Circuits & Systems

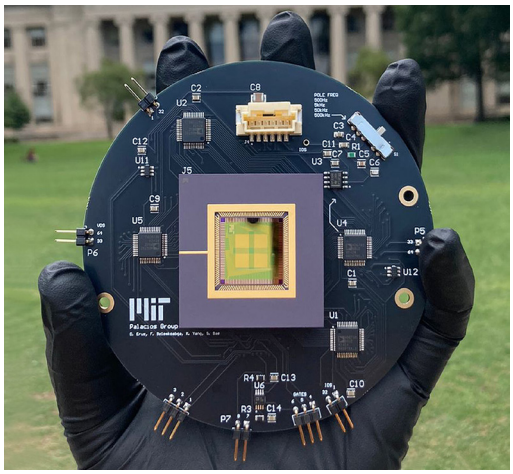
Integrated System to Control and Measure Graphene Field-effect Transistor Arrays 11
High-Angular-Resolution Sub-THz Imaging System with Antenna-in-Package (AiP) Technology 12
SPIPE: Differentiable SPICE-level Co-simulation Program for Integrated Photonics and Electronics 13
Terahertz Tactile and Non-contact Sensors for Robotic Manipulation and Training 14
Reinforcement Learning for Verilog Code Generation with Functional Feedback 15
Highly Selective Harmonic-resilient Sub-6G Front Ends 16
Defining and Optimizing Write/Clear Margin for a Nanoantenna-Based Petahertz Electronic Memory Cell 17
Memory-Efficient Gaussian Mapping on Micro-Robots: Algorithm and Chip 18
Design and Modeling of High Temperature Gallium Nitride RF Amplifiers 19
Harmonic-Resilient Low-Power Receiver Architecture with Pipeline Mixing for IoT Applications 20
A 28 GHz Coupled PLL-Based CMOS Quadrature Oscillator 21
Interface Circuits for Analog In-Memory Computing 22
An Analog Front End with Sparse-Image Capturing for Energy-Efficient Bladder Ultrasound Imaging 23
A 232-to-260GHz CMOS Amplifier-Multiplier Chain with a Matching-Sheet-Assisted Radiation Package and 11.1dBm Total Radiated Power 24
Efficientvit-SAM: Accelerated Segment Anything Model without Performance Loss 25

Integrated System to Control and Measure Graphene Field-effect Transistor Arrays

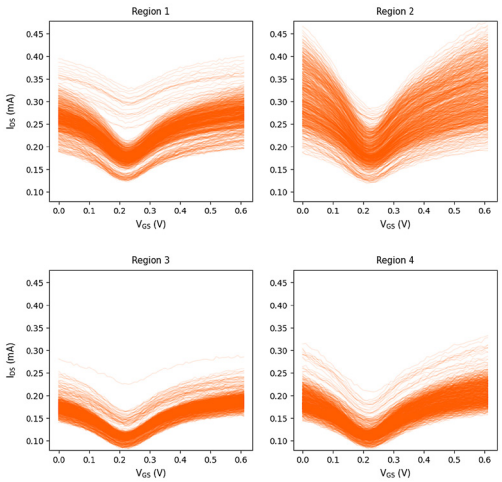
S. Bae, D. Erus, F. Belemkoabga, C. Lopez, C.-H. Liu, T. Palacios
Sponsorship: NSF CIQM, NSF Convergence

Two-dimensional materials such as graphene have shown great promise as biochemical sensors. As a zero-bandgap material, graphene field effect transistors (GFETs) exhibit a Dirac point shift when different ions (i.e., dopants) come into contact with the transistor. This shift can be observed from the GFET's I-V characteristics, enabling the detection of changes in chemical environments.

Our group has designed and fabricated a scalable, handheld chemical sensing system with 4096 sensing units, each consisting of a GFET (Figure 1). While previous research has demonstrated the feasibility of ion detection with electrolytegated transistors, we propose an integrated system that controls and measures both liquid-gated and gas-sensing transistors. Specifically, the custom-designed printed circuit board (PCB) enables extraction of the electrical characteristics of 4096 functional GFETs individually by multiplexing a single pair of drain and source connections to the measurement circuitry (Figure 2). This PCB facilitates automated, real-time analysis of data regardless of transistor quality, applied dopant, and sensing medium, advancing its potential for use as a portable biomedical diagnostic device.



▲Figure 1: Proposed integrated measurement system incorporating a GFET array chip. The platform supports both liquid-gated and gas-sensing modalities, enabling the characterization of 4,096 transistors in under three seconds.



▲Figure 2: I_{DS}-V_{GS} characteristics of 4096 functionalized GFET units measured in PBST. The system automatically determines the Dirac point—the gate voltage at minimum conductivity—of each unit using fifth-order polynomial fitting.

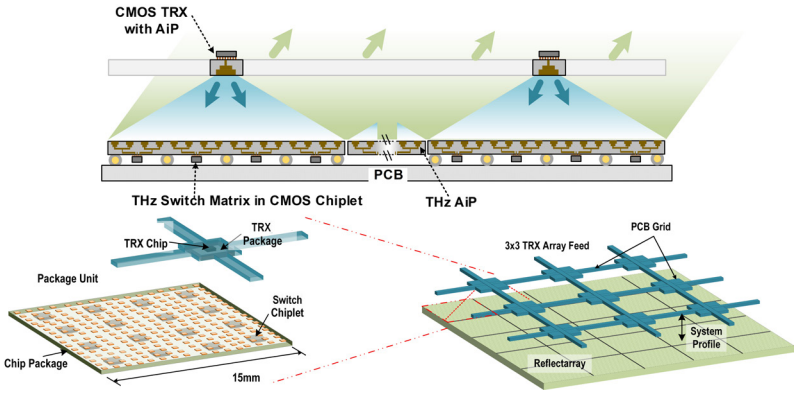
FURTHER READING

- M. Xue, C. Mackin, W.-H. Weng, J. Zhu, Y. Luo, S.-X. L. Luo, A.-Y. Lu, M. Hempel, et al., "Integrated Biosensor Platform Based on Graphene Transistor Arrays for Real-Time High-Accuracy Ion Sensing," *Nature Communications*, vol. 13, no. 5064, 2022.
- Y. Ohono, K. Maehashi, Y. Yamashiro, and K. Matsumoto, "Electrolytegated Graphene Field-effect Transistors for Detecting pH and Protein Adsorption," *Nano Letts.*, vol. 9, pp. 3318–3322, 2009.

