but also creates tradeoffs between bandwidth and detection range.

To avoid such limitations, we adopt a frequency-comb-based scalable architecture using a paralleled transceiver array as shown in Figure 1. The concept of the FMCW comb radar is illustrated as a wideband chirp signal is divided into N identical segments that sweep simultaneously using an array of transceivers with equally-spaced carrier frequencies. Each transceiver has its own on-chip antenna, and the received echo signal is mixed with the transmitted signal to generate an IF output. The presented high-parallelism scheme offers several advantages over single-transceiver radars. Firstly, it achieves scalable bandwidth extension and enables implementations in less advanced technologies as well as flatter frequency responses across the entire operation band. Secondly, the flat frequency response also leads to higher linearity of the equivalent chirp signal. Thirdly, the SNR of comb radar is improved by N for a given total detection time.

Implemented in a 65-nm bulk CMOS process, a five-transceiver radar chip is prototyped with seamless coverage of the entire 220-to-320GHz band as shown in Figure 2. Across the total chirp bandwidth of 100GHz, 0.6dBm/20dBm (with/without lens) multi-channel-aggregated EIRP with 8.8dB output power fluctuation, and 22.8dB minimum RX noise figure are achieved. With all five channels stitched together, 2.5-mm separation of two objects is clearly detected. This chip has an area of 5mm2 and consumes 840mW of power. This is the first demonstration of THz radar in CMOS process, and a record FMCW bandwidth is achieved.



▲ Figure 1: The basic concept and system diagram of the THz comb radar.



▶ Figure 2: Measured (a) TX EIRP, (b) RX gain & noise figure, and (c) range resolution. (d) Die photo and (e) FMCW measurement setup.

FURTHER READING

- X. Yi, C. Wang, M. Lu, J. Wang, J. Grajal, and R. Han, "A Terahertz FMCW Comb Radar in 65nm CMOS with 100GHz Bandwidth," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2020, pp. 90–91.

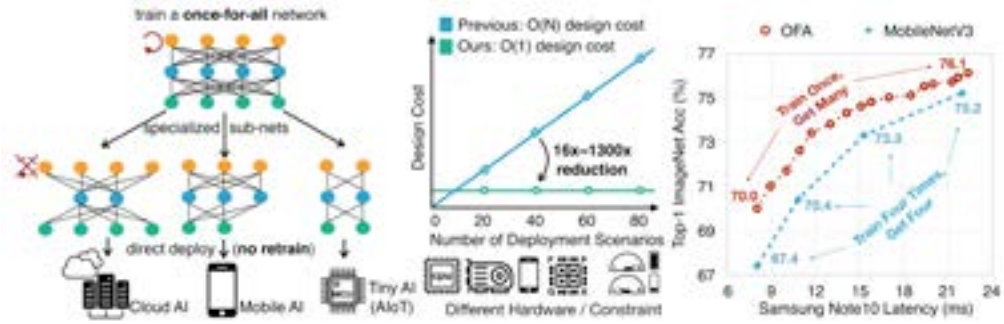
Efficient AutoML with Once-for-all Network

H. Cai, C. Gan, T. Wang, Z. Zhang, S. Han
Sponsorship: NSF, MIT-IBM Watson AI Lab, Google Research Award, AWS Machine Learning Research Award

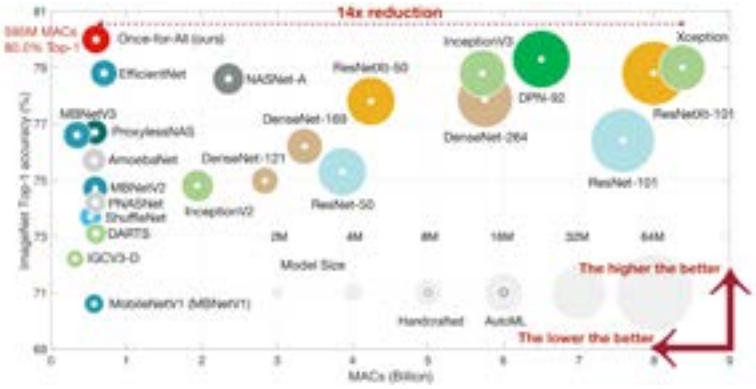
We address the challenging problem of efficient inference across many devices and resource constraints, especially on edge devices. Conventional approaches either manually design or use neural architecture search (NAS) to find a specialized neural network and train it from scratch for each case, which is computationally prohibitive (causing CO2 emission as much as 5 cars' lifetime) and thus unscalable. In this work, we propose to train a once-for-all (OFA) network that supports diverse architectural settings by decoupling training and search, to reduce the cost. We can quickly get a specialized sub-network by selecting from the OFA network without additional training. To efficiently train OFA networks, we also propose a novel progressive shrinking algorithm, a generalized pruning method that reduces the model size across many more dimensions than pruning (depth, width, kernel size, and resolution). It can obtain a surprisingly large number of sub-networks that can fit different hardware platforms and latency constraints while maintaining the same level of accuracy as training independently.

