IoT² — the Internet of Tiny Things: Realizing mm-Scale Sensors through 3D Die Stacking

Sechang Oh, Minchang Cho, Xiao Wu, Yejoong Kim, Li-Xuan Chuo, Wootaek Lim, Pat Pannuto,
Suyoung Bang, Kaiyuan Yang, Hun-Seok Kim, Dennis Sylvester, David Blaauw
University of Michigan, Ann Arbor, MI, USA

Abstract—**The Internet of Things (IoT) is a rapidly evolving application space. One of the fascinating new fields in IoT research is mm-scale sensors, which make up the Internet of Tiny Things (IoT²). With their miniature size, these systems are poised to open up a myriad of new application domains. Enabled by the unique characteristics of cyber-physical systems and recent advances in low-power design and bare-die 3D chip stacking, mm-scale sensors are rapidly becoming a reality. In this paper, we will survey the challenges and solutions to 3D-stacked mm-scale design, highlighting low-power circuit issues ranging from low-power SRAM and miniature neural network accelerators to radio communication protocols and analog interfaces. We will discuss system-level challenges and illustrate several complete systems and their merging application spaces.**

I. INTRODUCTION

We are living in a new era of the Internet of Things (IoT), where environmental and human activity data are sensed by everyday objects and connected to a network. One of the unique characteristics of IoT devices is that they can interface with "things" instead of "people" [1]. This feature eliminates human interfaces that require large physical dimensions (e.g., touch screens), shrinking IoT device size and enabling mm-scale sensors, which make up the Internet of Tiny Things (IoT²). With recent advances in low-power circuits and 3D die stacking technology, the IoT² is becoming a reality and opening up new applications in a variety of areas such as healthcare, surveillance, micro-robots, and industrial and environmental monitoring (Fig. 1).

However, such small form factors pose challenges to IoT² system design. An IoT² system must address power challenges because battery density does not scale well into such a small size. The energy budget of mm-scale batteries is limited to a couple µWh (thin-film battery) or mWh (coin-cell battery). Moreover, the IR voltage drop experienced due to their high battery resistance (1s to 10s kΩ) limits the maximum instantaneous power. Therefore, the system power must be extremely low in both active and standby modes. In addition, the physical size and number of discrete components should be minimized to achieve a compact millimeter-scale system. Furthermore, a modular design with seamless inter-layer communication is desirable because it enables inter-operability of diverse components, enabling a variety of applications at a
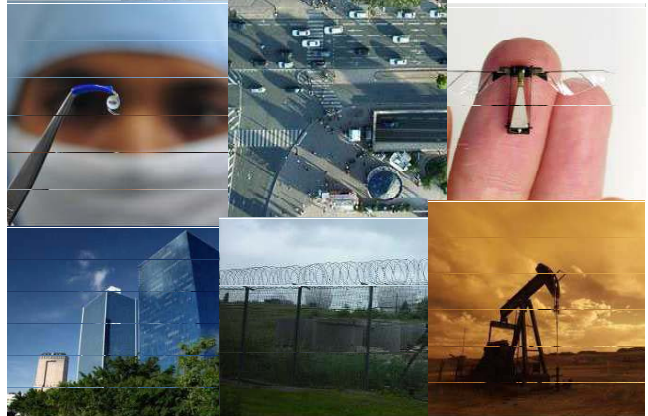

Fig. 1. IoT² application spaces, including healthcare, surveillance, micro robots, and industrial and environmental monitoring.

low development cost. By grouping the integrated circuits into multiple chips, we can integrate heterogeneous chips fabricated in different technologies based on their purpose (e.g., flash or image) or optimal performance and cost (e.g., low leakage or high speed).

This paper will discuss all of these challenges and demonstrate modular IoT² system implementation using 3D die stacking. The remainder of this paper is organized as follows. Section II discusses circuit techniques to address IoT² challenges. Section III presents system-level implementation of IoT² devices. Finally, Section IV concludes the paper.

