Circuits & Systems for Communications, IoT, and Machine Learning

A Sampling Jitter-tolerant Pipelined ADC

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

In a conventional pipelined ADC, the input signal is sampled upfront as shown in Figure 1. Any jitter in the sampling clock directly affects the sampled input and degrades the signal-to-noise ratio (SNR). Therefore, for fast varying input signals, the sampling jitter severely limits the SNR. The error in sampled voltage due to clock jitter is

$$\Delta v = (dv/dt) \cdot \Delta t$$

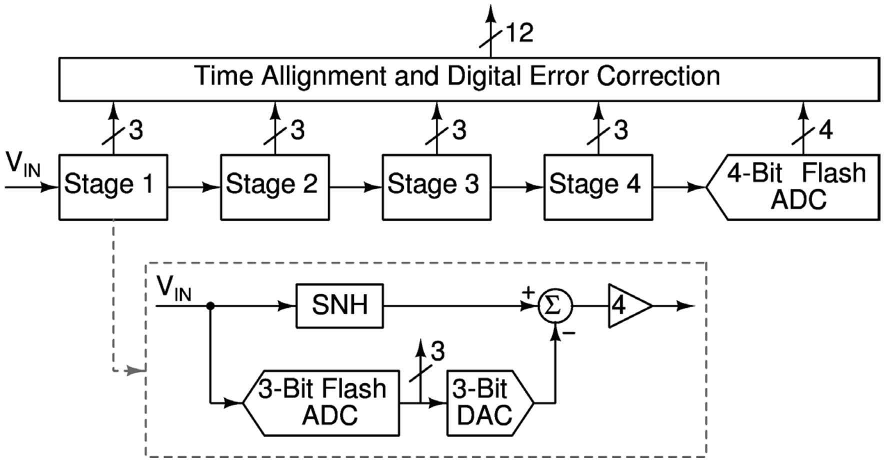
where dv/dt is the time-derivative of the input signal at the sampling instant and Δt is the jitter in the sampling clock. Since the sampling clock jitter is random, it introduces a random noise in the sampled input signal. Also, the error voltage is proportional to dv/dt and hence to the amplitude and frequency of the input signal. Thus, as the frequency of the input signal increases, the effect of sampling clock jitter becomes more pronounced. In fact, it can be shown that for a known rms sampling jitter σ_t the maximum SNR is limited to

$$SNR_{max} = 1/(2\pi f_{in}\sigma_{\tau})$$

where f_{in} is the input signal frequency. Typically, it is difficult to reduce the rms jitter below 100 fs. This limits the maximum SNR to just 44 dB (which is equivalent to 7 bits) for a 10 GHz signal. Therefore, unless the effect of sampling jitter is reduced, the performance of an ADC would be greatly limited for high frequency input signals.

It has been shown that continuous-time delta-sigma modulators (CTDSM) reduce the effect of sampling jitter. But since CTDSMs rely on oversampling, they are not suitable for high frequency signals. Therefore it is imperative to develop sampling jitter-tolerant architectures for Nyquist-rate data converters.

In this project, we propose a new topology that provides increased tolerance to sampling jitter. At present, we are designing the pipelined ADC in 16-nm CMOS technology to give a proof-of-concept for tolerance to sampling jitter.



▲ Figure 1: A conventional 12-bit pipelined ADC.

FURTHER READING

- H. Shibata, V. Kozlov, Z. Ji, A. Ganesan, H. Zhu, and D. Paterson. "16.2 A 9GS/s 1GHz-BW Oversampled Continuous-time Pipeline ADC Achieving-161dBFS/Hz NSD," *Solid-State Circuits Conference (ISSCC), IEEE International*, pp. 278-279. 2017.
- R. van Veldhoven, "A tri-Mode Continuous-time/Spl Sigma//Spl Delta/Modulator with Switched-capacitor Feedback DAC For A GSM-EDGE/CDMA2000/UMTS Receiver," *Solid-State Circuits Conference, 2003. Digest of Technical Papers, ISSCC, IEEE International*, pp. 60-477. IEE 2003.

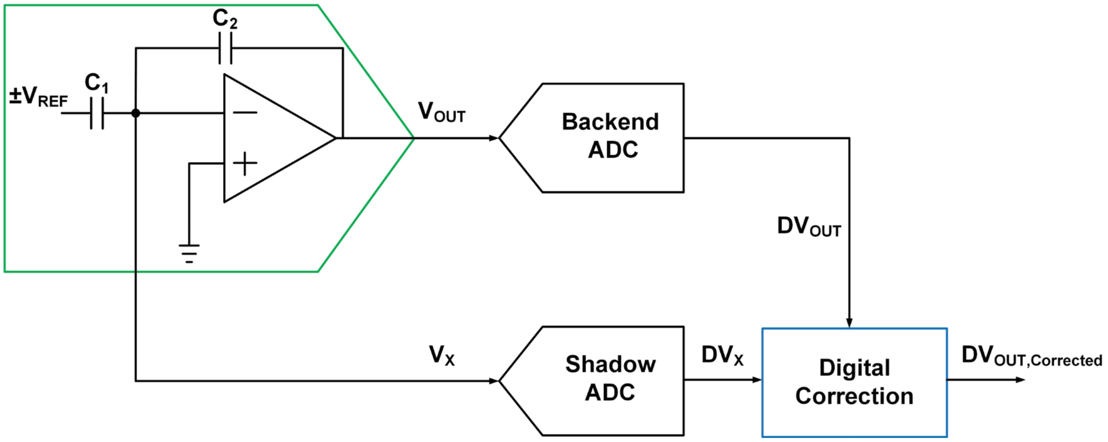
A Pipelined ADC with Relaxed Op-amp Performance Requirements

T. Jeong, A. P. Chandrakasan, H.-S. Lee

Sponsorship: CICS, STMicroelectronics, Korea Foundation for Advanced Studies

Among various analog to digital converter (ADC) architectures, pipelined ADCs are well suited for applications that need medium to high resolution above hundreds-of-megahertz sampling rate. To obtain good linearity, conventional pipelined ADCs must minimize multiplying digital to analog converter (MDAC) charge-transfer error by employing high-gain, fast-settling op-amps. However, such an op-amp design has become increasingly difficult due to the reduced intrinsic gain and voltage headroom in a fine-line CMOS technology. With low intrinsic gain devices, either a gain-boosting technique or a multi-stage topology is necessary to make the op-amp meet the gain requirement. Decreased power supply demands a larger capacitance to maintain the same level of SNR. As a result, the power consumption of these op-amps becomes prohibitively large.

Op-amp non-idealities have been removed or relaxed in digital domain by taking advantage of digital computation to address this issue. In this project, we propose a digital calibration scheme for op-amp-based pipelined ADCs. The ADC relaxes first stage op-amp performance requirements by using a shadow ADC and a simple digital domain calibration algorithm. To validate the functionality of the proposed calibration technique, a proof-of-concept ADC has been designed in 28nm CMOS technology and is currently being tested.



▲ Figure 1: Concept View of Proposed Calibration.

FURTHER READING

- H. H. Boo, et al., "A 12b 250MS/s Pipelined ADC with Virtual Ground Reference Buffers," *IEEE J. of Solid-State Circuits*, vol. 50, no. 12, pp. 2912-2921, Dec. 2015.
- A. Panigada and I. Galton, "A 130mW 100MS/s Pipelined ADC with 69dB SNDR Enabled by Digital Harmonic Distortion Correction," *IEEE J. of Solid-State Circuits*, vol. 44, no. 12, pp. 3314-3328, Dec. 2009.
- L. Brooks and H.-S. Lee, "A Zero-Crossing-Based 8-bit 200MS/s Pipelined ADC," *IEEE J. of Solid-State Circuits*, vol. 42, pp.2677-2687, Dec. 2009.

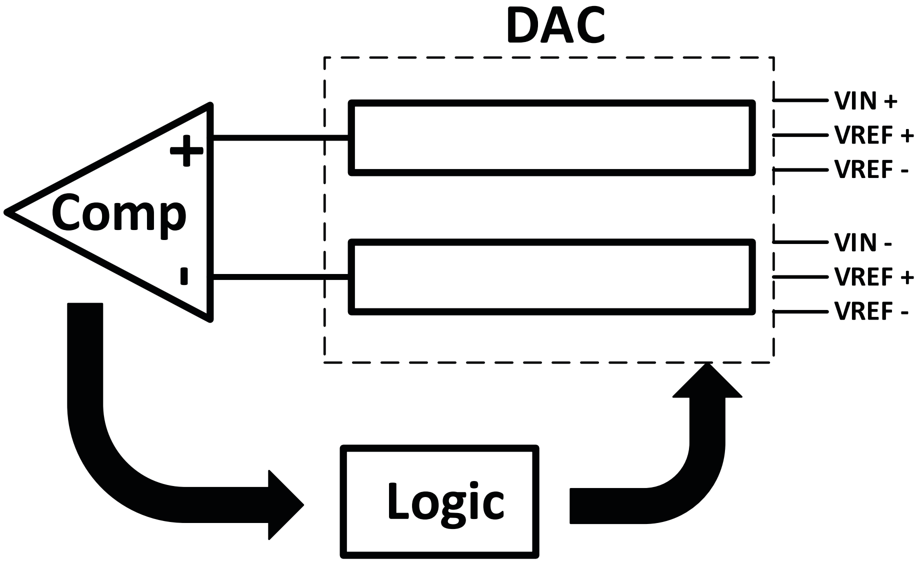
Data-dependent Successive-Approximation-Register Analog-to-Digital Converter

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

This work on successive-approximation-register (SAR) analog-to-digital converters (ADCs) (Figure 1) aims at improving data-dependent savings in energy in key components of a SAR ADC by leveraging the information available from signal's immediate past samples and the signal type. The dominant energy consuming components are the digital-to-analog converter (DAC) and the comparator.

Energy expenditure in the DAC per sample conversion depends on the DAC topology and sequence of steps taken during successive approximation. Energy in the comparator is directly proportional to the number of comparisons done per sample conversion. A design with data-dependent savings takes advantage of the correlation between successive samples in completing the conversion in fewer bit-cycles and also operates the DAC more energy-efficiently.

Previous work presented data-dependent savings by doing least-significant-bit (LSB)-first successive approximation to convert an input sample. By starting with a previous sample and using LSB-first, the algorithm converges in a fewer number of cycles than conventional most-significant-bit (MSB)-first SAR conversion when the present signal is close to the previous signal. Fewer cycles translate into energy savings in the comparator and the DAC. Another work developed successive approximation algorithms to find a sub-range from the full range in a few cycles before carrying on a binary search in this small range. In this work, we investigate a SAR ADC with a search algorithm based on the statistical characteristics of the signal for optimum energy expenditure.



▲ Figure 1: SAR ADC.

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.