On diverse edge devices, OFA consistently outperforms state-of-the-art (SOTA) NAS methods (up to 4.0% ImageNet top1 accuracy improvement over MobileNetV3, or same accuracy but 1.5x faster than MobileNetV3, and 2.6x faster than EfficientNet w.r.t measured latency) while reducing GPU hours and CO2 emission by many orders of magnitude. In particular, OFA achieves a new SOTA 80.0% ImageNet top1 accuracy under the mobile setting (<600M MACs).

OFA is the winning solution for the 3rd Low Power Computer Vision Challenge (LPCVC, classification DSP track) and the 4th LPCVC (both classification track and detection track).



▲Figure 1: The OFA network can produce diverse specialized sub-networks without retraining. It removes the need for repeated architecture design and model training, saving orders-of-magnitude GPU training cost, and also produces efficient models for fast inference on mobile devices.



◀Figure 2: OFA network achieves high accuracy at low computation cost, being at the top-left corner of the accuracy-computation trade-off curve.

FURTHER READING

- H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-All: Train One Network and Specialize It for Efficient Deployment," *ICLR*, 2020.
- H. Cai, L. Zhu, and S. Han, "Proxyless NAS: Direct Neural Architecture Search on Target Task and Hardware," *ICLR*, 2019.

An Efficient and Continuous Approach to Information-theoretic Exploration

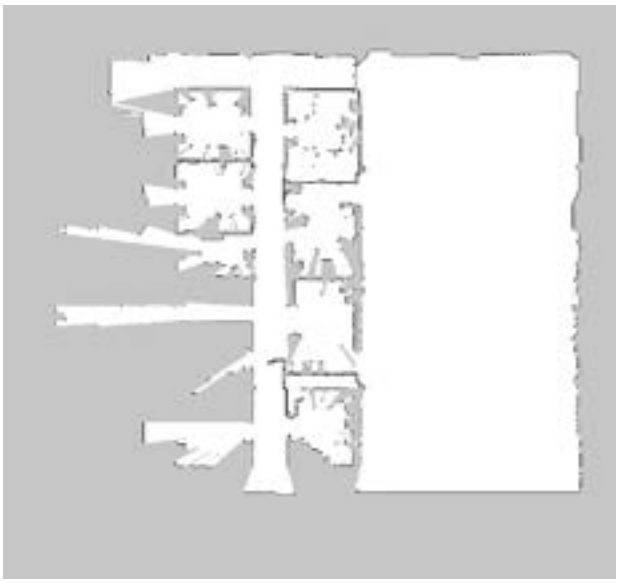
T. Henderson, V. Sze, S. Karaman
Sponsorship: NSF Cyber-Physical Systems (CPS) Program

Exploration of unknown environments is embedded in many robotics applications: search and rescue, crop survey, space exploration, etc. The central problem an exploring robot must answer is "where should I move next?" The answer should balance travel cost with the amount of information expected to be gained about the environment. Traditionally, this question has been answered by a variety of heuristics that provide no guarantees on their exploration efficiency. Information-theoretic methods can produce an optimal solution, but until now they were thought to be computationally intractable.