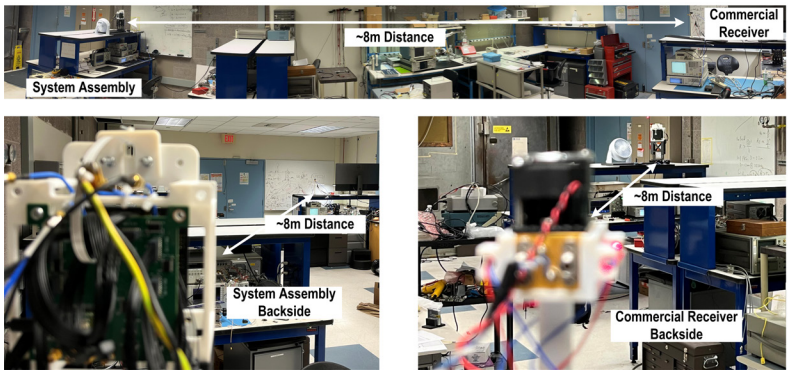
High-Angular-Resolution Sub-THz Imaging System with Antenna-in-Package (AiP) Technology

X. Chen, G. C. Dogiamis, R. Han
Sponsorship: Intel Corporation

The high-angular-resolution imaging capability of future automotive and security sensing systems is in favor of compact, fully integrated, and reconfigurable antenna arrays. The adoption of sub-terahertz (sub-THz) frequencies in such systems relieves the hardware requirements for the physical size and fractional bandwidth. This project proposes a sub-THz four-dimensional (4D) imaging system that decouples the designs of active circuits (transceivers) and a large passive antenna array (reflectarray), which naturally circumvents those challenges of circuit complexities, electronic density, and computation power in traditional phased/Multiple-Input Multiple-Output arrays. On the transceiver (TRX) side, a 250-GHz TRX system with antenna-in-package (AiP) technology is developed and assembled; it shows a significant radiation efficiency improvement over the on-chip solution. On the reflectarray side, AiP is used for the antenna array. Only digital circuits and memories are integrated on 400 tiny chiplets mounted on 25 package modules, which reduces the total silicon area by more than 10x compared to a full-silicon solution. To further enhance the aperture size and address the small gaps between adjacent package modules in the assembly process, the array is designed in an aperiodic manner, so that the grating lobes can be well suppressed. Combining the two sub-systems (TRX & reflectarray) with a full-duplex technique based on reflectarray-TRX cooperation enables the complete 4D imaging system with a needle-beam. Simulation shows a 0.5˚ round-trip beamwidth, which leads to a 0.25˚ angular resolution for radar imaging. Figure 1 shows the system architecture, and Figure 2 shows one typical measurement setup of the assembled system.



◀ Figure 1: Proposed sub-THz system with multi-transceiver feeds and AiP technology.



▶ Figure 2: Die micrograph of the proposed harmonic rejection receiver.

FURTHER READING

- X. Chen, "A 265-GHz CMOS Reflectarray With 98×98 Elements for 1˚-Wide Beam Forming and High-Angular-Resolution Radar Imaging," *IEEE J. of Solid-State Circuits*, vol. 59, no. 11, pp. 3655-3669, Nov. 2024. DOI: 10.1109/JSSC.2024.3393021
- X. Chen, "A 140-GHz FMCW TX/RX-Antenna-Sharing Transceiver With Low-Inherent-Loss Duplexing and Adaptive Self-Interference Cancellation," *IEEE J. of Solid-State Circuits*, vol. 57, no. 12, pp. 3631-3645. Dec. 2022. DOI: 10.1109/JSSC.2022.3202814

SPIPE: Differentiable SPICE-level Co-simulation Program for Integrated Photonics and Electronics

Z. Gao, D. S. Boning

Heterogeneous photonic-electronic systems, such as co-packaged optics and photonic-electronic artificial intelligence (AI) accelerators, are rapidly gaining traction but also pose significant design challenges due to distinct design methodologies. Digital and analog electronics are typically described using hardware description languages and Software Improvement and Capability Determination (SPICE), respectively, whereas photonic devices and systems are represented using permittivity tensors on the Yee grid and the scattering matrix formulation. This disparity necessitates an end-to-end photonic-electronic co-simulation tool to streamline co-design. Most preliminary co-simulation approaches rely on translating photonic compact models into Verilog-A or SPICE models to simulate everything there, which not only introduces the additional complexity of model conversion but also has potential numerical stability problems. Additionally, another critical functionality missing from the current implementation is enabling gradient calculation in these co-simulators, which will be crucial for end-to-end gradient-based electronic-photonic system optimization.