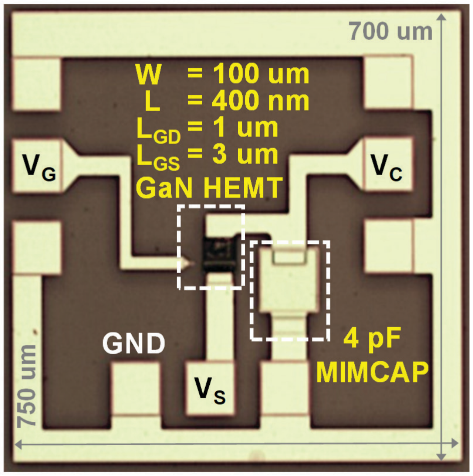
GaN HEMT Track-and-Hold Sampling Circuits with Digital Post-correction on Dynamic Nonlinearity for High-Performance ADCs

S. Chung, P. Srivastava, X. Yang, T. Palacios, H.-S. Lee
Sponsorship: MIT/MTL GaN Energy Initiative, CICS

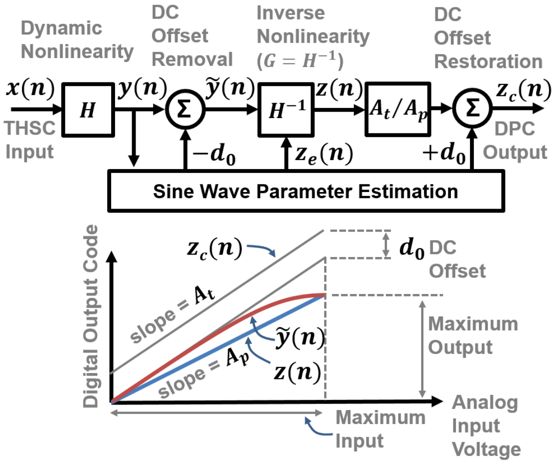
Analog-to-digital converters (ADCs) often limit the performance of integrated systems for emerging applications such as next-generation communication systems, data centers, and quantum computing. The ADC performance is, in turn, limited at least partly by a track-and-hold sampling circuit (THSC). The low supply voltage of deeply scaled complementary metal-oxide-semiconductor (CMOS) transistors determines the THSC input signal range, therefore becoming a fundamental upper bound to the effective number of bits (ENOBs) of CMOS ADCs.

This research work envisions to realize THSCs in GaN-on-Si technology, which monolithically integrates GaN high-electron-mobility transistors (HEMTs) with Si-CMOS transistors, for future ultrahigh-performance ADCs. Operating GaN HEMTs at a high voltage (>30 V) allows a very large input swing (>16 V), providing signal-to-noise ratio (SNR) performance orders of magnitudes beyond the limit of CMOS THSCs. We designed and implemented two GaN HEMT THSCs. The first THSC was fabricated in a commercial GaN foundry technology on SiC substrate, providing 98-dB SNR at 200 MS/s. The second THSC design was fabricated in a GaN technology that was developed at MTL on Si substrate, which operates at 1 GS/s thanks to a higher current-gain cutoff frequency fT and external gate-bootstrapping clock (Figure 1). While these GaN HEMT THSCs achieved an unprecedentedly high SNR at a given input frequency, they suffer from dynamic nonlinearity from the GaN HEMT source-follower buffers for gate-bootstrapping sampling clock generation. Although dynamic nonlinearity correction techniques are mature with RF power amplifiers (PAs), these conventional pre-distortion techniques have high sensitivity to DC offsets, and thus, cannot be directly applied to GaN HEMT THSCs.

To overcome this challenge, we are developing a digital post-correction (DPC) technique, which will demonstrate improved linearity of GaN HEMT THSCs without using a dedicated reference ADC. By applying a DPC technique based on modified Volterra series (Figure 2), we have recently demonstrated that THSC linearity can be improved by more than 20 dB. We are presently working to enhance the linearization performance by applying advanced DPC techniques.



▲Figure 1: Track-and-hold sampling circuit in a GaN technology developed at MTL on Si substrate, which provides over 700-MHz track-mode bandwidth and operates at 1 GS/s.



▲Figure 2: Digital post-correction (DPC) technique for GaN HEMT dynamic nonlinearity compensation, which improves the GaN track-and-hold linearity by more than 20 dB.

FURTHER READING

- S. Chung, P. Srivastava, X. Yang, H.-S. Lee, and T. Palacios, "Digital Post-correction of Nonlinearity with Memory Effects in GaN HEMT Track-and-Hold Circuits for High-performance ADCs," to be presented at *2018 IEEE Radio Frequency Integrated Circuits Symposium*, Philadelphia, PA, Jun. 2018.
- S. Chung, P. Srivastava, X. Yang, H.-S. Lee, and T. Palacios, "GaN HEMT Track-and-Hold Sampling Circuits with Digital Post-correction on Dynamic Nonlinearity," *Proceedings of IEEE Compound Semiconductor Integrated Circuit Symposium*, pp. 1-4, 2017.
- P. Srivastava, S. Chung, D. Piedra, H.-S. Lee, and T. Palacios, "GaN High Electron Mobility Transistor Track-and-Hold Sampling Circuit with Over 100-dB Signal-to-Noise Ratio," *IEEE Electron Device Letts*, vol. 37, no. 10, pp. 1314-1317, Nov. 2016.

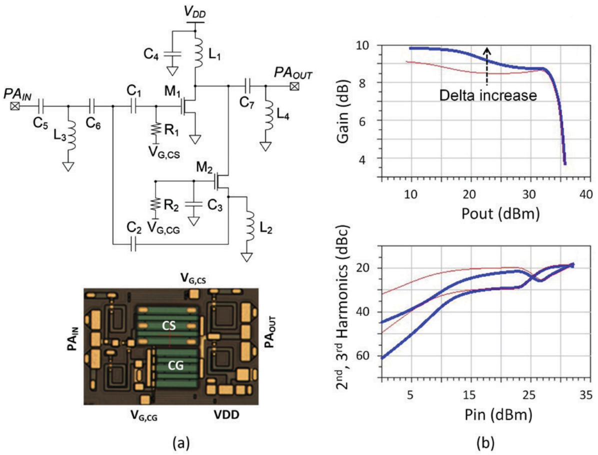
GaN Circuit-device Interaction in Fully Integrated RF Power Amplifiers

P. Choi, U. Radhakrishna, D. A. Antoniadis, E. A. Fitzgerald
Sponsorship: SMART LEES

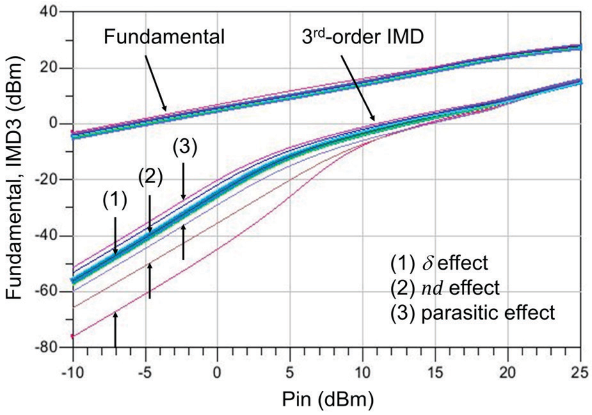
Highly integrated GaN RF power amplifiers (PAs) have been developed for mobile devices and connected cars applications using the physics-based RF transistor compact model, MIT Virtual Source GANFET (MVSG). RF power amplifiers are required to operate in a linear region to prevent signal distortion and resultant data loss, which is mainly affected by inherent device-level nonlinear behavior. Since the second derivative of transconductance, *g3*, is an intrinsic source of intermodulation distortion, many studies aimed to cancel it, especially in CMOS technology. However, the high mobility and thermal effect of GaN devices make the device nonlinearity compensation harder than in CMOS devices. Thus, we have looked into the large signal linearization considering both power gain and third-order harmonics rather than *g3* alteration techniques that cannot be properly functional in a high-power amplifier with large signal input.

In our previous design, the Class-AB + Class-C configuration was proposed for a fully integrated GaN RF amplifier, demonstrating improved linearity and efficiency. Recently, we designed another GaN RF power amplifier with the Common-Source & Common Gate (CS-CG) configuration to further improve the intermodulation distortion by optimizing the third order harmonics performance from the viewpoint of compensating for the large signal distortion. The CS-CG outperforms the Class-AB + Class-C in terms of the third order harmonics and intermodulation distortion, which means that the average time-varying composite *g3* of the CS–CG is lower than that of the Class-AB + Class-C.

To study the impact of the device and technology parameters on the circuit performance, we used both the MVSG model and the CS-CG amplifier and isolated some device parameters which affect the DC and RF performance at both device and circuit levels. Figure 1 shows the circuit implementation using 0.25μm GaN technology and its gain and third- order harmonics with varying DIBL, δ. Intermodulation distortion is further investigated with varying δ, short channel effects such as moderate punch-through, *nd*, and parasitics, i.e., *Cds*, and *Cdg*, as depicted in Figure 2.



▲Figure 1: (a) The CS-CG RF PA, (b) DIBL, δ, changes the threshold voltage as a function of *Vds* and thus makes the circuit behave in a different class, which also affects the composite gain and harmonics output of the CS-CG RF PA.



▲Figure 2: Two-tone intermodulation simulation with varying parameters - δ, *nd*, and parasitics, i.e., *Cds*, and *Cdg*.

FURTHER READING

- P. Choi, S. Goswami, U. Radhakrishna, D. Khanna, C. C. Boon, H.-S. Lee, D. A. Antoniadis, and L.-S. Peh, "A 5.9-GHz Fully Integrated GaN Frontend Design with Physics-based RF Compact Model," *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 4, pp. 1163-1173, Apr. 2015.
- P. Choi, U. Radhakrishna, C. C. Boon, L.-S. Peh, and D. A. Antoniadis, "Linearity Enhancement of a Fully Integrated 6-GHz GaN Power Amplifier," *IEEE Microwave and Wireless Components Lett.*, vol. 27, no. 10, pp. 927-929, Oct. 2017.
- P. Choi, U. Radhakrishna, D. A. Antoniadis, and E. A. Fitzgerald, "GaN Device-circuit Interaction on RF linear Power Amplifier Designed using the MVSG Compact Model," *IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, Miami, FL, Oct. 2017.

Cryptographically Secure Ultra-fast Bit-level Frequency Hopping for Next-generation Wireless Communications

R. T. Yazicigil, P. Nadeau, D. Richman, C. Juvekar, K. Vaidya, A. P. Chandrakasan

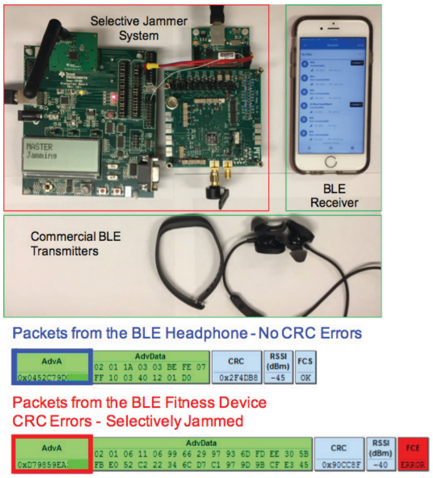
Sponsorship: Hong Kong Innovation and Technology Fund, Texas Instruments, NSF, TSMC

Current Internet-of-Things devices communicate via Bluetooth Low Energy (BLE). Unfortunately, BLE-connected devices are vulnerable to a wide range of attacks; this work specifically addresses selective jamming denial of service where the adversary corrupts transmitted messages targeting a single victim. Selective jamming is particularly challenging as it conceals the attacker's identity contrary to broadband-wireless jamming. To illustrate this type of attack, we demonstrate selective jamming against a commercial fitness BLE-device as shown in Figure 1. This form of attack can cause serious harm such as in the case of insulin pump medical devices.