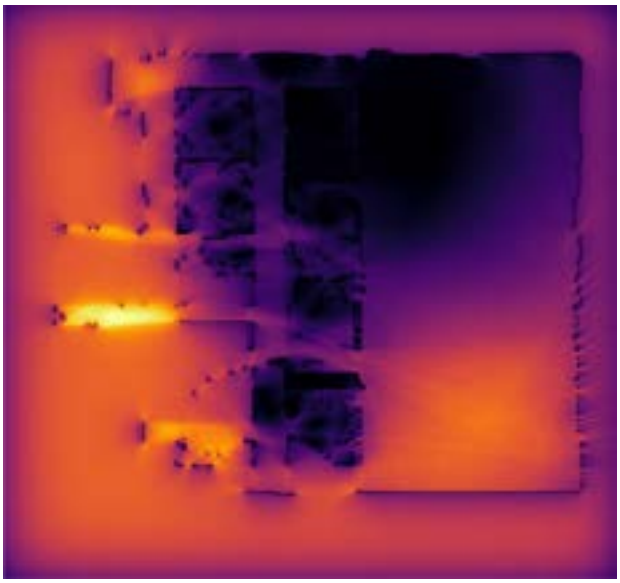
In our recent work we describe the Fast Continuous Mutual Information (FCMI) algorithm, which computes the information-theoretic exploration metric efficiently. FCMI takes as input an incomplete occupancy map like the one shown in Figure 1, where white pixels indicate free space, black pixels indicate occupied space, and gray pixels indicate unknown space. It then returns an information surface as shown in Figure 2, where the brightness of each pixel indicates how much information is expected to be gained by exploring at that location. The algorithm also works on multi-resolution or 3-dimensional maps. FCMI has a lower asymptotic complexity than existing methods and our experiments demonstrate that it is hundreds of times faster than the state-of-the-art for practical inputs.

The key insight that enables FCMI is to consider the occupancy map as a continuous random field rather than a discrete collection of cells. This reveals a nested information structure that makes it possible to recursively reuse computation from one map location in adjacent locations. The continuous structure also provides more general insights that are relevant to any occupancy mapping system.

For practical map sizes, FCMI runs in seconds on a single threaded laptop CPU which is well within the timing constraints for most robotic applications. It provides considerable savings to energy constrained systems by reducing both the exploration travel cost and the computation cost. FCMI is also highly parallelizable and suited for a rapid, low energy, embedded implementation.



▲ Figure 1: An incomplete occupancy grid map of MIT's building 31.



▲ Figure 2: An information surface produced by the FCMI algorithm.

FURTHER READING

- T. Henderson, V. Sze, S. Karaman "An Efficient and Continuous Approach to Information-Theoretic Exploration," Proceedings of the the *2020 Informational Conference on Robotics and Automation (ICRA)*, 2020.

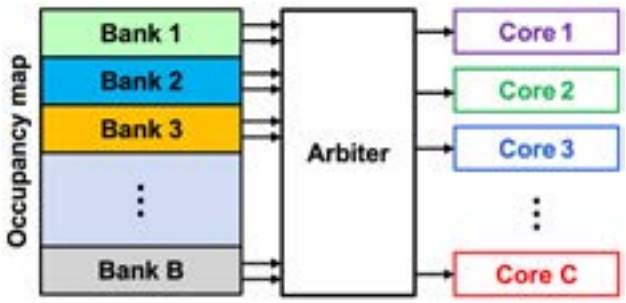
A Mutual Information Accelerator for Autonomous Robot Exploration

P. Z. X. Li, S. Karaman, V. Sze
Sponsorship: AFOSR YIP, NSF

Robotic exploration problems arise in various contexts, ranging from search and rescue missions to underwater and space exploration. In these domains, exploration algorithms that allow the robot to rapidly create the map of the unknown environment can reduce the time and energy for the robot to complete its mission. Shannon mutual information (MI) at a given location is a measure of how much new information of the unknown environment the robot will obtain given what the robot already know from its incomplete understanding of the environment. In a typical exploration pipeline, robot starts with an incomplete map of the environment. At every step, the robot computes the MI across the entire map. Then, the robot can select the location with the highest mutual information for exploration in order to gain the most information about the unknown environment.

However, on the CPUs and GPUs typically found on mobile robotic platforms, computing MI using the state-of-the-art Fast Shannon Mutual Information (FSMI) algorithm across the entire map takes more than one second, which is too slow for enabling fast autonomous exploration. As a result, the emerging literature considers approximation techniques, and many practitioners rely on heuristics that often fail to provide any theoretical guarantees.

To eliminate the bottleneck associated with the computation of MI across the entire map, we propose a novel multicore hardware architecture (Figure 1) with a memory subsystem that efficiently organizes the storage of the occupancy grid map and an arbiter that effectively resolves memory access conflicts among MI cores so that the entire system achieves high throughput. In addition, we provide rigorous analysis of memory subsystem and arbiter in order to justify our design decisions and provide provable performance guarantees. Finally, we thoroughly validated the entire hardware architecture by implementing it using a commercial 65nm ASIC technology (Figure 2).