We propose a differentiable SPICE-level co-simulation program for integrated photonics and electronics (SPIPE). We develop a customized differentiable frequency-domain scattering matrix simulation for the photonic components and, leveraging the time-domain adjoint method, enable differentiable transient simulation for the electronic components as well. SPIPE accepts a text file using an extended SPICE syntax to describe the photonic-electronic circuit and outputs both the analytical optical signal values (i.e., complex numbers representing the magnitude and phase of the electromagnetic (EM) mode coefficients) and the electrical signal values (i.e., real numbers representing voltage or current). SPIPE is the first co-simulator to overcome model conversion issues (e.g., eliminating the need to convert photonic compact models into Verilog-A or SPICE) and provide differentiability. Numerical experiments on several circuits confirm the accuracy of SPIPE when compared to analytical solutions and real-world experimental data. Further, in cases where existing simulators are applicable, SPIPE achieves a runtime reduction of 2~85x compared to an industry-standard simulator. In summary, SPIPE has two major advantages. First, SPIPE eliminates model conversion issues (i.e., translating photonic models into Verilog-A or SPICE). Instead, it directly utilizes SPICE for electronic simulation and the scattering matrix method for photonic simulation, with a customized interface to facilitate seamless interaction between the two domains. Second, SPIPE is differentiable, enabling the computation of derivatives of an optical signal with respect to voltage or current signals.

FURTHER READING

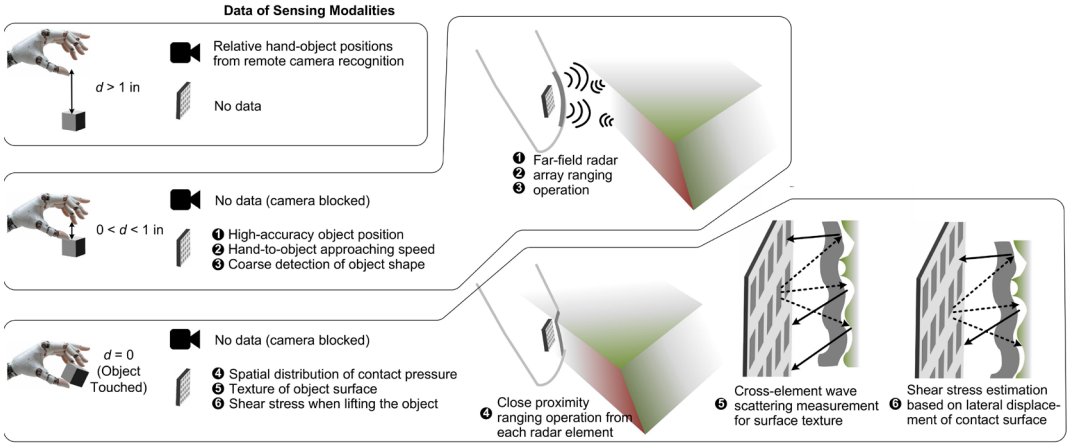
- C. S. Agaskar, J. Leu, M. R. Watts, and V. Stojanovic, "Electro-optical Co-simulation for Integrated CMOS Photonic Circuits with VerilogA," *Optics Express*, pp. 27180-27203, 2015.

Terahertz Tactile and Non-contact Sensors for Robotic Manipulation and Training

K. R. Pochana, R. Han
Sponsorship: Analog Devices/MIT Generative AI Impact Consortium

Similar to Large Language Models, Large Behavior Models in robotics have shown promise in learning general capabilities for interacting in dynamic environments, essential for robots that will operate in human environments and alongside humans. Within these efforts, tactile sensors, which provide contact-based shape and force information about objects as they are touched, have become widely regarded as critical to robust performance. However, current mainstream tactile sensors use optical and vision systems, resulting in bulky and rigid-frame packages that limit robot form factors and learning approaches. We aim to develop a thin-package solution for tactile sensing that reduces robot design constraints and lowers the barrier to data collection needed for robotic learning methods. In addition to high-resolution (<0.5 mm) contact pressure spatial mapping, we aim to provide pre-contact position and shape information, a new modality of information for robotic manipulators.

The key enabler of this technology is the use of an on-chip micro-radar array that operates in the sub-terahertz probing frequency. The corresponding small wavelength (~1 mm) allows for a dense array (0.25 mm element pitch) and sub-mm-level accuracy in distance and shape detection. The proposed system consists of a 5x5 micro-radar array implemented in a custom sensor chip and coupled to in-package antennas to minimize the amount of silicon used for a potentially flexible sensor system. A deposited layer of elastic material on top of these antennas serves as a skin-like interface to contact environmental objects. As the surface deforms, the micro-radar array measures it. Since distance measurements can be taken before an object contacts the sensor, pre-contact shape information can be constructed as well. This modality of information can allow more reliable and precise maneuvers in conditions where vision is occluded and complements vision-based sensing in ideal conditions, ultimately expanding the capabilities of robotic manipulation.



▲ Figure 1: Operation of the proposed sensor. The sensor consists of a 2D-antenna array with bonded chiplets and an elastomer contact surface. In ranges where a camera cannot provide data for behaviors, our sensor provides pre-contact information and high-resolution tactile information, including important components such as shear forces.

FURTHER READING

- B. Romero, H.-S. Fang, P. Agrawal, and E. Adelson, "EyeSight Hand: Design of a Fully Actuated Dexterous Robot Hand with Integrated Vision-Based Tactile Sensors and Compliant Actuation," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.
- Toyota Robotics Institute, "Large Behavior Models," *[Link]*

Reinforcement Learning for Verilog Code Generation with Functional Feedback

J. Shi, Z. Gao, D. S. Boning

The increasing complexity of hardware systems has made the automatic generation of register-transfer-level code an important yet challenging problem. While large language models (LLMs) have demonstrated strong capabilities in generating syntactically correct hardware code, ensuring functional correctness and optimizing for design constraints remains a major challenge.