The primary vulnerability of BLE is founded in the communication protocol which uses frequency hopping to send a message, which is decomposed into data packets, over rapidly changing sub-frequencies. The carrier frequency hops among these sub-frequencies at a relatively slow rate of 612µs per data packet (Figure 2). Conversely, an attacker needs only 1µs to identify the carrier frequency, then block the remainder of the data packet sent on that sub-frequency. To counter this attack, we developed physical-layer security through an ultra-fast bit-level frequency hopping scheme which sends every data bit on a unique carrier frequency while achieving a 1µs hop period (Figure 2).

In addition, a challenging issue is that traditional modulation schemes, such as the BLE Gaussian frequency shift keying (GFSK) modulation with fixed carrier offset of ± 250kHz for Bit 1 and Bit 0, permit the attacker to selectively overwrite individual bits in a packet once the carrier frequency is localized. The attacker gains control over the packet that will be received by the victim. We protect against this attack by implementing a cryptographically secure data-driven dynamic channel selection scheme that enables 80-way pseudorandom FSK modulation and provides data encryption in the physical layer.

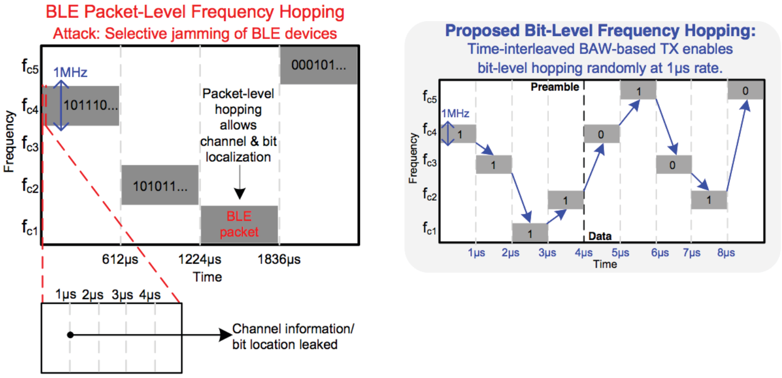
In this work, we demonstrated the first integrated bit-level frequency-hopping transmitter that hops at 1µs period and uses data-driven random dynamic channel selection to enable secure wireless communications with data encryption in the physical layer.


Selective Jammer System
MASTER Jammer
BLE Receiver
Commercial BLE Transmitters
Packets from the BLE Headphone - No CRC Errors
Packets from the BLE Fitness Device CRC Errors - Selectively Jammed

AdvA	AdvData	CRC	RSSI (dBm)	FCS
0x0452C79D	02 01 1A 03 03 BE FE 07 FF 10 03 40 12 01 D0	0x2F4DB8	-45	OK

AdvA	AdvData	CRC	RSSI (dBm)	FCE
0xD79859EA	02 01 06 11 06 99 66 29 97 93 6D FD EE 30 5B FB E0 52 C2 22 34 6C D7 C1 97 9D 9B CF E3 45	0x90CCBF	-40	ERROR

▲ Figure 1: BLE vulnerability to selective jamming attacks.


BLE Packet-Level Frequency Hopping
Attack: Selective jamming of BLE devices
Frequency
f_{c5}
f_{c4}
f_{c3}
f_{c2}
f_{c1}
1MHz
101110...
101011...
000101...
BLE packet
Packet-level hopping allows channel & bit localization
612µs
1224µs
1836µs
Time
1µs 2µs 3µs 4µs
Channel information/ bit location leaked
Proposed Bit-Level Frequency Hopping:
Time-interleaved BAW-based TX enables bit-level hopping randomly at 1µs rate.
Preamble
Data
1
1
1
1
0
1
0
1
0
1µs
2µs
3µs
4µs
5µs
6µs
7µs
8µs

▲ Figure 2: Proposed ultra-fast bit-level frequency hopping with 1µs hop period in contrast to BLE packet-level frequency hopping with a relatively slow hop period of 612µs.

FURTHER READING

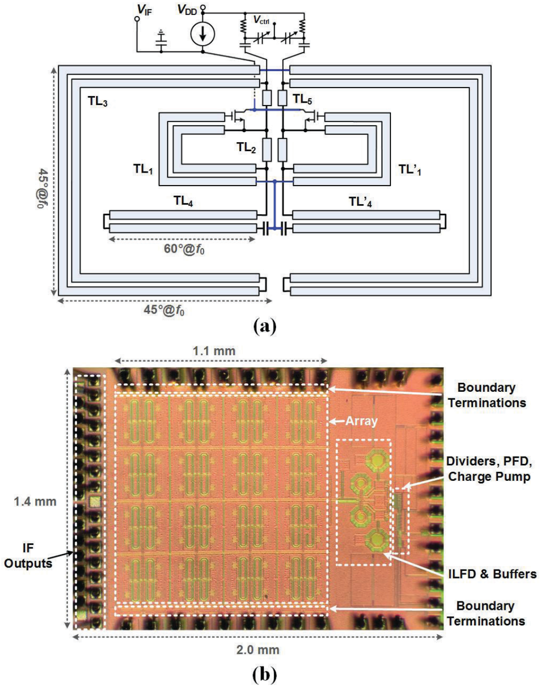
- R. T. Yazicigil, P. Nadeau, D. Richman, C. Juvekar, K. Vaidya, and A. P. Chandrakasan, "Ultra-fast Bit-level Frequency-hopping Transmitter for Securing Low-power Wireless Devices," *IEEE Radio Frequency Integrated Circuits Symposium,* Philadelphia, Pennsylvania, Jun. 2018.

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

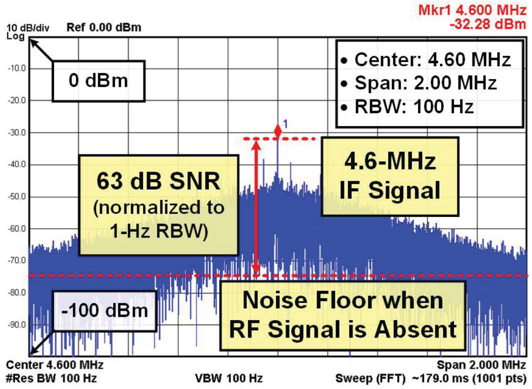
Z. Hu, C. Wang, R. Han
Sponsorship: TSMC, MIT-SMART, NSF

There is a growing interest in pushing the frequency of beam-steering systems towards terahertz range, in which case narrow-beam response can be realized at chip scale. However, this calls for disruptive changes to traditional terahertz 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). In the latter case, corporate feed for generating and distributing the local oscillation signals (LO)— typically a necessary component—can be very lossy at large scale. Here, we report a highly scalable 240-GHz 4×8 heterodyne array achieved by replacing the LO corporate feed with a network that couples LOs generated locally at each unit. A major challenge for this architecture is that each unit should fit into a tight λ/2×λ/2 area to suppress side lobes in beamforming, making the integration of the mixer, local oscillator, and antenna into a unit extremely difficult. This challenge is well-addressed in our design. We have built highly-compact units, which ultimately enables the integration of two interleaved 4×4 phase-locked sub-arrays in 1.2-mm².

The schematic of the circuit of one unit 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. The SOHM is connected to both an intra-unit slot antenna (TL_4 and TL_4') for RF receiving and a co-planar waveguide (CPW)/slotline mesh (TL_3) for strong LO coupling with neighboring SOHMs. Owing to the coupling, LOs generated in each unit can be all locked to an external reference signal so that the array is coherent. Die photo showing the placement of the array and the PLL is given in Figure 1(b). Measured spectrum of 4.6-MHz (below the noise corner frequency) baseband signal is shown in Figure 2, from which we obtain a sensitivity (required incident RF power to achieve SNR=1 at baseband) over 1-kHz detection bandwidth of 38.8pW – more than 6× improvement over state-of-the-art large-scale homodyne arrays.



▲ Figure 1: (a) Circuit schematic (with EM structures) of one receiving unit; (b) die photo of the chip.



▲ Figure 2: Measured 4.6-MHz baseband spectrum.

FURTHER READING

- C. Jiang, A. Mostajeran, R. Han, M. Emadi, H. Sherry, A. Cathelin, and E. Afshari, "A Fully Integrated 320 GHz Coherent Imaging Transceiver in 130 nm SiGe BiCMOS," *IEEE J. of Solid-State Circuits,* vol. 51, no. 11, pp. 2596-2609, 2016.
- K. Sengupta, D. Seo, L. Yang, and A. Hajimiri, "Silicon Integrated 280 GHz Imaging Chipset with 4×4 SiGe Receiver Array and CMOS Source," *IEEE Transactions on Terahertz Science and Technology,* vol. 5, no. 3, pp. 427-437, 2015.
- S. Maas, "*Nonlinear Microwave and RF Circuits,* Boston Artech House, 2003.

THz-Comb-Based Radar for Ultra-Broadband 3-D Imaging

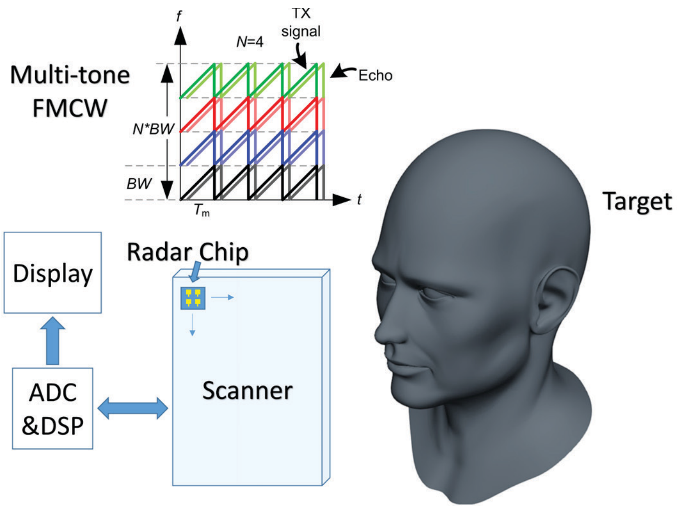
X. Yi, C. Wang, R. Han

Sponsorship: NSF, TSM

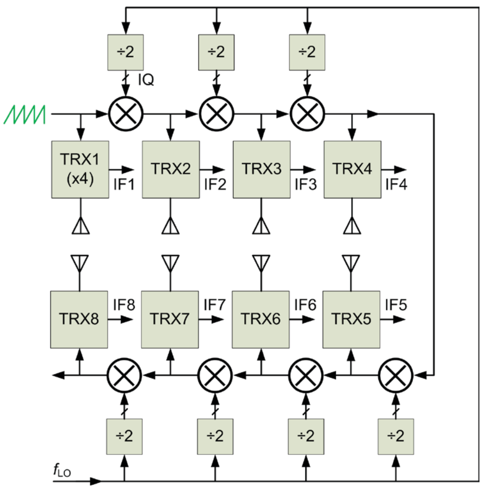
Low-cost 3-D imaging recently becomes increasingly attractive because of its enormous potential in security applications. In particular, waves in the low terahertz (THz) range provide powerful capabilities for 3-D imaging due to the large available bandwidth and improved angular resolution (compared with radio frequency and mm-wave signals), and good transmission (<0.01 dB/m) through extreme weather conditions (compared with infrared and visible light).

