Abstract—This paper provides an overview of the SONIC Center’s research vision, outcomes, and future prospects. SONIC’s research focused on establishing a Shannon-inspired statistical foundation for computing platforms realized on nanoscale process technologies (both CMOS and beyond CMOS) to meet the needs of emerging data-centric learning-based applications. SONIC’s outcomes include architectural concepts of deep in-memory and deep in-sensor computing, the use of stochastic nanofunctions coupled with statistical models of computing to realize systems in beyond CMOS, and fundamental limits on energy-latency-accuracy of statistical computing systems and design principles to approach those limits. Numerous laboratory prototypes have validated these outcomes thereby demonstrating the hallmark of SONIC’s research—the systems-to-devices journey. These outcomes also set the stage for developing design methodologies to realize complex autonomous cognitive agents that learns to observe, infer, reason, create, explain, and act by engaging with its immediate environment, while operating at the limits of energy, latency, and accuracy trade-off by leveraging statistical computing design principles.

Keywords—Shannon-inspired computing, nanoscale devices, in-memory computing, in-sensor computing, nanofunctions.

I. INTRODUCTION

SONIC was established in 2013 to take a fresh look at the design of computing platforms in deeply scaled CMOS and beyond CMOS nanoscale fabrics in order to address the unique needs of emerging applications. These applications (Fig. 1) tend to be learning-based, requiring the sensing, storage, transportation, and processing of massive volumes of data. The value of today’s computing platforms is judged primarily in terms of their ability to extract information and actionable intelligence from data rather than their raw processing capability as has been the case traditionally.

While today’s systems rely on the cloud in order to extract information from data captured by edge devices, SONIC chose to focus on embedding intelligence directly into the edge or the swarm. Edge devices that can process data locally avoid the high energy and latency costs involved in data transmission to and from the cloud, and alleviate concerns related to privacy and security.

Indeed, today’s machines have begun to approach and exceed human performance in many complex inference tasks, e.g., AlexNet [1], ResNet [2] which outperformed humans in recognition tasks, and Google DeepMind’s AlphaGo [3], which beat human champion Sedol Lee in the ancient game of Go.

Despite these well-acclaimed successes, machines have a lot of catching up to do when energy costs are accounted for. It is estimated that the computing platforms employed to realize these wins (1202 CPUs and 176 GPUs for AlphaGo) consume about four-orders-of-magnitude higher energy as compared to the human brain. Therefore, the fundamental question facing us today is: How to design intelligent machines that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain?

SONIC’s research was motivated by the realization that present day computing platforms based on the von Neumann architecture will not be able to address the applications challenge shown in Fig. 1. The reasons are multifold: (a) the unique nature of modern-day cognitive applications/tasks, which emphasize exploration-, learning-, and association-based computations based on statistical representations and processing of massive volumes of data vs. conventional data processing; (b) the traditional von Neumann architecture [4], its associated Turing model of computation [5], with its requirement of a deterministic, Boolean-switch-based, semiconductor process...
technology that has become hard to realize because of (c) the realities of semiconductor feature-size scaling (Fig. 2(a)), e.g., diminishing energy-delay benefits via CMOS scaling, and increased stochasticity (variations and failures) (Fig. 2(b)), which squarely oppose the deterministic requirements of today’s von Neumann-style computing. While many beyond CMOS devices have been invented, none surpass the silicon MOSFET in terms of the metrics for a deterministic switch. Additionally, the data-centric nature of emerging applications places an undue burden on the memory-processor interface aggravating the well-known ‘memory wall’ problem (Fig. 2(c)) in von Neumann architectures.

SONIC’s vision (Fig. 3) was based on the understanding that information and nanoscale stochasticity (noise) needs to be, and can be, addressed within a common statistical framework in order to be able to design information processing systems in nanoscale technologies that can approach fundamental limits on energy efficiency, latency and accuracy. The reason being that both information and noise are statistical quantities and therefore are in some sense ‘compatible’, and because the feasibility and benefits of such a unified view have already been demonstrated in the area of communications through the work of Claude Shannon [9] (Fig. 4). SONIC’s vision stands in contrast with traditional von Neumann computing, where nanoscale noise and computational errors are viewed as problems needing to be heavily suppressed to present a deterministic fabric for computing platforms. This suppression is the root cause of the unfavorable energy-latency-accuracy trade-offs inherent in von Neumann computing.