In this work, we propose a reinforcement learning (RL)-based framework for Verilog code generation, where an LLM is fine-tuned with functional feedback. The model takes as input a natural language description of a hardware module's functionality, along with its interface specification, and generates the corresponding Verilog implementation. Our framework employs proximal policy optimization. We leverage simulation tools like Icarus Verilog and functional equivalence checking with Yosis to compute feedback for each generated candidate. The reward is shaped to encourage both successful compilation and functional equivalence with a reference design, allowing the model to learn effective exploration strategies.

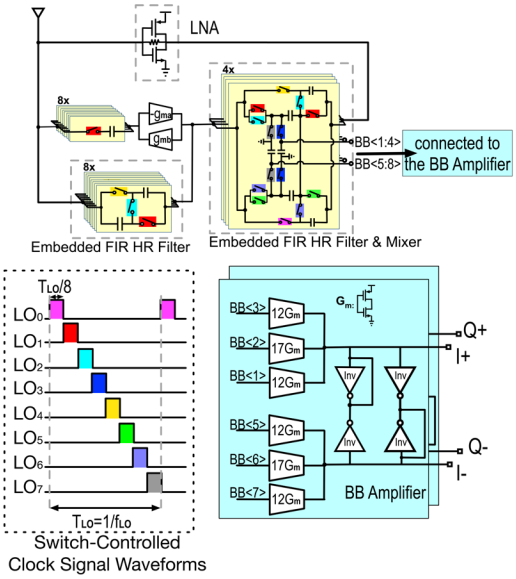
Preliminary results demonstrate that RL training significantly improves functional pass rates compared to baselines. The trained model is capable of generating Verilog programs that not only compile successfully but also satisfy formal equivalence checks with reference designs

This work contributes to the development of AI-driven electronic design automation (EDA) tools and highlights the role of exploration in program synthesis. Our method offers a scalable and generalizable framework that can be extended to other hardware description language domains or integrated into existing EDA pipelines for agile hardware design.

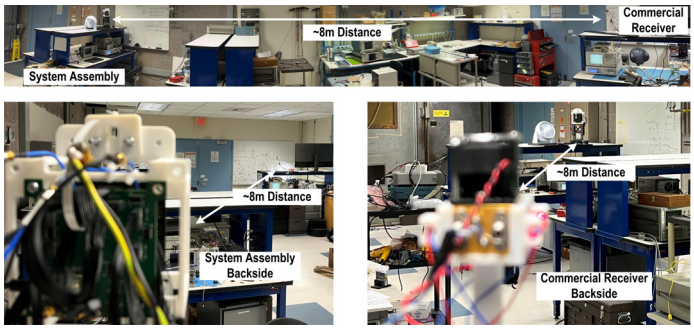
Highly Selective Harmonic-resilient Sub-6G Front Ends

H. Yang, S. Araei, M. Barzgari, N. Reiskarimian

Widely tunable and reconfigurable front ends are heavily in demand to support multi-band communication, such as the Internet of Everything. Recent developments have shown increased radio selectivity to improve linearity performance with respect to interference in the adjacent channel. One of the crucial features required for a widely tunable receiver is the ability to suppress harmonic blockers efficiently and at the earliest opportunity, because powerful harmonic blockers can saturate the front end early in the chain before being rejected, hindering receiver functionality. We have proposed a highly selective sub-6GHz reconfigurable receiver resilient to both harmonic blockers and close-in blockers, achieving a 40dB/decade filtering along with >10 dB 3rd/5th order harmonic rejection directly at the antenna input, a radio frequency (RF) bandwidth of 100 MHz, a +15dBm out-of-band (OOB) third-order intercept point (IIP3) at the frequency twice of the RF bandwidth, and a -3dBm OOB blocker 1-dB compression point (B1dB). The proposed receiver includes a single-stage low-noise amplifier (LNA) crossed by a high-order feedback composed of a gyrator, N-path filters, and mixers, followed by samplers to extract the baseband signal and single-stage baseband amplifiers to realize the I/Q recombination. Harmonic-rejecting mixers previously proposed by our group are added across the LNA, as a built-in structure to generate transmission zeros in the high-order feedback and simultaneously achieve harmonic resonance. The proposed design, to be taped out in 65-nm complementary metal-oxide-semiconductor technology, showcases a novel approach that simultaneously rejects close-in and harmonic blockers.



▲ Figure 1: Architecture of the Proposed RX.



▲ Figure 2: Conversion Gain.