II. TECHNIQUES TOWARD THE IoT²

A. Ultra-Low Standby Power 7T SRAM

IoT² devices are often heavily duty-cycled, and therefore low standby power is critical to achieve overall low-power operation. One circuit block that draws large standby current is SRAM. A regular 6-transistor (6T) SRAM bit-cell is compact but requires high VDD to satisfy all trade-offs, including noise margin, write margin, and read current, resulting in high standby power consumption. On the other hand, a 10T SRAM bit-cell [2] achieves extremely low leakage (fW/bit) at the expense of the bit-cell area.
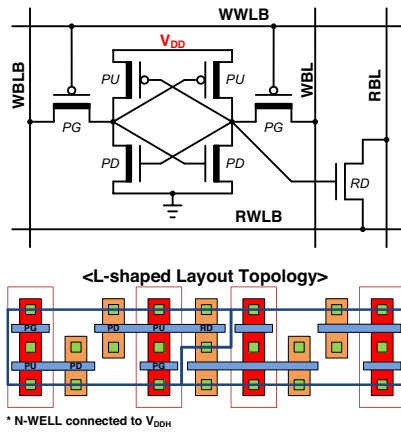
Fig. 2. 7T SRAM bit-cell design with L-shaped layout topology.

A 7T design adds only one decoupled read interface to the original 6T bit-cell, as shown in Fig. 2, allowing low VDD operation similar to that obtained with the 10T SRAM. A separate high VDD (VDD$_H$) is used to bias NWELL and boost and precharge the bit-cell. By reducing the number of devices and adopting an L-shaped topology, the bit-cell is 50% and 18% smaller than the 10T and 8T bit-cells, respectively. An 8 kB SRAM macro was fabricated in 180-nm CMOS technology. It achieved 3.35 fW/bit leakage power, 0.39 pJ/bit access power, and 320 mV minimum operating VDD.

B. Miniature Deep Learning Accelerator

The demand for deep learning on mobile devices is increasing. It provides intelligence at the device level instantly without relying on online servers and without energy overhead from data communications. In a deep learning accelerator, memory access is very frequent and statically scheduled. Based on this observation, a new non-uniform memory architecture (NUMA) to balance the trade-off between memory area and access energy is proposed [3].

Fig. 3 shows the memory design trade-off between bit density and access energy across bank size. A small bank SRAM consumes low access energy per bit at the cost of low memory density and vice versa. The statistical nature of deep learning memory usage is used to determine the number of NUMA hierarchical levels and the size of each hierarchy. In this architecture, NUMA is optimized to implement a fully connected layer that performs matrix-vector multiplication, offset addition, and a non-linear activation function. The input vector of deep learning is mapped to the nearest memory for frequent usage, while infrequently accessed bits, such as weight value, are loaded from dense high hierarchy. This strategy results in >40% energy saving with only 2% of the area overhead compared to a typical uniform memory architecture in simulation. The prototype design fabricated in 40-nm CMOS technology showed 60% less power consumption in L1 than in L4. It achieved 374 GOPS/W peak efficiency and 288 µW power consumption at 0.65 V supply and 3.9 MHz clock [3].
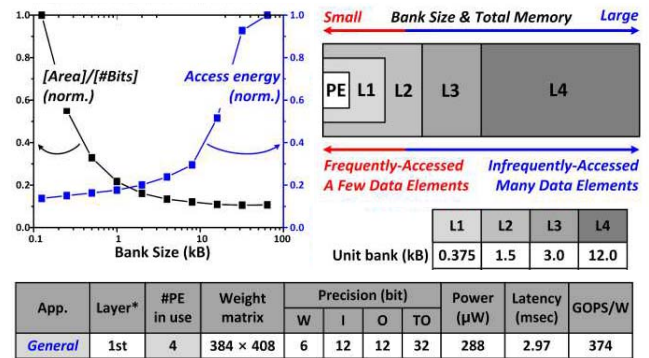

Fig. 3. Memory area and access energy trade-off (top left). NUMA memory for a processing element (PE) (top right). Measured neural network performance (bottom).
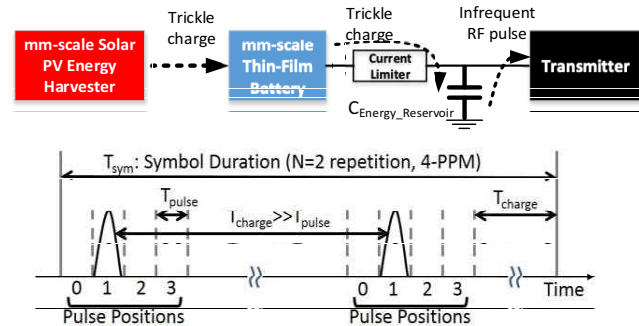

Fig. 4. Millimeter-scale radio transmitter with an energy reservoir capacitor, battery and harvester (top). Timing diagram of the proposed pulse position modulation with an example of 2-repetition of 4-PPM (bottom).