We propose a comb radar architecture to increase the bandwidth to more than 0.1 THz without using ultra-wideband components. Shown in Figure 1, it utilizes equally-spaced signal tones with frequency modulation; the generated IF signals are then combined in the digital domain. The proposed comb radar architecture has many advantages compared with conventional Frequency-Modulated Continuous-Wave (FMCW) radar in silicon: peak performance is maintained across a large bandwidth, finer Doppler frequency resolution, larger intermediate frequency (thus smaller flicker noise) and higher linearity. Similar to our previous frequency-comb-based THz spectrometer, in this radar, all components including antennas can be integrated on a single chip, our solution has merits of low cost, small volume, and lightweight.

Figure 2 shows the architecture of the proposed comb radar. It consists of multiple channels with a suitable bandwidth in each channel, leading to an aggregated bandwidth that is larger than 0.1 THz. Note that the number of channels is not limited by the architecture, so the aggregated bandwidth is only limited by the bandwidth of a single channel. The FMCW signal is fed into the first channel directly and up-converted through single sideband mixers to the subsequence channels step by step. The transmitter and the receiver share one on-chip antenna to save the area and power. The mixer first receiver utilizes the transmit power as local oscillator signal and down-converts the received echo signal to IF for further image processing. In addition, since backside radiation has asymmetric radiation pattern and multiple reflections in the attached silicon lens, front-side radiation is desired. To this end, we adopt a substrate-integrated-waveguide antenna utilizing its multiple high-order resonance modes in orthogonal directions. Compared with patch antenna, the new on-chip antenna design has much wider bandwidth (>10% fractional bandwidth).



▲ Figure 1: Proposed chip scale THz comb radar.



▲ Figure 2: Architecture of the THz comb radar.

FURTHER READING

- A. Mostajeran, A. Cathelin, and E. Afshari, "A 170-GHz Fully Integrated Single-chip FMCW Imaging Radar with 3-D Imaging Capability," *IEEE J. Solid-State Circuits,* vol. 52, no. 10, pp. 2721–2734, Oct. 2017.
- J. Grajal, G. Rubio-Cidre, A. Badolato, L. Ubeda-Medina, F. Garcia-Rial, A. Carcia-Pino, and O. Rubinos, "Compact Radar Front-end for an Imaging Radar at 300 GHz," *IEEE Transactions on Terahertz Science and Technology,* vol. 7, no. 3, pp. 268–273, May 2017.
- C. Wang and R. Han, "Dual-Terahertz-Comb Spectrometer on CMOS for Rapid, Wide-range Gas Detection with Absolute Specificity," *IEEE J. Solid-State Circuits,* vol. 52, no. 12, pp. 3361–3372, Dec. 2017.

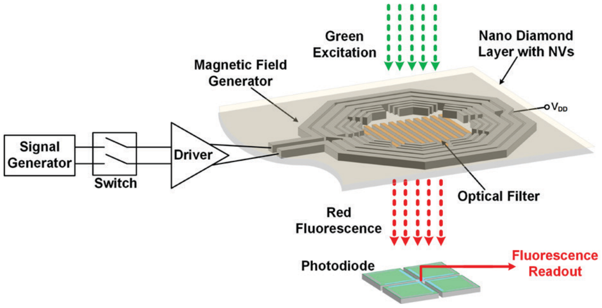
CMOS Chip-scale Vector Ambient Magnetic Field Sensing Based on Nitrogen-vacancy (NV) Centers in Diamond

M. I. Ibrahim, C. Foy, D. Kim, D. R. Englund, R. Han

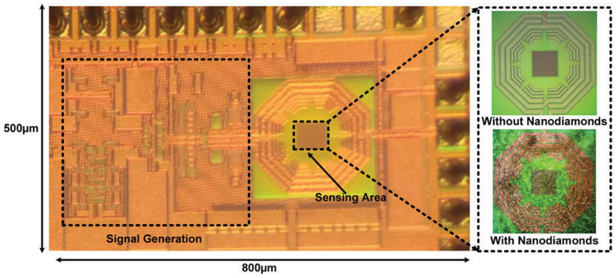
Sponsorship: CICS

Nitrogen-vacancy (NV) centers in diamond have attracted attention for spin-based quantum sensing in ambient conditions. They have demonstrated outstanding nanoscale sensing and imaging capabilities for magnetic-fields. However, these sensing systems require many discrete devices to operate. This limits their scalability. In this work, we demonstrate a chip-scale CMOS and NV integrated platform for magnetic field sensing. The CMOS chip performs the required spin manipulation and readout functions for NV sensing protocols.

Magnetic field sensing is accomplished by determining the spin states of the NV. The frequency of the spin states is determined by through optically detected magnetic resonance (ODMR). The magnetic field is proportional to the frequency splitting of the spin states (2.8 MHz/Gauss). Our system has an on-chip microwave (MW) signal generator, operating from 2.6 GHz to 3 GHz. In addition, an on-chip coil with parasitic loops radiates the AC magnetic field with an amplitude up to 10 Gauss with 95% uniformity over 50 µm x 50 µm. This MW radiation efficiently manipulates the NV spin ensembles. This is followed by on-chip optical readout of the spin state. A CMOS-compatible metal-dielectric structure filters out the optical pump (532 nm) with an isolation of 10 dB. An on-chip patterned P+/N-Well photodiode, beneath the MW coil and the filter, detects the NV red fluorescence. This photodiode is patterned to reduce the unwanted coupling to the MW coil. The measured photodiode responsivity is 230mA/W. The proposed system opens the door for a highly integrated quantum system with applications in the life sciences, tracking, and advanced metrology.



▲Figure 1: Block diagram of the proposed magnetic field sensor. It shows the loop inductor with parasitic loops, the optical filter, and the patterned photodiode.



▲Figure 2: Chip die photo. It shows the sensing area with and without nanodiamonds particles (right inset).

FURTHER READING

- M. I. Ibrahim, C. Foy, D. Kim, D. R. Englund, and R. Han, "Room-temperature Quantum Sensing in CMOS: On-chip Detection of Electronic Spin States in Diamond Color Centers for Magnetometry," *IEEE VLSI Circuits Symposium*, Honolulu, HI, Jun. 2018.
- G. Balasubramanian, I. Y. Chan, R. Kolesov, M. Al-Hmoud, J. Tisler, C. Shin, C. Kim, A. Wojcik, P. R. Hemmer, A. Krueger, T. Hanke, A. Leitenstorfer, R. Bratschitsch, F. Jelezko, and J. Wrachtrup. "Nanoscale Imaging Magnetometry with Diamond Spins under Ambient Conditions," *Nature*, vol. 455, no. 7213, pp. 648-651, 2008.
- J. R. Maze, P. L. Stanwix, J. S. Hodges, S. Hong, J. M. Taylor, P. Cappellaro, L. Jiang, M. G. Dutt, E. Togan, A. S. Zibrov, and A. Yacoby. "Nanoscale Magnetic Sensing with an Individual Electronic Spin in Diamond," *Nature*, vol. 455, no. 7213, pp. 644-647, 2008.

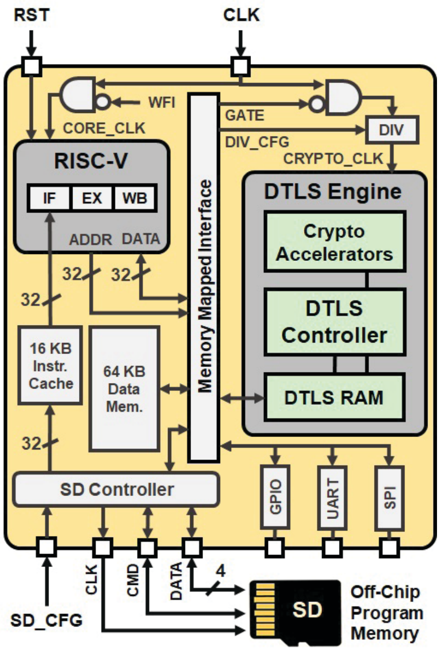
An Energy-efficient Reconfigurable DTLS Cryptographic Engine for End-to-End Security in IoT Applications

U. Banerjee, C. Juvekar, A. Wright, Arvind, A. P. Chandrakasan
Sponsorship: Texas Instruments, Qualcomm Innovation Fellowship

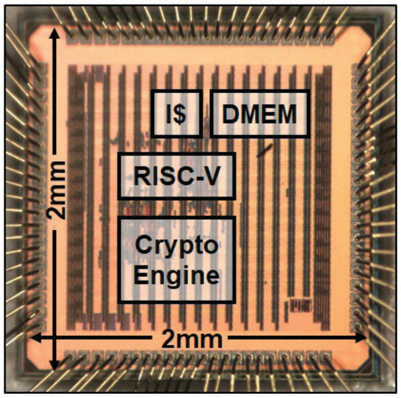
End-to-end security protocols, like Datagram Transport Layer Security (DTLS), enable the establishment of mutually authenticated confidential channels between edge nodes and the cloud, even in the presence of untrusted and potentially malicious network infrastructure. While this makes DTLS an ideal solution for IoT, the associated computational cost makes software-only implementations prohibitively expensive for resource-constrained embedded devices. We address this challenge through the design of energy-efficient hardware to accelerate the DTLS protocol along with associated cryptographic computations.

Figure 1 shows a block diagram of our system, which consists of a 3-stage RISC-V processor, and a memory-mapped DTLS engine supporting the AES-128 GCM, SHA-256, and prime curve elliptic curve cryptography (ECC) primitives. We demonstrate hardware-accelerated DTLS which is 438x more energy-efficient and 518x faster than software implementations. The use of dedicated hardware for DTLS also reduces code size by 78KB and data memory usage by 20KB, thus increasing processor resources available to the application stack.

The test chip, shown in Figure 2, was fabricated in a 65nm LP CMOS process, and it supports voltage scaling from 1.2V down to 0.8V. The RISC-V processor achieves 0.96DMIPS/MHz, consuming 40.36µW/MHz at 0.8V. The DTLS engine consumes 44.08µJ per DTLS handshake, and 0.89nJ per byte of application data, both at 0.8V. Therefore, through the design of reconfigurable energy-efficient cryptographic accelerators and a dedicated protocol controller, this work makes DTLS a practical solution for implementing end-to-end security on resource-constrained IoT devices.



▲ Figure 1: System block diagram, showing the RISC-V processor and the DTLS cryptographic engine, along with instruction cache, data memory, and peripherals.