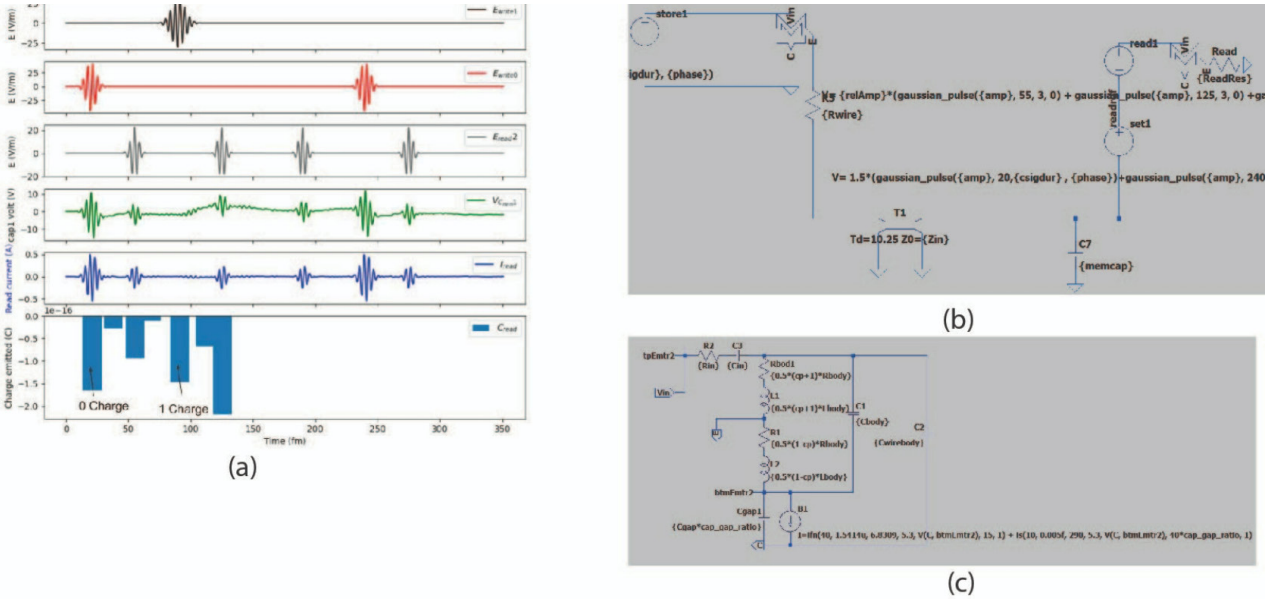
FURTHER READING

- S. Araei, S. Mohin, and N. Reiskarimian, "5.2 0.25-to-4GHz Harmonic-Resilient Receiver with Built-In HR at Antenna and BB Achieving +14/+16.5dBm 3rd/5th IB Harmonic B1dB," *2024 IEEE International Solid-State Circuits Conference (ISSCC).*

Defining and Optimizing Write/Clear Margin for a Nanoantenna-Based Petahertz Electronic Memory Cell

A. Chen, A. R. Bechhofer, J. Simonaitis, F. Ritzowsky, K. K. Berggren, P. D. Keathley
Sponsorship: MIT UROP, NSF, AFOSR

Nanoantennas leverage the unique properties of metallic nanostructures to enable tunneling of low-voltage currents. The compact and integrable design of the nanoantennas makes them a promising approach for realizing ultrafast electronic components. It is possible to construct complex structures using these nanoantennas. One possible structure is a memory cell. In this study, SPICE circuit models will be employed to explore and optimize various parameters, such as input intensity and input signal frequency, in an operational nanoantenna-based memory cell. Performance metrics will be defined and numerical techniques will be applied to refine these parameters, ensuring a robust and clear distinction between the 'write' and 'clear' states for reliable memory cell performance. Modelling the nanoantenna as a circuit component would take significantly less time than traditional physical simulations.



▲Figure 1: The current model of the nanoantenna has indistinguishable write 1/0 as can be seen from (a). A simplified model of the nanoantenna component (b) can be used to create memory cells (c).

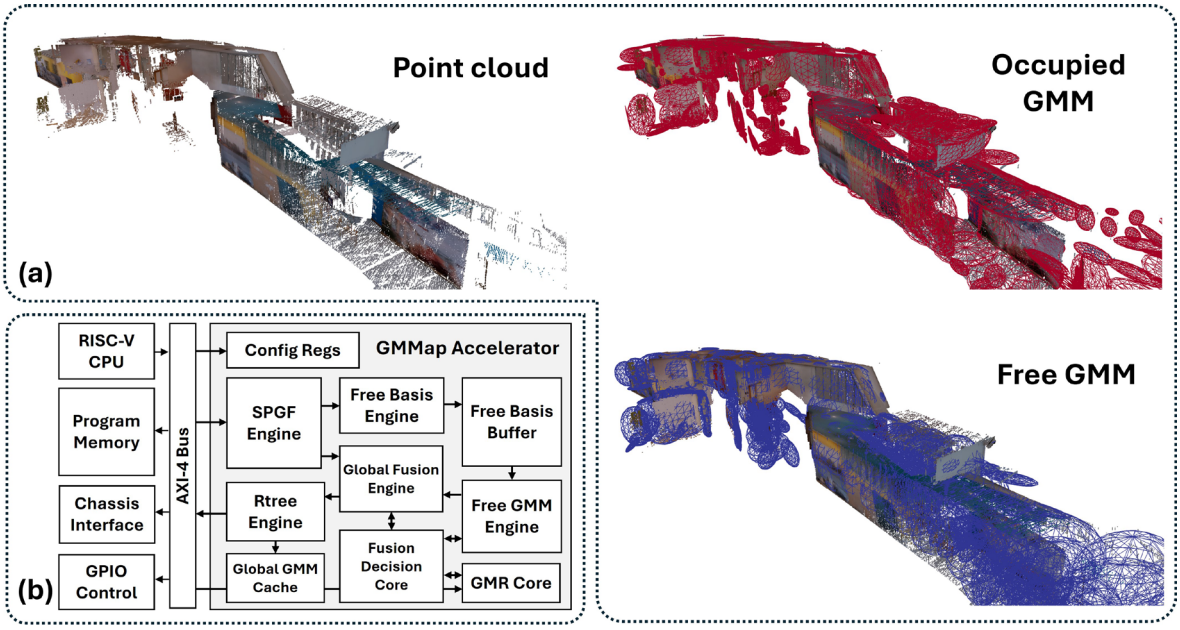
Memory-Efficient Gaussian Mapping on Micro-Robots: Algorithm and Chip

P. Z. X. Li, Z. S. Fu, K. Gupta, S. Karaman, V. Sze
Sponsorship: Amazon, MathWorks Fellowship, National Science Foundation CPS

Constructing a compact map of 3D environments in real-time is essential for enabling autonomy on energy-constrained robots. During construction, the memory usage is not limited to map storage, but also includes overheads for storing the sensor measurements and temporary variables. Prior works reduce the map size while incurring a large memory overhead (MBs) which increases energy consumption and limits throughput.