C. Millimeter-scale Radio Communication

Wireless communication is a dominating energy consumption factor in IoT[2] systems under stringent size constraints. Thus, optimizing radio communication energy is critical to extending the lifespan of IoT[2] devices. However, the ultra-small size constraints impose substantial challenges to enabling long-range (>10 m) wireless communication. A conventional millimeter-size radio typically operates at >>1 GHz, and hence its performance is limited due to low antenna efficiency (<1%). Power hungry PA, LNA, and PLL also make this millimeter radio impractical in IoT[2] systems, which can only supply up to ~100 µA instantaneous current because of their high battery internal resistance.

To address these challenges, a low-power crystal-less radio system based on pulse-position modulation was proposed [4], [5] as shown in Fig. 4. The battery is charged by harvested energy. To avoid abrupt voltage drop, radio transmit pulse energy is drawn from an energy reservoir capacitor (1 µF), and the capacitor is charged by the battery through a current limiter. This scheme requires a reservoir capacitor charging time between adjacent pulses and uses M-ary pulse position modulation (PPM) to increase energy efficiency. Since the pulse width (T$_{pulse}$) is much shorter than the charging time (T$_{charge}$), resolution of M is achieved with almost no time overhead in a reasonable M range (e.g., ≤ 64). Its symbol error rate is limited by the energy in the reservoir capacitor. By N
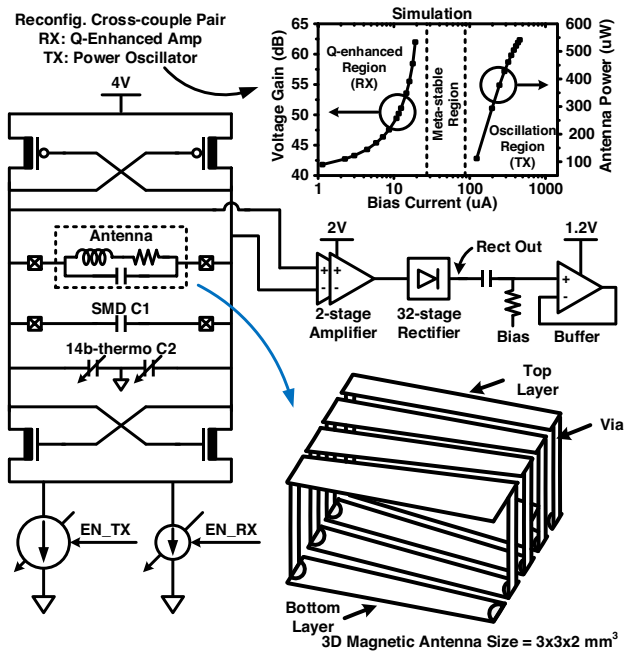
Fig. 5. Radio transceiver circuits using a re-configurable cross-coupled pair with a 3×3×2 mm³ 3D magnetic antenna.

Fig. 6. Optical receiver block diagram.

repetition of a same symbol, the communication distance is enhanced at the expense of reduced data throughput.