▲ Figure 1: Proposed multi-core hardware architecture that provides sufficient memory bandwidth so that the computation cores are active.

▲ Figure 2: Layout of the proposed hardware architecture with 16 cores using a commercial 65nm technology.

FURTHER READING

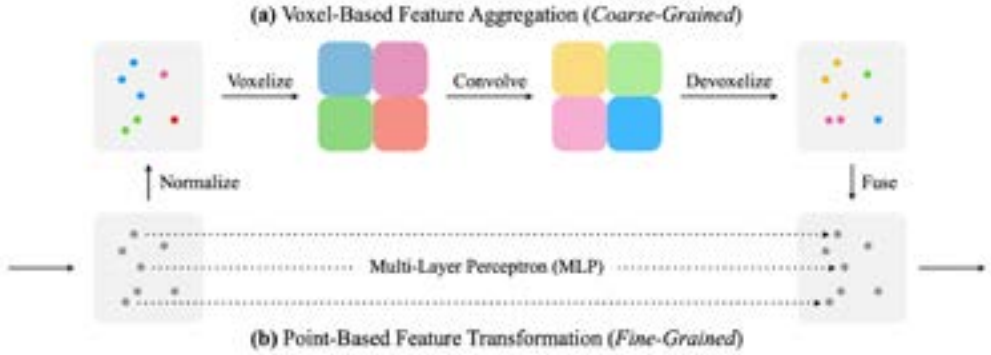
- P. Z. X. Li*, Z. Zhang*, S. Karaman, V. Sze, "High-throughput Computation of Shannon Mutual Information on Chip," *Robotics: Science and Systems (RSS)*, Jun. 2019.
- Z. Zhang, T. Henderson, V. Sze, S. Karaman, "FSMI: Fast Computation of Shannon Mutual Information for Information-theoretic Mapping,"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

Efficient 3D Deep Learning with Point-voxel CNN

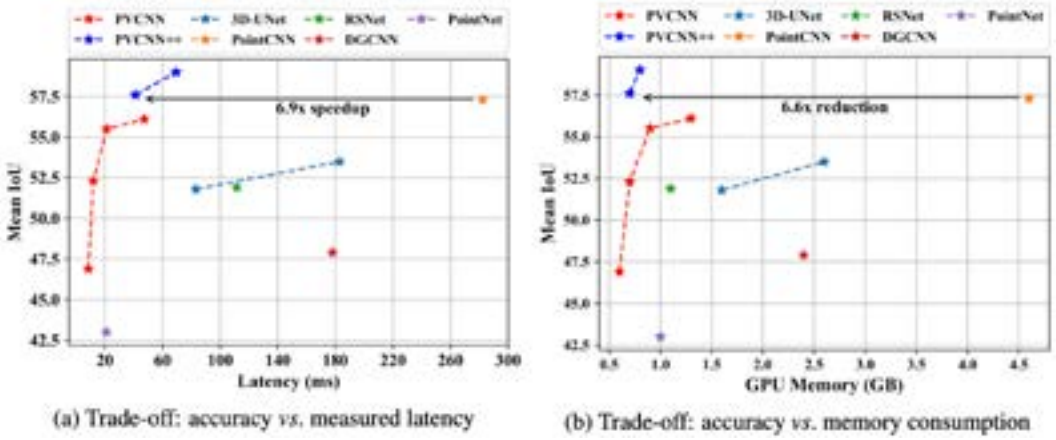
Z. Liu, H. Tang, Y. Lin, S. Han
Sponsorship: MIT Quest for Intelligence, MIT-IBM Watson AI Lab, Samsung, Facebook, SONY

3D deep learning has received increased attention thanks to its wide applications: e.g., AR/VR and autonomous driving. These applications need to interact with people in real time and therefore require low latency. However, edge devices (such as AR/VR headsets and self-driving cars) are tightly constrained by hardware resources and battery. Previous work processes 3D data using either voxel-based or point-based NN models. However, both approaches are computationally inefficient. The computation cost and memory footprints of the voxel-based models grow cubically with the input resolution, making it memory-prohibitive to scale up the resolution. As for point-based networks, up to 80% of the time is wasted on structuring the sparse data which have rather poor memory locality, not on the actual feature extraction.