To reduce memory and energy, we present GMMap, a memory-efficient algorithm that compresses each depth image into Gaussians which are directly fused across multiple images to form a compact 3D map. Using a low-power ARM Cortex-A57 CPU, GMMap can be constructed in real-time with comparable accuracy as prior works while reducing the map size by at least 56%, memory overhead by at least 88%, and energy by at least 69%.

To further reduce energy, we present LEANC, a System-On-Chip with an accelerator for GMMap. On an FPGA, LEANC enables real-time 3D mapping using only milliwatts of power. Thus, LEANC not only enables autonomy on micro-robots but also illustrates the importance of memory-efficient algorithms and specialized hardware design for low-energy applications.



▲ Figure 1: (a) Visualization of the first floor of Stata Center, MIT, and its GMMap representation consisting of Gaussians representing occupied (red) and free (blue) regions. Each Gaussian is visualized as an ellipsoid in 3-D. (b) Hardware architecture of LEANC.

Design and Modeling of High Temperature Gallium Nitride RF Amplifiers

A. Goodnight, J. Niroula, M. Taylor, P. Yadav, T. Palacios
Sponsorship: AFOSR (Grant No. FA9550-22-1-0367), NSF Graduate Research Fellowship

High temperature electronics play an important role in many applications such as space exploration, hypersonic flight, geothermal energy, and automotive industry. However, an RF circuit operating at 500˚C has never been demonstrated, which, in part, is due to a lack of sufficient high temperature device modeling and process design kits (PDKs). Such frameworks are critical in designing robust circuits with high first pass yield as they enable accurate estimations of circuit performance and allow exploration of design space parameters.

Over the last couple years, our group has been developing and exercising a CAD framework of gallium nitride technologies [1]. In this work, we expand upon this modeling effort, exploring the design space of high temperature GaN RF HEMTs. We use the MIT Virtual Source GaN FET (MVSG) model to predict the influence of temperatures from a wide temperature range on the electrical and RF device performance. Based on this model, we are developing a PDK for design optimization within industry-standard CAD tools such as Keysight ADS and Cadence. These device characterizations can then be used to investigate different amplifier topologies with a focus on the output power, efficiency, and linearity operating over a wide temperature range from 25˚C to 500˚C.

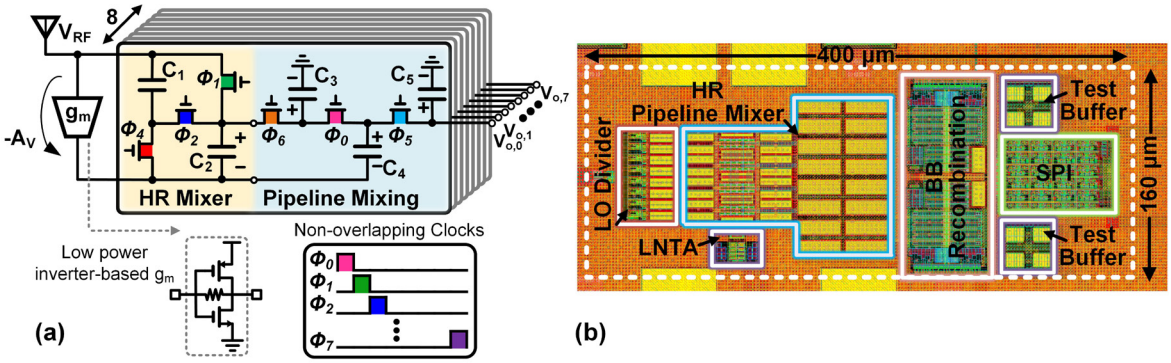
FURTHER READING

- Q. Xie et al., "Towards DTCO in High Temperature GaN-on-Si Technology: Arithmetic Logic Unit at 300˚C and CAD Framework up to 500˚C," *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Kyoto, Japan, 2023, pp. 1-2.

Harmonic-Resilient Low-Power Receiver Architecture with Pipeline Mixing for IoT Applications

S. Araei, M. Barzgari, N. Reiskarimian

The rapid growth of Internet of Things (IoT) applications in wearable devices, smart homes, and healthcare demands receivers (RXs) that integrate seamlessly into this expanding network. Achieving low power consumption, high linearity, and wide tunability is crucial for efficient operation in the congested sub-7GHz frequency band. In this work, we present an RX topology tailored for IoT needs, featuring a harmonic-resilient N-path filter embedded within negative feedback to provide robust blocker protection for all active circuits. By leveraging the Miller effect, the proposed RX achieves a reduced capacitor area and minimal dynamic power consumption. Furthermore, the use of pipeline down-mixing eliminates the need for power-hungry baseband amplifiers, enabling passive gain in a low-noise manner. This RX topology consumes sub-mW power, relies solely on switches and capacitors, and is highly scalable, making it an ideal choice for battery-operated IoT devices.