Shannon’s 1948 paper established information as a statistical quantity and laid out a complete theory of communications (Fig. 4). A key contribution was to introduce a simple yet pragmatic abstraction of the problem of communicating information over a noisy channel. This problem abstraction enabled Shannon to define channel capacity as a function of noise statistics, and obtain a surprising result that reliable, i.e., probability of error approaching zero, communication can be achieved if the information transmission rate is less than the channel capacity. Prior to this work, it was commonly assumed that the only way...
to approach zero error probability was via a large (potentially infinite) number of repeated transmission of the same information bit (repetition codes) which forced the information rate (rate at which new bits are transmitted) also to approach zero. Shannon also showed that error control codes exist that approach channel capacity. Indeed, the discovery of turbo codes [10] and the re-discovery of low-density parity-check (LDPC) codes [11], [12] in the 1990s have all but eliminated the gap to Shannon capacity for point-to-point links. In this manner, Shannon theory completely revolutionized the area of communications.

SONIC’s research agenda was designed to similarly revolutionize the area of computing, particularly for learning-based systems realized on nanoscale fabrics. The emergence of data-centric learning-based applications in conjunction with ‘noise’ in the nanoscale makes SONIC’s research highly relevant at the present time and for the foreseeable future.

SONIC’s research leveraged ideas from communications (Shannon-inspired) and brain (neuro-inspired/principled) at the system level (Fig. 5(a)), translating the powerful insights offered by these domains into realizable architectures and circuits, followed by experimental demonstrations (design, fabrication, and test) of system prototypes in both scaled CMOS and beyond CMOS fabrics. The principle underlying the SONIC approach was to employ statistical models of computation to compose reliable and energy-efficient systems using stochastic components. This principle was shared with what might be arguably two existing information processing systems that operate at the limits of energy-latency-accuracy: a communication receiver operating over a stochastic channel and the human brain that is composed of stochastically behaving neurons and synapses. SONIC’s approach, referred to as statistical information processing or statistical computing, employed information-based metrics such as mutual information for system design and evaluation and to fully explore the trade-offs between energy-efficiency, latency, and robustness (Fig. 5(b)). An important goal in SONIC’s research endeavors was to develop principles of statistical information processing and demonstrate their effectiveness in reducing energy, enhancing robustness and increasing functional density, via experimental realizations.

The SONIC’s systems-driven and top-down approach enabled a principled embrace of component stochasticity. To develop a Shannon-inspired foundation for computing, SONIC brought together a diverse group of researchers from neuroscience, information and coding theory, machine learning, microarchitecture, integrated circuit design, and devices. SONIC’s vertically (from applications/systems through architectures and circuits down to nanoscale devices) integrated research organization was critical for developing a statistical foundation for computing in the nanoscale era, and now has become a de facto standard for all major semiconductor research programs including Joint University Microelectronics Program (JUMP), Nanoelectronic Computing REsearch (nCore), and DARPA’s Electronics Resurgence Initiative (ERI). This research organization is described next.

SONIC has three system-driven research themes (Themes 1, 2, and 3) firmly anchored in a circuits/devices theme (Theme 4). Theme 1 focused on developing Shannon-inspired design principles and fundamental limits for realizing statistical information processing systems using stochastic components. Theme 1 outcomes guided the practical realizations of systems in the rest of the SONIC themes. Theme 2 employed Theme 1 provided design principles to realize analog mixed-signal systems and circuits for inference applications. Theme 3 developed neuro-inspired approaches for the design of statistical information processing systems on nanoscale fabrics encompassing computation, communication, and both discrete and mixed-signal representations. Thus, all three systems themes drove towards the development of statistical foundations for information processing. Theme 4 was focused on experimental demonstrations of Shannon-inspired nanofunctions and systems in both CMOS and beyond CMOS fabrics. Theme 4 demonstrated that the Shannon/neuro-inspired
system design approaches from Themes 1, 2, and 3, enabled deeply-scaled device/circuit fabrics to be deployed effectively in a manner not feasible today under the von Neumann model.