To this end, we propose Point-Voxel CNN (PVCNN) that represents the 3D input data as point clouds to take advantage of the sparsity to reduce the memory footprint, and leverages the voxel-based convolution to obtain the contiguous memory access pattern (Figure 1). Evaluated on semantic and part segmentation datasets, it achieves a much higher accuracy than the voxel-based baseline with 10× GPU memory reduction; it also outperforms the state-of-the-art point-based models with 7× measured speedup on average (Figure 2). We validate its general effectiveness on 3D object detection: Frustrum PVCNN outperforms Frustrum PointNet++ by up to 2.4% mAP with 1.8× measured speedup and 1.4× GPU memory reduction.



▲ Figure 1: PVCNN is composed of several PVConv's, each of which has a low-resolution voxel-based branch and a high-resolution point-based branch. The voxel-based branch extracts coarse-grained neighborhood information, which is supplemented by the fine-grained individual point features extracted from the point-based branch.



▲ Figure 2: Results of indoor scene segmentation on S3DIS. On average, our PVCNN and PVCNN++ outperform the point-based models with 8× measured speedup and 3× memory reduction, and outperform the voxel-based baseline with 14× measured speedup and 10× memory reduction.

SpArch: Efficient Architecture for Sparse Matrix Multiplication

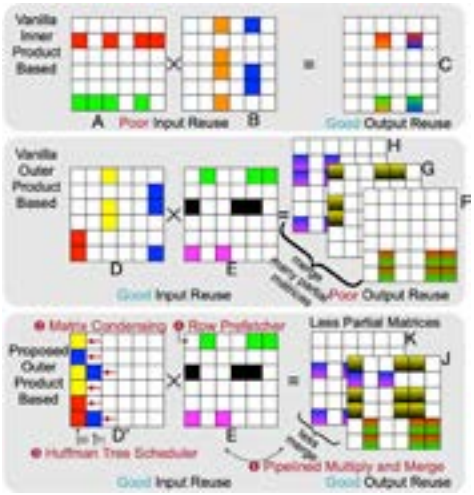
Z. Zhang*, H. Wang*, S. Han, W. J. Dally
(*Equal Contributions)
Sponsorship: NSF, DARPA

Generalized sparse matrix-matrix multiplication (SpGEMM) is the key computing kernel for many algorithms such as compressed deep neural networks. However, the performance of SpGEMM is memory-bounded on the traditional general-purpose computing platforms (CPU, GPU) because of the irregular memory access pattern and poor locality brought by the extremely sparse matrices. For instance, the density of Twitter's adjacency matrix is as low as 0.000214%. Previous accelerator OuterSPACE proposed an outer product method that has perfect input reuse but poor output reuse due to enormous partial matrices, thus achieving only 10.4% of the theoretical peak.

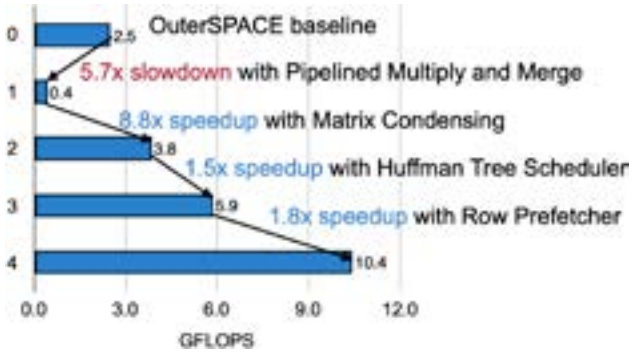
Therefore, we propose SpArch (HPCA'2020) to jointly optimize input and output data reuse. We obtain input reuse by using the outer product and output reuse by on-chip partial matrix merging (Figure 1).

We first design a highly parallelized merger to pipeline the two computing stages, Multiply and Merge. However, the number of partial matrices can easily exceed the on-chip merger's parallelism and incurs even larger DRAM access. We thus propose a condensed matrix representation for the left input matrix, where all non-zero elements are pushed to the left, forming much denser columns and fewer partial matrices. Unfortunately, the condensed representation can still produce more partial matrices than the merger's parallelism. Since the merge order impacts DRAM access, we should merge matrices with fewer non-zeros first. To this end, we design a Huffman tree scheduler to decide the near-optimal merge order of the partial matrices. Finally, we propose a row prefetcher to prefetch rows of the right matrix and store to a row buffer, thus improving the input reuse.