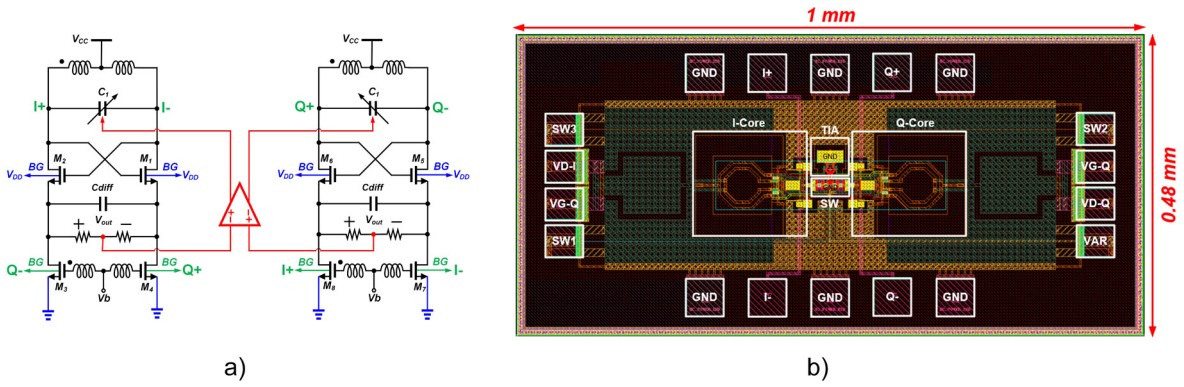
▲ Figure 1: (a) Block diagram of the proposed architecture (b) Chip Layout.

A 28 GHz Coupled PLL-Based CMOS Quadrature Oscillator

M. Barzgari, S. Araei, H. Yang, N. Reiskarimian

One of the most critical challenges in RF and mm-wave transceivers for the further development of wireless communications is generating quadrature signals for receiver down-conversion. Quadrature voltage-controlled oscillators are among the most promising candidates for 5G/6G frequency bands due to their lower power consumption and reduced area. However, achieving a higher Figure-of-Merit (FoM) for the oscillator while maintaining perfect quadrature accuracy remains a significant challenge.

In this work, a coupled PLL-based approach for generating quadrature signals at the local oscillator (LO) is implemented and fabricated using 22nm fully depleted silicon-on-insulator (FDSOI) CMOS technology. In this approach, each oscillator serves as a reference for the other, and the entire structure functions as a type-II phase-locked loop (PLL). This topology relaxes the trade-off between phase noise and quadrature accuracy while achieving a high FoM. It not only improves the image rejection ratio in mm-wave transceivers but also enhances overall system efficiency.

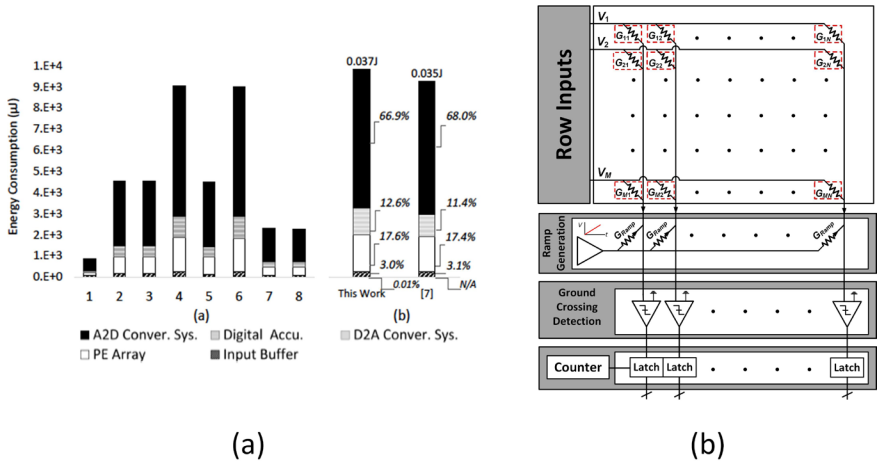


▲ Figure 1: (a) Simplified schematic and (b) Chip Micrograph.

Interface Circuits for Analog In-Memory Computing

M. A. G. Elsheikh, H.-S. Lee
Sponsorship: MIT/MTL Samsung Semiconductor Research Fund

The increased adoption of machine learning (ML) in mobile devices demands high performance at low energy consumption. General purpose hardware is limited in speed and efficiency by the data movement between the memory and the processing unit. Alternatively, compute-in-memory (CIM) accelerators perform the required arithmetic by the integration of memory and processing elements to mitigate this problem. Analog CIM accelerators use voltage and current laws for the parallel generation and addition of partial sums, rapidly and efficiently. However, the read-out circuitry poses a substantial overhead in energy consumption. In this work, we propose single-slope an analog-to-digital converter with non-linear characteristics, which maintains the inference accuracy at a reduced number of bits, to leverage the statistical properties of partial sum outputs to optimize performance and efficiency. This will enable more widespread adoption of ML in more energy-constrained applications.

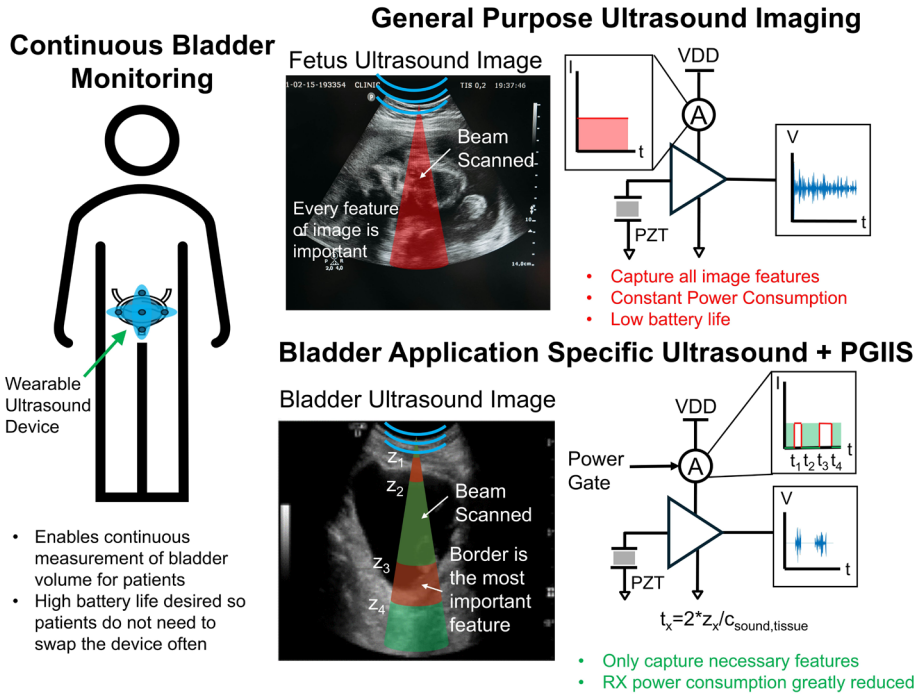


▲ Figure 1: (a) Energy breakdown in analog CIM systems (b) proposed single-slope analog to digital converter readout circuits for analog CIM.