▲ Figure 2: Micrograph of test chip fabricated in 65nm LP CMOS process.

FURTHER READING

- U. Banerjee, C. Juvekar, A. Wright, and A. P. Chandrakasan, "An Energy-efficient Reconfigurable DTLS Cryptographic Engine for End-to-End Security in IoT Applications," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 42-44, Feb. 2018.
- U. Banerjee, C. Juvekar, S. H. Fuller, and A. P. Chandrakasan, "eeDTLS: Energy-efficient Datagram Transport Layer Security for the Internet of Things," *IEEE Global Communications Conference (GLOBECOM)*, pp. 1-6, Dec. 2017.
- E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.3," *IETF Internet [draft]*, Mar. 2018.

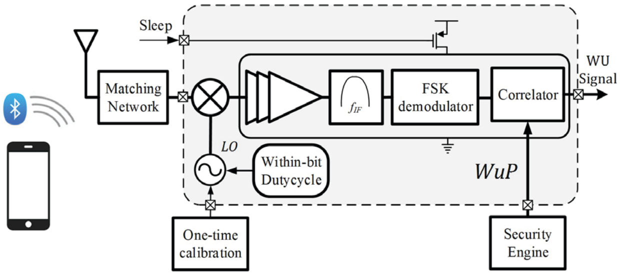
Ultra-Low-Power, High-sensitivity Secure Wake-up Transceiver for the Internet of Things

M. R. Abdelhamid, A. Paidimarri, A. P. Chandrakasan

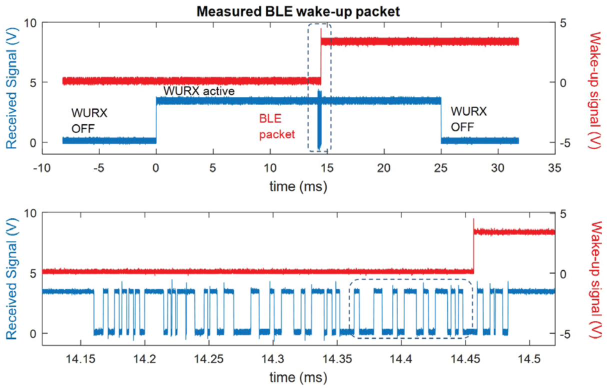
Sponsorship: Delta Electronics

The Internet of Things (IoT) connects together an exponentially growing number of devices with an estimate of more than 70 billion devices in less than ten years from now. Such devices revolutionize the personal heart monitoring, home automation, as well as the industrial monitoring systems. Unfortunately, the wireless IoT nodes consume a huge portion of their energy on communicating with other devices. On the other hand, a longer battery lifetime or even a batteryless energy-harvesting operation requires a sub-microwatt consumption without significant performance degradation. In this work, we propose protocol optimizations as well as circuit-level techniques in the design of a -80dBm sensitivity ultra-low power wake-up receiver for on-demand communication with IoT nodes.

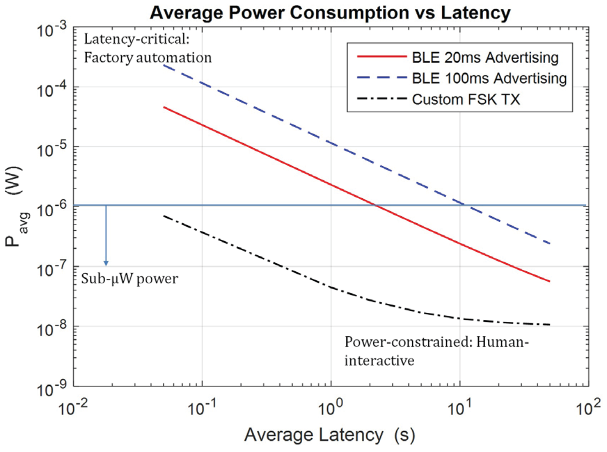
Wireless protocols such as Bluetooth low-energy (BLE) are optimized for short-length packets with small preambles and reduced header sizes. However, the power consumption of a low duty-cycled node in the default connected-mode is limited by the periodic beacons dictated by the protocol. Commercial BLE chips are then limited to tens of microwatts even though their standby power is in the nanowatt range. This wake-up receiver exploits the lower limit of the standby power to achieve significant power reduction through a wake-up scheme wrapped around the BLE advertising protocol. The receiver, shown in Figure 1, employs such duty-cycled wake-up scheme to mitigate the power/sensitivity trade-off achieving sub-microwatt average power at the required BLE sensitivity. When the receiver decodes its wake-up pattern inside the BLE advertising packet, depicted in Figure 2, it generates a wake-up signal then reconfigures its correlator with a new pattern. Figure 3 illustrates the power/latency trade-off where a user with a commercial app can use a cellphone to wake any sleeping IoT node up using the BLE standard according to the application at hand.



▲ Figure 1: Block diagram of the proposed wake-up receiver.



▲ Figure 2: Wake-up using a BLE advertising packet.



▲ Figure 3: Latency-power trade-off.

FURTHER READING

- M. Ding, et al., "A 2.4GHz BLE-Compliant Fully-integrated Wakeup Receiver for Latency-critical IoT Applications using a 2-Dimensional Wakeup Pattern in 90nm CMOS," *IEEE RFIC Symposium Digest of Papers*, pp. 168-171, Jun. 2017.
- N. Roberts, et al., "A 236 nW -56.5dBm-Sensitivity Bluetooth Low-energy Wakeup Receiver with Energy Harvesting in 65nm CMOS," *ISSCC Digest of Technical Papers*, pp. 450-451, Feb. 2016.
- C. Salazar, A. Kaiser, A. Cathelin, and J. Rabaey, "A -97dBm-Sensitivity Interferer-Resilient 2.4ghz Wake-up Receiver using Dual-IF Multi-N-Path Architecture in 65nm CMOS," *ISSCC Digest of Technical Papers*, pp. 242-243, Feb. 2015.

Contactless Current Sensing for Industrial IoT

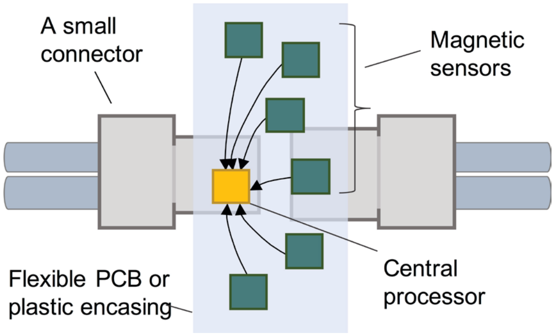
P. Garcha, B. Haroun, S. Ramaswamy, V. Schaffer, D. Buss, J. Lang, A. P. Chandrakasan

Sponsorship: Texas Instruments

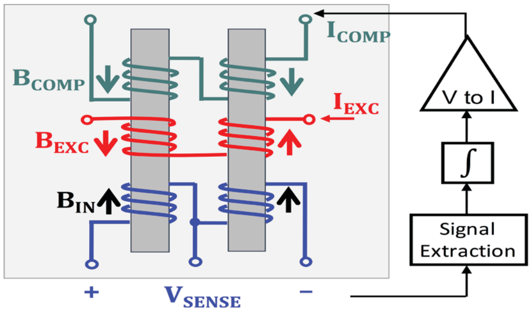
The ability to sense current is crucial to many industrial applications including power line monitoring, motor controllers, battery fuel gauges, etc. We are developing smart connectors with current sensing abilities for use in the industrial internet of things (IoT). These connectors can be used for 1) power quality management: to measure real power, reactive power, and distortion, and 2) machine health monitoring applications for continuous monitoring, control, prevention, and diagnosis. At the system level, the smart connectors need to 1) measure AC, DC, and multiphase currents, 2) reject stray magnetic fields, and 3) detect impending connector failure. On the sensor level, they need to provide high measurement bandwidth (BW) and low power operation.

Current can be sensed directly by using a shunt resistor, but it leads to large power dissipation for measuring high current levels (10-100 A). Indirect/contactless sensing, which senses the magnetic field strength, is a better option as it offers galvanic isolation and the ability to operate safely in high voltage applications. Examples of contactless current sensors include Hall, magneto-resistive (MR), and fluxgate (FG) sensors. FG sensors with integrated magnetics offer higher sensitivity than Hall sensors (nT vs. µT) and higher linearity and lower offset hysteresis than MR sensors, making them a good choice for industrial current sensing.

The proposed system consists of a central processor and multiple low-power, high-BW FG sensors to make synchronous measurements (Figure 1). The measured data from all sensors is stored on the central processor, which runs preliminary analytics on the data before sending it to the cloud. Figure 2 shows the workings of a basic fluxgate sensor design. The proposed sensor makes use of various power saving techniques to reduce the energy per measurement, as well as digitally assisted analog circuits to push for high BW and BW scalability with duty cycling, from >100 kHz BW for machine health monitoring to <1 kHz for power quality management.



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



▲ Figure 2: Typical fluxgate sensor design, with two magnetic cores and three sets of coils: excitation, sense, and compensation. When excited, one core saturates before the other, and $V_{SENSE} \propto `B_{IN}-B_{COMP}`$. Compensation provides feedback to improve linearity and measurement accuracy.

FURTHER READING

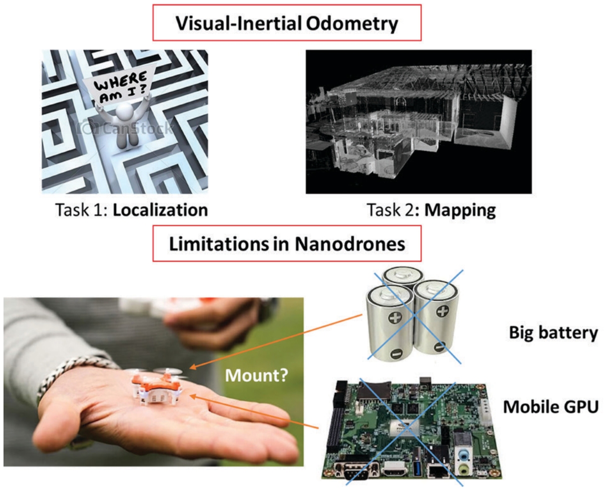
- HARTING. 04. Industrial Connector Han e-Catalogue [Online]. Available: https://www.harting.com/sites/default/files/2018-02/DevCon_07_5_E_Kap04_Industrial-Connectors-Han.pdf, Feb. 2018.
- M. F. Snoeij, V. Schaffer, S. Udayashankar, and M. V. Ivanov, "Integrated Fluxgate Magnetometer for use in Isolated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

Navion: An Energy-efficient Accelerator for NanoDrones Autonomous Navigation in GPS-denied Environments

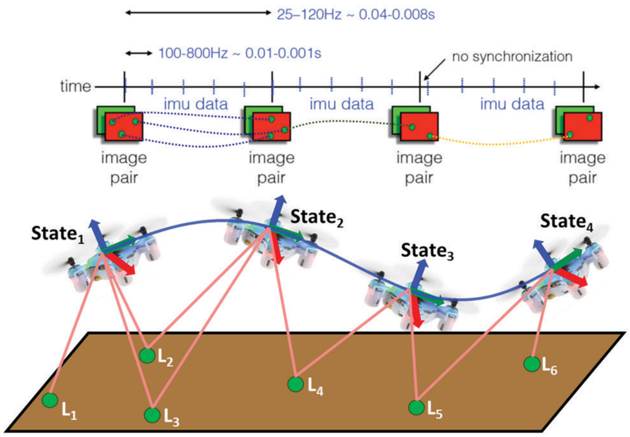
A. Suleiman, Z. Zhang, L. Carlone, S. Karaman, V. Sze
Sponsorship: AFOSR

Drones are getting increasingly popular nowadays. Nanodrones specifically are easily portable and can fit in your pocket. Equipped with multiple sensors; the drone functionality is getting more powerful and smart (e.g., track objects, build 3-D maps, etc.). These capabilities can be enabled by powerful computing platforms (CPUs and GPUs), which consume a lot of energy. The size and battery limitations of Nanodrones make it prohibitive to deploy.