In the final year (YR5) of its tenure, SONIC’s Shannon and brain-inspired approach to computing was consolidated into a set of four broad outcomes – deep in-memory computing, deep in-sensor computing, systems-in-beyond CMOS, and fundamental limits/design principles.

III. SONIC OUTCOMES

A. Deep In-Memory Computing

SONIC’s deep in-memory architecture (DIMA) [13], [14] directly attacks the memory wall problem intrinsic to the von Neumann architecture. DIMA (Fig. 7) does so by reading functions of data stored in multiple rows, and processing them via analog computations embedded in a pitch-matched manner in close proximity to SRAM bit-cell array (BCA). DIMA has four sequential processing stages: 1) multi-row functional read (FR): fetches a function of bits stored in multiple rows per read cycle from each column, 2) BL processing (BLP): performs word-level arithmetic operations, via column pitch-matched analog processors in parallel, 3) cross BL processing (CBLP): aggregates multiple BLP outputs via charge-sharing to obtain a scalar output, 4) ADC and residual digital: generates the digital output (decision) from the previous analog computations.

By substantially eliminating an explicit processor-memory interface, DIMA simultaneously enhances energy, delay, and latency. In doing so, it trades-off the circuit signal-to-noise ratio (SNR) making it an ideal fabric for demonstrating Shannon-inspired concepts. Additionally, it turns out that DIMA is well-matched to the data-flow of machine learning and cognitive applications which are based heavily on projection-based computations. However, DIMA has multiple design challenges primarily due the need for analog computations under severe area constraints imposed by the BCA pitch. The SONIC research team successfully overcame these design challenges and demonstrated several prototypes. SONIC’s DIMA integrated circuit (IC) prototypes [15]–[19] showed energy savings ranging from 10× to 113×; throughput gains of up to 6× and energy-delay product (EDP) reduction ranging from 52× to 162× over custom von Neumann architectures. These gains were obtained prior to the application of design optimizations in DIMA, i.e., these designs have room for improvement.

B. Deep In-Sensor Computing

Deep in-sensor computing aims to overcome the energy, scalability, and fidelity limitations that arise by separating the sensing functionality from feature-extraction and inference functions. This is because the resulting large volume data transfer from sensor to processor dominates the energy and latency costs. Deep in-sensor computing extracts information by embedding computing functions into the sensory substrate by realizing those via devices native to those substrates and which tend to be slow and unreliable.

SONIC researchers developed a range of technologies for diverse, distributed, and form-fitting sensors. For example, SONIC’s system-level principles and demonstrations based on large-area electronics for image sensing using boosted embedded weak classifiers and embedded random projection-based compression, and electroencephalogram (EEG) acquisition and processing via compressed-domain transformations [20], demonstration of a photoplethysmographic (PPG) acquisition via a sensing-and-processing epidermal-electronics patch [21], design and test of a nanopower chip designed for potentiometric ion sensing consumes only 5.5 nW for full functionality, which is by far the lowest power reported by any wireless ion sensing system [22], design and test of a 12.5 Gb/s (over 5 m) 130 GHz link achieving an energy efficiency of 7.8 pJ/bit and a link efficiency of 1.55 pJ/bit/m, which is 40× improvement over the current state-of-the-art [23]. In addition, the team demonstrated an ultrasound front-end that compresses the image signals within the sensor interface, leading to a 30-50× data reduction [28].