Fig. 5 shows the proposed cross-coupled pair transceiver circuit. In the transmission mode, it functions as a power oscillator with a bias current >100 μA. The radio uses 902-928 MHz band for optimal Non-Line-of-Sight (NLOS) path loss. The cross-coupled driver resonates a 3×3×2 mm³ magnetic dipole antenna with a high-Q off-chip capacitor and delivers power to the antenna directly. This design offers the advantages of inherent frequency generation, inherent antenna matching, and high (>30%) transmitting efficiency by eliminating the power-hungry PA and PLL. It also removes a bulky external crystal. However, this approach results in slight carrier frequency drift, and the gateway needs to compensate it periodically. In this usage scenario, the sensor node initiates communication by sending a header, and the gateway estimates accurate timing and frequency and synchronizes with the IoT² node. It moves the complexity and power from the IoT² node to the gateway with these advanced signal processing capabilities. This is reasonable since the gateway energy efficiency is not our main concern. In the receive mode, the cross-coupled pair is biased in the non-oscillating region (<20 μA) as opposed to the oscillation region (>100 μA), increasing the Q of the resonant tank to 300 and resulting in a 49-dB voltage gain. This approach significantly lowers the required gain in later stages and replaces a conventional power-hungry LNA and off-chip saw filters. After the cross-coupled amplifier, the signal is further amplified and demodulated by a 32-stage passive rectifier.
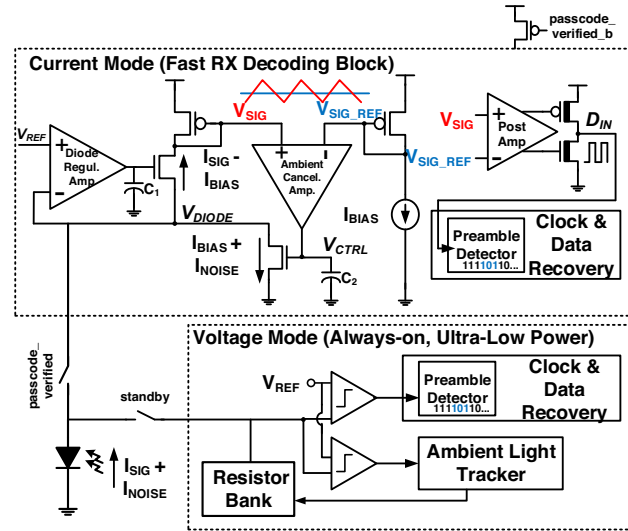
D. Ultra-Low-Power Optical Communication

IoT² devices are in idle mode much of the time and often lack an accurate clock source. Thus, they need an ultra-low-power wakeup receiver that is always on to wake up when requested and synchronize the time with a base station. Fully encapsulated IoT² devices for deployment have no physical access pins. Initial direct SRAM programming is required since there is often no read-only memory and no firmware. For these purposes, an optical receiver offers a significant advantage in terms of area and power when the communication distance is short (~1 m). Even with a parasitic diode, an optical receiver operates with an area of 100×100 μm² and consumes 380 pW standby power and 110 pJ/bit energy efficiency [6]. On the other hand, a conventional RF-based receiver is limited to 10s of μW standby power and a couple nJ/bit energy efficiency due to the use of oscillators and amplifiers and requires a much larger antenna.

Fig. 6 shows a block diagram of the proposed receiver in [6], which employs dual modes. In the voltage mode, the current mode circuit is power-gated, an ultra-low-power voltage comparator directly compares the diode voltage with the reference voltage, and a digital logic verifies a valid passcode by matching the 16-bit on-off keying Manchester coded signal. It achieves 380 pW standby power by eliminating power-hungry analog blocks such as amplifiers. However, the maximum bit rate of the voltage mode is limited by the capacitance-to-current ratio of the photovoltaic cell, which is physically constrained. To enhance the speed, the current mode uses operational transconductance amplifiers to regulate V_{DIODE} voltage regardless of light intensity, and the regulation loop determines the maximum speed, which is proportional to the amplifier currents. The loop also cancels up to 100 klx ambient light by high-pass filtering. The remaining fast switching signal is amplified to rail-to-rail voltage by a comparator and decoded by the clock-data-recovery logic. It achieves 110 pJ/bit energy efficiency at 250 kbps.
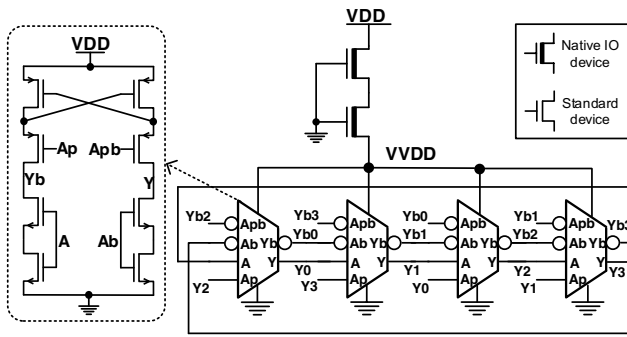
Fig. 7. Temperature sensor core using rotated diode connected PMOS loads and native NMOS source.