This work presents Navion, an energy-efficient accelerator for visual-inertial odometry (VIO) that enables autonomous navigation of miniaturized robots, and augmented reality on portable devices. The chip fuses inertial measurements and mono/stereo images to estimate the camera's trajectory and a sparse 3-D map. VIO implementation requires large irregularly structured memories and heterogeneous computation flow. The entire VIO system is fully integrated on-chip to eliminate costly off-chip processing and storage. This work uses compression and exploits structured and unstructured sparsity to reduce on-chip memory size by 4.1x. Navion is fabricated in 65nm CMOS. It can process 752x480 stereo images at 171 fps and inertial measurements at 52 kHz, consuming an average 24mW. It is configurable for maximizing accuracy, throughput, and energy-efficiency across different environments. This is the first fully integrated VIO in an ASIC.



▲Figure 1: Visual-inertial odometry output of localization and mapping are the first steps to enable autonomous navigation. With Nanodrones, this is extremely challenging with both power and weight limitations.



▲Figure 2: VIO pipeline takes the visual 3-D point cloud and the IMU data over a time horizon and solves a global optimization problem to refine the drone pose (also called "state").

FURTHER READING

- Z. Zhang, A. Suleiman, L. Carlone, V. Sze, and S. Karaman, "Visual-inertial Odometry on Chip: an Algorithm-and-Hardware co-Design Approach," *Robotics: Science and Systems (RSS)*, 2017.
- C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard, "Past, Present, and Future of Simultaneous Localization and Mapping: Towards the Robust-perception Age," *arXiv*, e:1606.05830, 2016.
- J. Bonnet, P. Yin, M. E. Ortiz, P. Subsoontorn, and D. Endy, "Controlled Flight of a Biologically Inspired, Insect-scale Robot," *Science*, vol. 340, no. 6132, pp. 599-603, 2013.

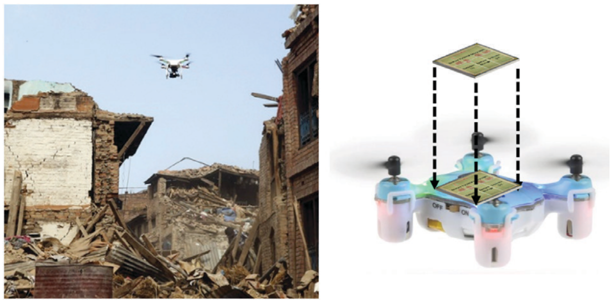
Fast Frontier-exploration for Unmanned Autonomous Vehicles with Resource Constraints

Z. Zhang, S. Karaman, V. Sze
Sponsorship: AFOSR

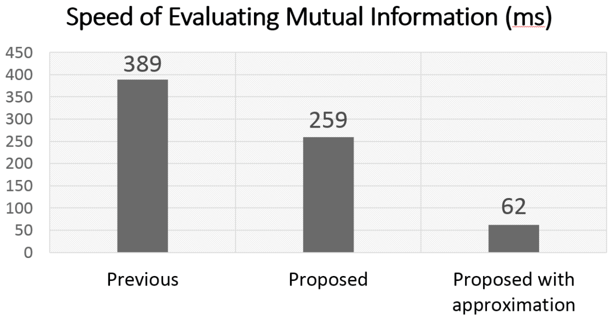
Unmanned Autonomous Vehicles (UAV) have received wide attention. Their capability to autonomously navigate around the environment enables many applications including search-and-rescue, surveillance, wildlife protection and environment mapping. The key technique to empower such capabilities is the frontier-exploration algorithm, which periodically makes decisions on where the vehicle should explore next in an unknown environment based on previously acquired knowledge. However, such algorithms are computationally expensive. In a practical system, the computation is usually offloaded to a powerful computer, causing a significant delay in the response time. This also makes the system strongly dependent on the presence of a stable wireless connection. These factors prohibit the application of the frontier-exploration algorithm to resource-constrained miniature UAVs with limited battery and computation power.

In this work, we present an algorithm to reduce the computation cost of the state-of-the-art mutual information based frontier-exploration algorithm. The key idea behind the algorithm is to use the same computations between different parts of the mutual information computation and reduce redundant computations. Additionally, our approach seeks a more compact representation of the environment, which minimizes the number of operations required to run the algorithm.

In practice, our algorithm enables the complicated frontier-exploration algorithm to be deployed to a battery-powered miniature UAVs with limited computation power. The algorithm makes it possible for the UAV to explore a closed unknown environment with no stable wireless connections. Thanks to the capability of local computation, the latency of running the algorithm is reduced, enabling the UAVs to explore faster and quickly react to the changes in the environment. The saved computation power can be allocated to the actuators of the UAV, enabling the UAVs to stay in the air longer and therefore explore a larger area given fixed power budget.



▲ Figure 1: Fast frontier-exploration algorithm empowers a drone to map an unknown environment autonomously and quickly. The lower complexity of the proposed algorithm enables on-chip computation on a battery-constrained device. It is a key component for applications such as search and rescue.



▲ Figure 2: The proposed algorithm is faster than the previous state-of-the-art algorithm [Charrow 2015]. With reasonable approximation, the algorithm accelerates the computation of mutual information, the bottleneck of the exploration pipeline, by 6x. Therefore this enables the drone to fly and explore the environment more quickly.

FURTHER READING

- B. Charrow, S. Liu, V. Kumar, and N. Michael, "Information-theoretic Mapping using Cauchy-Schwarz Quadratic Mutual Information," *IEEE International Conference on Robotics and Automation*, 2015.
- B. J. Julian, S. Karaman, and D. Rus, "On Mutual Information-based Control of Range Sensing Robots for Mapping Applications," *Int'l J. Robotics Research*, vol. 33, no. 10, pp. 1375-1392, 2014.
- B. Yamauchi, "A Frontier-based Approach for Autonomous Exploration," *Computational Intelligence in Robotics and Automation (CIRA)*, 1997.

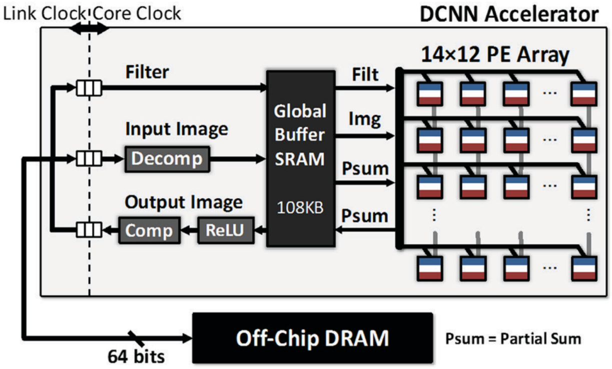
Efficient Processing for Deep Neural Networks

Y.-H. Chen, T.-J. Yang, Y. N. Wu, J. Emer, V. Sze

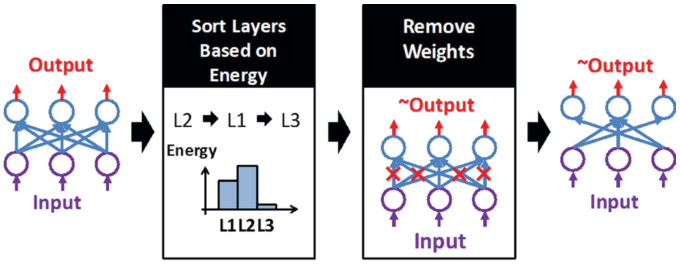
Sponsorship: DARPA YFA, MIT CICS, Intel, Nvidia

Artificial Intelligence powered by deep neural networks (DNNs) has shown great potential to be applied to a wide range of industry sectors. Due to DNNs' high computational complexity, energy efficiency has ever-increasing importance in the design of future DNN processing systems. However, there is currently no standard to follow for DNN processing; the fast-moving pace in new DNN algorithm and application development also requires the hardware to stay highly flexible for different configurations. These factors open up a large design space of potential solutions with optimized efficiency, and a systematic approach becomes crucial.

To solve this problem, we address the co-optimization among the three most important pillars in the design of DNN processing systems: architecture, algorithm, and implementation. First, we present Eyeriss, a fabricated chip that implements a novel data flow architecture targeting energy-efficient data movement in the processing of DNNs (Figure 1). Second, we develop Energy-Aware Pruning (EAP), a new strategy of removing weights in the network to reduce computation so that it becomes more hardware-friendly and yields higher energy efficiency (Figure 2). Finally, we present a tool to realize fast exploration of the architecture design space under different implementation and algorithmic constraints.



▲ Figure 1: System diagram of Eyeriss implementing a novel dataflow architecture targeting energy-efficient data movement in the processing of DNNs. It demonstrated over ten times higher energy efficiency over mobile GPUs that are widely benchmarked for such a task.



▲ Figure 2: EAP removes weights based on their associated energy consumption. EAP demonstrates 3.7 and 1.7 times higher energy efficiency compared to the unpruned DNNs and DNNs pruned with conventional methodologies, respectively.