We evaluate SpArch on real-world datasets from SuiteSparse, SNAP, and rMAT, achieving 4×, 19×, 18×, 17×, and 1285× speedup and 6×, 164×, 435×, 307×, and 62× energy saving over OuterSPACE, MKL, cuSPARSE, CUSP and ARM Armadillo, respectively. Figure 2 shows the speedup breakdown of SpArch over OuterSPACE.



▲ Figure 1: Four Innovations in SpArch.



▲ Figure 2: SpArch Speedup Breakdown.

FURTHER READING

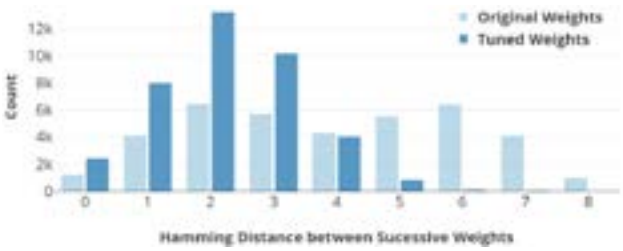
- Z. Zhang*, H. Wang*, S. Han, and W. J. Dally, "SpArch: Efficient Architecture for Sparse Matrix Multiplication," *HPCA 2020*, pp. 261-274, 2020.
- S. Pal, J. Beaumont, D. H. Park, A. Amarnath, S. Feng, C. Chakrabarti, C., et, al, "OuterSPACE: An Outer Product Based Sparse Matrix Multiplication Accelerator," *HPCA 2018*, pp. 724-736, 2018.
- J. Cong, Z. Fang, M. Lo, H. Wang, J. Xu, and S. Zhang, "Understanding Performance Differences of FPGAs and GPUs," *FCCM 2018*, pp. 93-96, 2018.

Flexible Low Power CNN Accelerator for Edge Computing with Weight Tuning

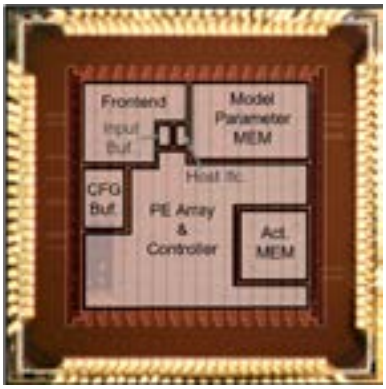
M. Wang, A. P. Chandrakasan
Sponsorship: Foxconn Technology Group

Smart edge devices that support efficient neural network (NN) processing have recently gained public attention. With algorithm development, previous work has proposed small-footprint NNs achieving high performance in various medium complexity tasks, e.g. speech keyword spotting (KWS), human activity recognition (HAR), etc. Among them, convolutional NNs (CNNs) perform well, which gives rise to the deployment of CNNs on edge devices. A hardware platform for edge devices should be (1) flexible to support various NN structures optimized for different applications; (2) energy efficient to operate within the power budget; (3) achieving high accuracy to minimize spurious triggering of power-hungry downstream processing, since it is often part of a large system.

This work proposes a weight tuning algorithm to improve the energy efficiency by lowering the switching activity of weight-related components, e.g. weight buses and multipliers. To achieve that, the algorithm reduces the Hamming distance between successive weights as shown in Figure 1. A flexible and runtime-reconfigurable CNN accelerator is co-designed with the algorithm. The system is fully self-contained for small CNNs. Speech keyword spotting is shown as an example with an integrated feature extraction frontend. As shown in Figure 2, a fully integrated custom ASIC is fabricated for this system. Based on post place-and-route simulation of the ASIC, the weight tuning algorithm reduces the energy consumption of weight delivery and computation by 1.70x and 1.20x respectively with little loss in accuracy.



▲ Figure 1: The histogram of the distribution of Hamming distance between successive weights.



▲ Figure 2: Chip micrograph.

FURTHER READING

- M. Wang, and A. P. Chandrakasan. "Flexible Low Power CNN Accelerator for Edge Computing with Weight Tuning." *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 209-212, Nov., 2019.