An Analog Front End with Sparse-Image Capturing for Energy-Efficient Bladder Ultrasound Imaging

M. Manohara, S. Schoen, D. U. Yildirim, M. Perrot, P. Garcha, A. Samir, A. Bahai, A. P. Chandrakasan
Sponsorship: Texas Instruments

Continuous bladder monitoring is important for patients unable to excrete their urine. One method for bladder volume calculation is capturing ultrasound images and utilizing image segmentation algorithms for bladder volume estimation. These algorithms typically search for the boundary of the bladder and ignore other regions of the image. In this work, we design an analog front end (AFE) utilizing an algorithm called Power Gating for Intra-Image Sparsity (PGIIS). This AFE simultaneously generates beamformed TX pulses and amplifies the RX signals with good signal-to-noise ratio. For each TX event, the PGIIS algorithm identifies time intervals corresponding to the bladder boundary and power-gates the RX amplifiers outside of these time intervals, generating a sparse image. The AFE was tested on a custom phantom with a bladder proxy. After calibration, the PGIIS algorithm demonstrated 75% power savings in RX amplification, enabling low-power imaging suitable for wearable ultrasound systems.



▲Figure 1: Motivation for continuous bladder monitoring and motivation behind Power-Gating for Intra Image Sparsity algorithm

A 232-to-260GHz CMOS Amplifier-Multiplier Chain with a Matching-Sheet-Assisted Radiation Package and 11.1dBm Total Radiated Power

J. Wang, R. Han
Sponsorship: Intel University Shuttle Program, Jet Propulsion Laboratory Strategic University Research Partnerships Program

Terahertz (THz) signal sources and radiators are essential for a variety of future applications, such as high-resolution radar imaging, molecular spectroscopy, and clocks, as well as high-speed or miniature-platform communications. For over a decade, the growing interest in complementary metal-oxide semiconductor (CMOS) compact THz radiation sources has been driven by their small form factor and integration with other analog/digital systems. However, the total radiated power of prior CMOS THz sources was still only several mW, not only due to the limited fmax and breakdown voltages of the CMOS transistors, but also due to the inefficient on-chip radiation approaches. Due to the large dielectric constant contrast between the silicon and air, the radiated waves undergo strong reflection at the silicon-air interface, and with a small outward angle, total internal reflection occurs. To alleviate this problem, cm-sized, high-resistivity silicon lenses affixed to the chip back have been used. However, they dramatically increase the cost and size of the overall assembly. In this work, we introduce a CMOS THz amplifier-multiplier chain array generating radiation between 232 and 260GHz. Instead of using a silicon lens, a patterned dielectric matching sheet is applied onto the flipped-chip back, which enhances the wave coupling from silicon to air and enables low-cost and planar packages. That, in conjunction with broadband gain-peaking power amplifiers built with a high-power radio frequency FinFET transistor technology in the CMOS process, enables a measured peak total radiated power of 11.1dBm.

Efficientvit-SAM: Accelerated Segment Anything Model without Performance Loss

Z. Zhang, H. Cai, S. Han
Sponsorship: National Science Foundation, MIT-IBM Watson AI Lab, MIT AI Hardware Program, Amazon, Samsung, Hyundai

Segment Anything Model (SAM) has gained widespread recognition as a milestone in the field of computer vision, showcasing its exceptional performance and generalization in image segmentation. SAM defines image segmentation as a promptable task, that aims to generate a valid segmentation mask given any segmentation prompt. SAM has shown its high versatility in a wide range of downstream applications, including image in-painting, object tracking, and 3D generation. Nevertheless, SAM imposes significant computational costs, leading to high latency that restricts its practicality in time-sensitive scenarios and edge devices. In particular, SAM's main computation bottleneck is its image encoder, which requires 2973 GMACs per image at the inference time. To accelerate SAM, numerous efforts (MobileSAM, EdgeSAM, EfficientSAM) have been made to replace SAM's image encoder with lightweight models. While these methods reduce the computation cost, they all suffer from significant performance drops. We introduce EfficientViT-SAM to address this limitation by leveraging EfficientViT to replace SAM's image encoder. Meanwhile, we retain the lightweight prompt encoder and mask decoder architecture from SAM. Our training process consists of two phases. First, we train the image encoder of EfficientViT-SAM using SAM's image encoder as the teacher. Second, we train EfficientViT-SAM end-to-end on the whole SA-1B dataset. We thoroughly evaluate EfficientViT-SAM on a series of zero-shot benchmarks. EfficientViT-SAM provides a significant performance/efficiency boost over all prior SAM models. In particular, on the COCO dataset, EfficientViT-SAM achieves 48.9× higher throughput on A100 GPU without mAP drop compared with SAM-ViT-H. We believe that EfficientViT-SAM enables the segment anything technique to be widely applied in time-sensitive scenarios, such as autonomous driving and robotic manipulation, while also making it easily deployable on edge devices like mobile phones.