FURTHER READING

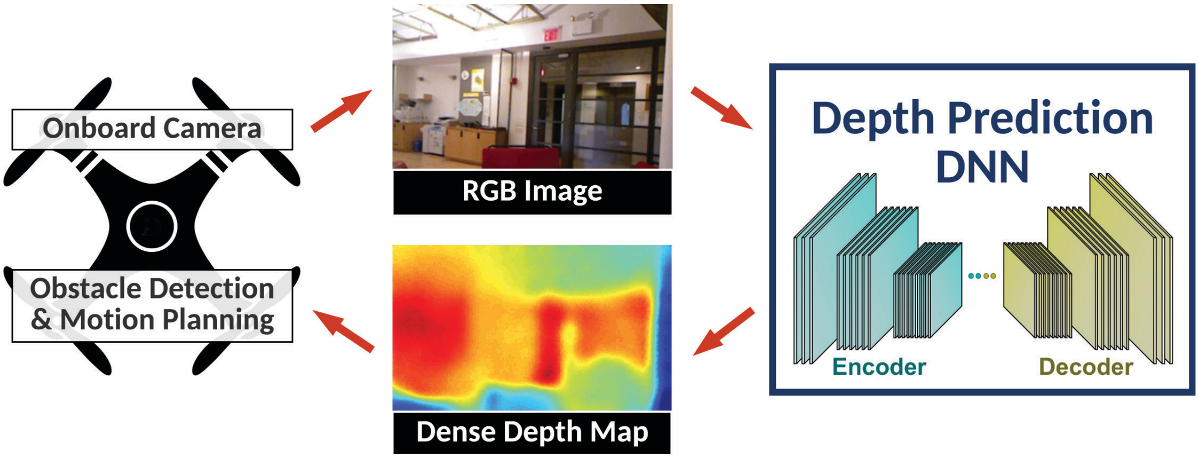
- T.-J. Yang, Y.-H. Chen, and V. Sze, "Designing Energy-efficient Convolutional Neural Networks using Energy-aware Pruning," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Y.-H. Chen, T. Krishna, J. Emer, and V. Sze, "Eyeriss: an Energy-efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE J. of Solid-State Circuits*, vol. 52, pp. 127-138, 2016.
- Y.-H. Chen, J. Emer, and V. Sze, "Eyeriss: A Spatial Architecture for Energy-efficient Data Flow for Convolutional Neural Networks," *Proceedings of the 43rd Annual International Symposium on Computer Architecture (ISCA)*, 2016.

Energy-efficient Deep Neural Network for Depth Prediction

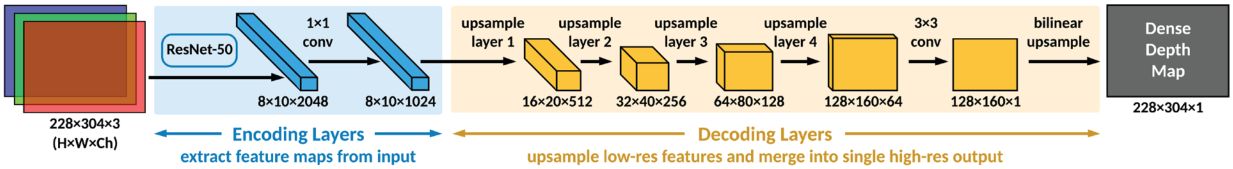
D. Wofk, F. Ma, T.-J. Yang, S. Karaman, V. Sze

Sponsorship: Analog Devices, Inc.

Depth sensing and estimation is a key aspect of positional and navigational systems in autonomous vehicles and robots. The ability to accurately reconstruct a dense depth map of a surrounding environment from RGB imagery is necessary for successful obstacle detection and motion planning. Since deep convolutional neural networks (DNNs) have proven to be successful at achieving high accuracy rates in image classification and regression, recent work in the deep learning space has focused on designing neural networks for depth prediction applications. However, the high accuracy of DNN processing comes at the cost of high computational complexity and energy consumption, and most current DNN designs are unsuitable for low-power applications in miniaturized robots. In this project, we aim to address this gap by applying recently developed methodologies for estimating and improving the energy-efficiency of DNNs to an existing depth-prediction DNN. We envision an outcome in which the depth-prediction DNN is modified to be better suited for a specialized hardware implementation that could be integrated with a low-power visual-inertial odometry system to result in a combined navigational system for miniaturized robots.



▲ Figure 1: Usage of a depth prediction DNN as part of a miniaturized robot's navigational system. In order for the DNN to be integrated onboard the robot, a faster and more energy-efficient design is needed.



▲ Figure 2: Baseline encoder-decoder DNN design used for depth prediction. Our goal is to improve the DNN's energy-efficiency by modifying the encoding and decoding layers and analyzing energy vs. accuracy tradeoffs.

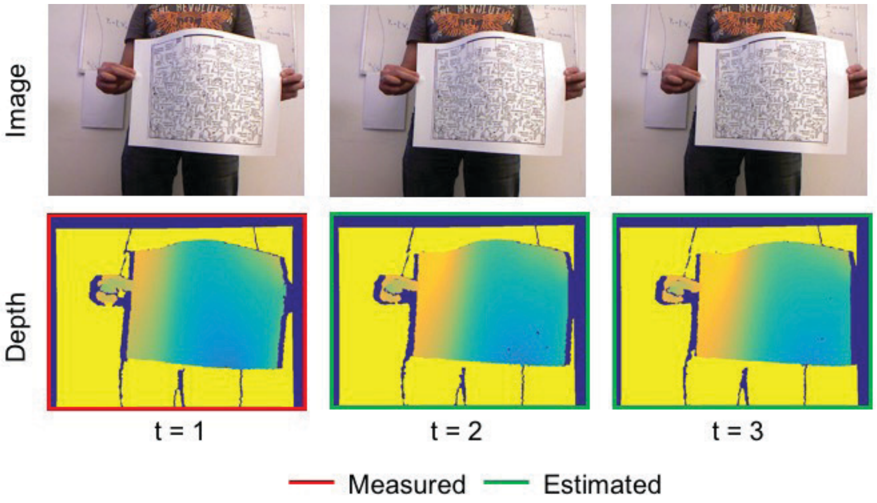
FURTHER READING

- F. Ma and S. Karaman, "Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image," *ICRA*, 2018.
- V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer, "Efficient Processing of Deep Neural Networks: a Tutorial and Survey," 2017.
- T.-J. Yang, Y.-H. Chen, and V. Sze, "Designing Energy-efficient Convolutional Neural Networks using Energy-aware Pruning," *CVPR*, 2017.

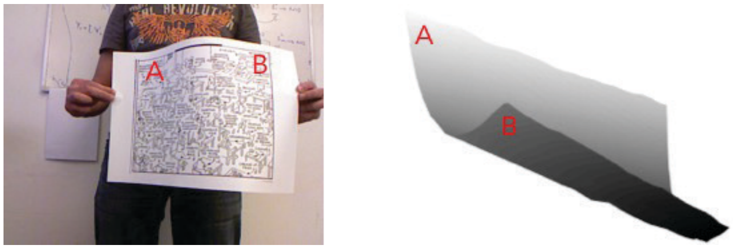
Depth Estimation of Non-Rigid Objects for Time-of-Flight Imaging

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

Depth sensing is used in a variety of applications that range from augmented reality to robotics. Time-of-flight (TOF) cameras, which measure depth by emitting and measuring the roundtrip time of light, are appealing because they obtain dense depth measurements with minimal latency. However, as these sensors become prevalent, one disadvantage is that many TOF cameras in close proximity will interfere with one another, and techniques to mitigate this can lower the frame rate at which depth can be acquired. Previously, we proposed an algorithm that uses concurrently collected optical images to estimate the depth of rigid objects. Here, we consider the case of objects undergoing non-rigid deformations. We model these objects as locally rigid and use previous depth measurements along with the pixel-wise motion across the collected optical images to estimate the underlying 3-D scene motion, from which depth can then be obtained. In contrast to conventional techniques, our approach exploits previous depth measurements directly to estimate the pose, or the rotation and translation, of each point by finding the solution to a sparse linear system. We evaluate our technique on a RGB-D dataset where we estimate depth with a mean relative error of 0.58%, which outperforms other adapted techniques.



▲ Figure 1: Our algorithm estimates depth using the pixel-wise motion across images and previous depth measurements (obtained either from the TOF camera or previously estimated).



▲ Figure 2: We can visualize our estimated depth map with a 3-D reconstruction. Here, the reconstruction is rotated to show the contours of the sheet, where the corresponding corners are indicated by A and B.

FURTHER READING

- J. Noraky and V. Sze, "Low Power Depth Estimation for Time-of-Flight Imaging," *International Conference on Image Processing,* 2017.

Small-footprint Automatic Speech Recognition Circuit

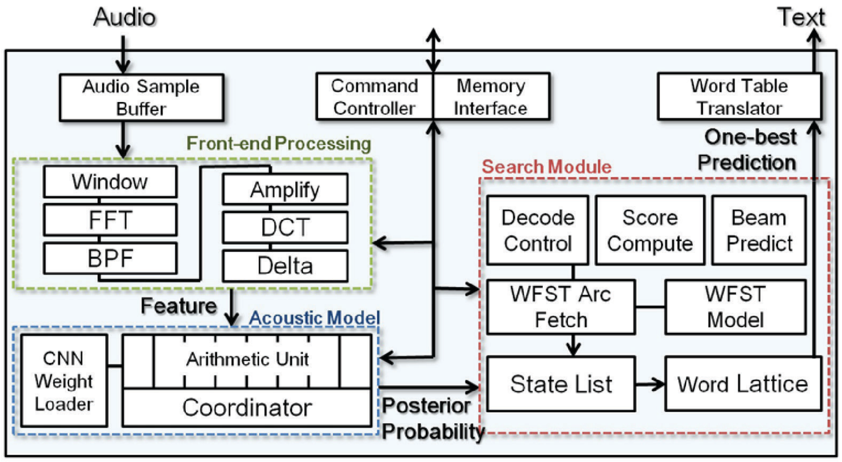
D.-C. Chueh, A. P. Chandrakasan

Sponsorship: Foxconn Technology Group

With the advanced technology of speech and natural language processing, spoken language has become a feasible way for human-machine interaction. Due to the high complexity of articulated speech signal, automatic speech recognition (ASR) generally requires intensive computation and memory size to achieve good performance. However, due to its widespread applications on robots, wearables, and mobile devices, it's desirable to design circuit to implement ASR locally in a resource-limited environment, particularly in which power consumption is a critical concern.

In this work, we first scrutinize software speech recognition procedure; evaluate the memory and computational resource needed when transferring to hardware, and take advantage of circuit design to minimize size and power usage. We design small-footprint ASR system (Figure 1) with cutting-edge neural network that can best perform acoustic modeling with memory restrictions, along with weight truncation and quantization. Dedicated arithmetic unit design, parallelization, and resource dispatching further reduce latency. We implement weighted finite-state transducer (WFST) to incorporate the phonetic probability with language model to select the best word transcription. Model compression, caching, and lattice truncation are adopted to adapt the ASR to circuit and optimize the design.

Our ASR design leveraging powerfulness and robustness of neural network in hybrid ASR model outperforms conventional model in recognition accuracy, whereas conducting ASR tasks on-chip sees great reduction in power compared to CPU. We show a 2.4X reduction in neural network weight size compared to previous hardware design. Our work demonstrates the feasibility to operate an ASR in a small-footprint environment in applications with small vocabulary size and optimized model.



▲ Figure 1: ASR chip design.

FURTHER READING

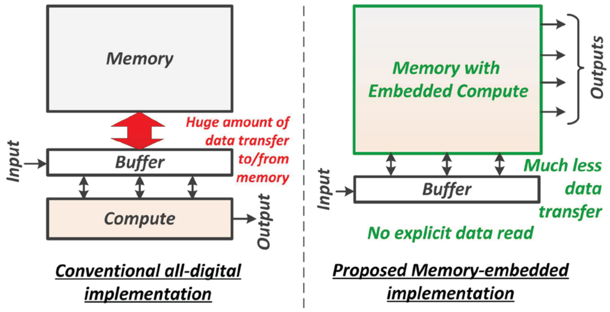
- M. Price, "Energy-scalable Speech Recognition Circuit" PhD Thesis, Massachusetts Institute of Technology, Cambridge, 2016.