Modern Microprocessor Built from Complementary Carbon Nanotube Transistors

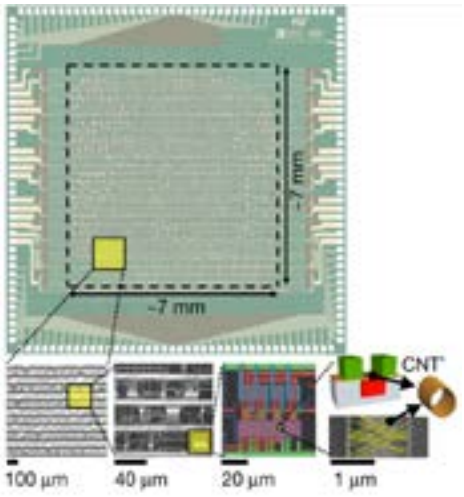
G. Hills, C. Lau, A. Wright, S. Fuller, M. D. Bishop, T. Srimani, P. Kanhaiya, R. Ho, A. Amer, Y. Stein, D. Murphy, Arvind, A. P. Chadrakasan, M. M. Shulaker
Sponsorship: Analog Devices, DARPA 3DSoC, NSF

Electronics is approaching a major paradigm shift because silicon transistor scaling no longer yields historical energy-efficiency benefits, spurring research towards beyond-silicon nano-technologies. In particular, carbon nanotube field-effect transistor (CNFET)-based digital circuits promise substantial energy-efficiency benefits, but the inability to perfectly control intrinsic nanoscale defects and variability in carbon nanotubes has precluded the realization of very-large-scale integrated systems. Here we overcome these challenges to demonstrate a beyond-silicon microprocessor built entirely from CNFETs: RV16X-NANO. This 16-bit micro-processor is based on the RISC-V instruction set, runs standard 32-bit instructions on 16-bit data and addresses, comprises more than 14,000 complementary metal–oxide–semiconductor CNFETs and is designed and fabricated using industry-standard design flows and processes. We propose a manufacturing methodology (MMC) for carbon nanotubes, a set of combined processing and design techniques for overcoming nanoscale imperfections at macroscopic scales across full wafer substrates. The key elements of MMC are:

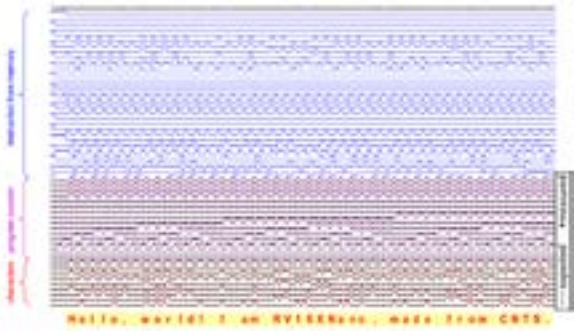
(1) RINSE (removal of incubated nanotubes through selective exfoliation). We propose a method of removing CNT aggregate defects through a selective mechanical exfoliation process. RINSE reduces CNT aggregate defect density by >250× without affecting non-aggregated CNTs or degrading CNFET performance.

(2) MIXED (metal interface engineering crossed with electrostatic doping). Our combined CNT doping process leverages both metal contact work function engineering as well as electrostatic doping to realize a robust wafer-scale CNFET CMOS process. We experimentally yield entire dies with >10,000 CNFET CMOS digital logic gates (2-input 'not-or' gates with functional yield 14,400/14,400, comprising 57,600 total CNFETs), and present a wafer-scale CNFET CMOS uniformity characterization across 150-mm wafers.

(3) DREAM (designing resiliency against metallic CNTs). This technique overcomes the presence of metallic CNTs entirely through circuit design. DREAM relaxes the requirement on metallic CNT purity by about 10,000× (relaxed from a semiconducting CNT purity requirement of 99.999999% to 99.99%),



▲Figure 1: Image of a fabricated RV16X-NANO chip. The die area is 6.912 mm × 6.912 mm, with in-put/output pads placed around the periphery. Scanning electron microscopy images with increasing magnification are shown below.



▲Figure 2: Experimentally measured waveform from RV16X-NANO, executing the 'Hello, World' program. The waveform shows the 32-bit instruction fetched from memory, the program counter stored in RV16X-NANO, as well as the character output from RV16X-NANO. Below the waveform, we convert the binary output (shown in red in hexadecimal code) to their ASCII characters.

FURTHER READING

- G. Hills, C. Lau, et al., "Modern Microprocessor Built from Complementary Carbon Nanotube Transistors," *Nature*, 572, 595–602, 2019.

IN APPRECIATION OF OUR MICROSYSTEMS INDUSTRIAL GROUP MEMBER COMPANIES:

Analog Devices, Inc.
Applied Materials
Draper
Edwards
HARTING
Hitachi High-Tech Corporation
IBM
Lam Research Co.
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
Raith
Waters