C. Systems in-beyond CMOS

SONIC significantly accelerated the deployment of beyond CMOS and CMOS nanoscale fabrics via the concept of nanofunctions. Nanofunctions are pervasively employed computational kernels in the inference application space that are sufficiently simple in structure (can be realized with a <10-20 devices) and yet functionally complex to enable useful systems research by enabling the composition of complex inference systems. Nanofunctions bridge the devices-to-systems gap in the SONIC way. SONIC researchers successfully developed the following set of nanofunctions: graphene dot-product [24], RRAM nano-oscillators for Oscillatory Neural Networks (ONNs) [25], graphene mux-based logic [26], Magnetic Tunnel Junction (MTJ)-based random number generators [27], which form a key component in deep neural networks, hyper-dimensional (HD) computing, and spin channel networks for inference applications. SONIC researchers also created

![Fig. 7. The deep in-memory architecture (DIMA).](image-url)
behavioral models of these nanofunctions and shared them with systems researchers who developed methods for designing statistical information processing systems using nanofunctions.

D. Fundamental Limits and Design Principles

SONIC explored the fundamental limits and design principles for Shannon-inspired statistical information processing. SONIC researchers developed systems models of stochastic nanofunctions in CMOS and beyond CMOS, obtained bounds and limits on information processing capacity, and developed capacity achieving design principles. SONIC researchers developed a number of principles for statistical information processing and demonstrated their broad applicability for guiding the design of systems. SONIC’s design principles took a system-level approach to statistical information processing systems, focusing on application-relevant metrics that expose statistical dependencies in sensory inputs data.

Specifically, SONIC researchers (1) leveraged ML algorithmic approaches that can exploit the relationships for efficiency and robustness to nanoscale upsets and non-idealities; (2) developed architectural principles that provide robust building blocks for systematizing design of reliable information kernels from unreliable nanofabrics; and (3) developed fundamental limits of performance, scaling laws, and performance predictors to provide insight into the “adjacent possible” leveraging today’s nanofunctional blocks.

IV. Conclusions & Future Prospects

SONIC’s Shannon-inspired statistical computing research is a complete and principled approach to non von Neumann computing, encompassing principles of statistical information processing at the system level, mapped to architectures and circuits, followed up by experimental demonstrations of system prototypes in both scaled CMOS and beyond CMOS fabrics. SONIC’s research has leveraged information-based metrics for design and evaluation, which enabled the exploration of trade-offs between energy-efficiency, robustness, and application-level accuracy.

SONIC’s research outcomes already had the following immediate impact:

- CMOS deep in-memory/in-sensor inference architectures are ready for deployment today, and are already called for in major semiconductor research programs.
- Statistical models of computation enabling systems in beyond CMOS have provided an alternative path for device scaling, one in which the device error rate is coupled with statistical approaches for error compensation.
- Fundamental limits, metrics, and design principles have provided a compelling framework for the principled design of AI systems in silicon and beyond.

SONIC’s research outcomes have set the stage for developing design methodologies to compose complex autonomous cognitive agents, which are sensor-rich autonomous entities, highly constrained in volume, energy, and computational resources, e.g., an unmanned aerial vehicle or untethered robot that learns to observe, infer, reason, create, explain, and act by engaging with its immediate environment, including meaningful interactions with humans. In this way, SONIC research will contribute to a society which today stands at the cusp of a Machine Revolution that is expected to lead to a future in which cognitive machines—i.e., machines that think like humans—will collaborate with humans and augment their capabilities. Furthermore, SONIC’s multidisciplinary and systems-to-devices research agenda, in addition to establishing a statistical foundation for computing in the nanoscale era, has also set a framework for conducting advanced research in semiconductor-based computing systems for the foreseeable future.

SONIC’s impact also extends into the future of semiconductor research. Its vertically integrated research (systems research grounded in experimental prototypes) has become the standard for conducting advanced semiconductor research. JUMP sought to establish four (out of six) vertically integrated centers. Key SONIC concepts such as Shannon-inspired computing, algorithmic noise-tolerance, and on-chip architectures were called out as key areas for future research. DARPA’s ERI similarly has a vertically integrated structure geared to translate outcomes from STARNet and JUMP. Furthermore, nCORE also embraces the SONIC philosophy of providing systems pull on device research.

REFERENCES