In-Memory Computation for Low Power Machine Learning Applications

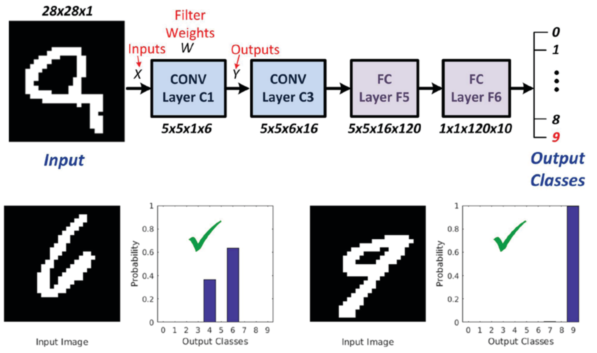
A. Biswas, A. P. Chandrakasan
Sponsorship: Intel Corporation

Convolutional Neural Networks (CNN) have emerged to provide the best results in a wide variety of machine learning (ML) applications, ranging from image classification to speech recognition. However, they require huge amounts of computation and storage. When implemented in the conventional von-Neumann computing architecture, there is a lot of data movement per computation between the memory and the processing elements. This leads to a huge power consumption and long computation time, making CNNs unsuitable for many energy-constrained applications, e.g., smartphones, wearable devices, etc. To address these challenges, we propose embedding computation capability inside the memory (Figure 1). By doing that, we can significantly reduce data transfer to/from the memory and also access multiple memory addresses in parallel, to increase processing speed. The basic convolution operation in a CNN layer can be simplified to a dot-product between the layer inputs (X) and the filter weights (w), to generate the outputs (Y) for that layer. In this work, CNNs are trained to use binary filter weights (w = +/- 1), which are stored as a digital '0' or '1' in bit-cells of the memory array. The digital inputs (X) are converted to analog voltages and sent to the array, where the dot-products are performed in the analog domain. Finally, the analog dot-product voltages are converted back into the digital domain outputs (Y) for further processing.

To demonstrate functionality for a real CNN architecture, the Modified National Institute of Standards and Technology (MNIST) handwritten digit recognition dataset is used with the LeNet-5 CNN (Figure 2). We demonstrated a classification accuracy of 98.35%, which is within 1% of what can be achieved with an ideal digital implementation. We achieved more than 16x improvement in the energy-efficiency in processing the dot-products vs. full-digital implementations. Thus our approach has the potential to enable low-power ubiquitous ML applications for smart devices in the Internet-of-Everything.


Memory
Huge amount of data transfer to/from memory
Input
Buffer
Compute
Output
Conventional all-digital implementation
Memory with Embedded Compute
Outputs
Much less data transfer
Input
Buffer
No explicit data read
Proposed Memory-embedded implementation

▲ Figure 1: Comparison of conventional approach vs. proposed approach of memory-embedded convolution computation, for processing of CNNs.



▲ Figure 2: Demonstrated test case: Handwritten digit recognition task on the MNIST dataset with the LeNet-5 CNN (shown at the top half) and two sample test cases which are correctly classified.

FURTHER READING

- 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.
- V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295-2329, Dec. 2017.
- M. Rastegari, et al., "XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks," *arXiv*, 1603.05279, https://arxiv.org/abs/1603.05279, 2016.

Reconfigurable Neural Network Accelerator using 3-D Stacked Memory Supporting Compressed Weights

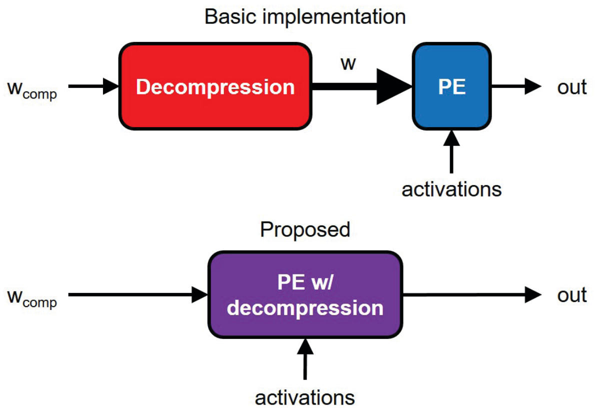
W. Jung, A. Ji, A. P. Chandrakasan

Sponsorship: Taiwan Semiconductor Manufacturing Company (TSMC)

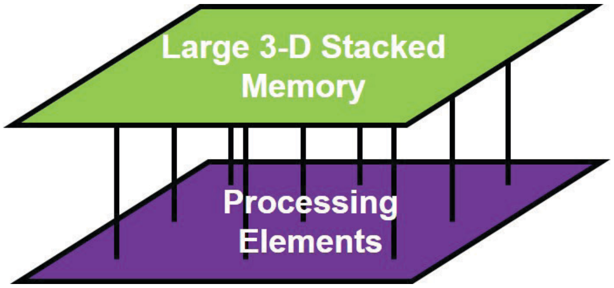
The recent success of machine learning, with the help of emerging techniques, such as convolutional neural networks, have been rapidly changing the way many traditional signal processing problems are being solved, including vision processing, speech recognition, and other prediction and optimization problems. However, neural networks require a large number of weight parameters and processing power that are difficult to accommodate efficiently using a normal CPU architecture. This necessitates dedicated on-chip solutions.

A major challenge in recent on-chip neural network processors is reducing the energy consumed by memory accesses, as the cost for data operations becomes relatively cheaper than the cost for data movement in recently advanced processes. One approach is to simply reduce the amount of data movement by using compression schemes (i.e., reducing the bit-width of weights and activations). Han, et al. develop a deep compression technique to non-uniformly quantize floating point weights to 4-bit values, without any loss of accuracy. This was further extended to quantizing to only 2-bit ternary weights. Another approach is to increase the memory capacity, for example with 3-D stacked memory, to reduce the required number of costly external DRAM accesses.

Our proposed design takes full advantage of these compression schemes by directly integrating the decompression within the processing element. In addition, the design can be reconfigured to perform more general fixed-point computations with variable bit-widths. Combining this with a closely integrated memory chip through 3-D stacking makes it possible to run large networks with less data movement to and from the external DRAM, resulting in improved energy efficiency compared to other implementations.



▲ Figure 1: Comparison between basic (top) and proposed (bottom) implementations of processing element (PE) with the deep compression scheme. The proposed design reduces the number of operations required as it operates on compressed (w_{comp}) rather than uncompressed weights (w).



▲ Figure 2: 3-D integration of the memory chip with the processor chip. This provides significant additional storage for the processor with relatively low access energy (compared to external DRAM).

FURTHER READING

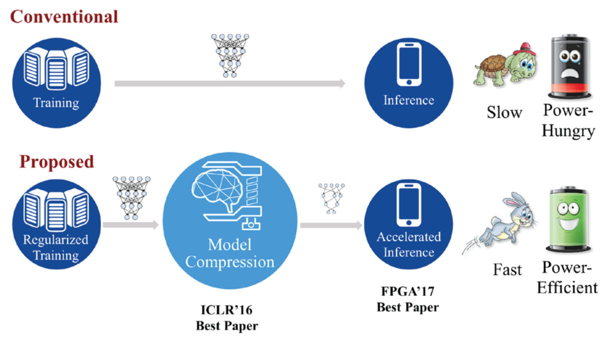
- J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU: A 50.6TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-variable Weight Bit-precision," *ISSCC*, 2018.
- S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization, 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.

Bandwidth-efficient Deep Learning: Algorithm and Hardware co-Design

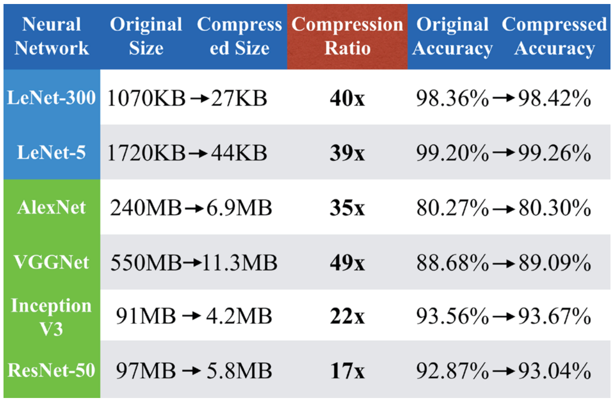
S. Han

In the post-ImageNet era, computer vision and machine learning researchers are solving more complicated Artificial Intelligence (AI) problems using larger data sets driving the demand for more computation. However, we are in the post-Moore's Law world where the amount of computation per unit cost and power is no longer increasing at its historic rate. This mismatch between supply and demand for computation highlights the need for co-designing efficient machine learning algorithms and domain-specific hardware architectures. By performing optimizations across the full stack from application through hardware, we improved the efficiency of deep learning through smaller model size, higher prediction accuracy, faster prediction speed, and lower power consumption.

Our approach starts by changing the algorithm, using "Deep Compression" that significantly reduces the number of parameters and computation requirements of deep learning models by pruning, trained quantization, and variable length coding. "Deep Compression" can reduce the model size by 18× to 49× without hurting the prediction accuracy. We also discovered that pruning and the sparsity constraint not only applies to model compression but also applies to regularization, and we proposed dense-sparse-dense training (DSD), which can improve the prediction accuracy for a wide range of machine learning tasks. To efficiently implement "Deep Compression" in hardware, we developed EIE, the "Efficient Inference Engine," a domain-specific hardware accelerator that performs inference directly on the compressed model which significantly saves memory bandwidth. Taking advantage of the compressed model, and being able to deal with the irregular computation pattern efficiently, EIE improves the speed by 13× and energy efficiency by 3,400× over GPU.



▲ Figure 1: Improving the latency and energy efficiency of deep learning by regularized training, model compression and accelerated inference with domain-specific hardware architecture.



Neural Network	Original Size	Compressed Size	Compression Ratio	Original Accuracy	Compressed Accuracy
LeNet-300	1070KB	→27KB	40x	98.36%	→98.42%
LeNet-5	1720KB	→44KB	39x	99.20%	→99.26%
AlexNet	240MB	→6.9MB	35x	80.27%	→80.30%
VGGNet	550MB	→11.3MB	49x	88.68%	→89.09%
Inception V3	91MB	→ 4.2MB	22x	93.56%	→93.67%
ResNet-50	97MB	→ 5.8MB	17x	92.87%	→93.04%

▲ Figure 2: The Deep Compression algorithm can compress modern deep neural networks by 17x-49x with no loss of accuracy, saving computation and memory bandwidth.

FURTHER READING

- S. Han, et al, "ESE: Efficient Speech Recognition Engine for Sparse LSTM," *International Symposium on FPGA (FPGA)*, 2017.
- S. Han, et al, "Deep Compression," *International Conference on Learning Representations (ICLR)*, 2016.
- S. Han, et al, "EIE: Efficient Inference Engine for Sparse, Compressed Neural Network," *International Conference on Computer Architecture (ICLR)*, 2016.