E. *Extremely Small Temperature Sensor*

Thermal sensing is an important feature for IoT[2] systems and used in applications such as chip performance regulation, thermal compensation for other sensors, factory and home automation, and body temperature monitoring for implantable healthcare. On-chip temperature sensor challenges include reducing energy consumption, area, and supply sensitivity while maintaining accuracy.

Fig. 7 shows a proposed temperature-sensing core using exponential dependency of sub-threshold current [7]. The sub-threshold oscillator frequency is expressed as

$$T \cdot \ln(Freq) = aT + b \qquad (1)$$

where a and b are coefficients from 2-point calibration.

To address the issue of high sensitivity to supply or bias voltage, a native NMOS is added as a header device for the ring oscillator. The temperature-sensing ring oscillator can be understood as a rotating diode-connected PMOS current load, and the fundamental equation is similar to the ultra-low-power voltage references proposed in [8], [9]. The PMOS switches of the delay cell are controlled by the preceding stages to avoid voltage fluctuation possibly derived from short current on the oscillator. Because of its simple structure, the temperature sensor occupies only 8,900μm^2 in 180-nm CMOS technology, achieving 70 mK resolution and 0.6 nJ/conversion.

F. *MBus: Low-Power Bus for Modular IoT[2] Design*

Conventional interconnect buses employ power-hungry pull-up in an open drain configuration or area-expensive chip select lines. Both render them fundamentally impractical for use in IoT[2] design with stringent budgets for power and area. To address these unique constraints of IoT[2] devices, a low-power bus, referred to as MBus, is proposed in [10], [11].

Fig. 8 shows the MBus block diagram. The MBus nodes are arranged in two rings, clock and data, driving the same line high and low. With two rings, the number of nodes is easily scalable, and any node is able to initiate a message transmission to any other node at any time without conflicts. This design achieves ultra-low power by eliminating a need for local oscillators in member nodes and by using power-gating control. The only special node, the mediator, is responsible for generating the clock and resolving arbitration. The clock and data signals pass through only a minimal amount of combinational logic from
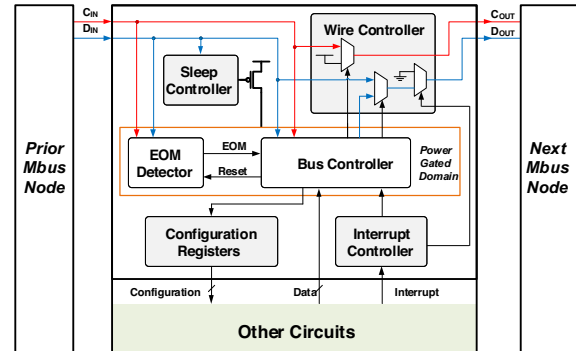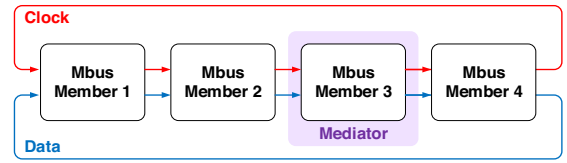

Fig. 8. Block diagram of MBus, which forms two rings. Every member node has four I/O pins and only mediator has a local oscillator.

one node to the next, and the other transistors are power-gated. MBus wake up is hierarchical for maximum power efficiency. When an MBus event occurs, all bus controllers in the ring are active and match the message address to determine if it is destined for the node. Once the message address is matched, the bus controller wakes up the rest of the node. The MBus block was implemented in 180-nm CMOS technology with an active area of 37,200 μm^2. The measured energy consumptions for mediator sending, member receiving, and member forwarding modes at 620 kHz were 27, 23, and 18 pJ/bit, respectively, which are two orders of magnitude better than that obtained with standard I^{2}C.

III. 3D System Implementations

Pressure [12], audio [13], and cellular temperature [14] sensing systems are implemented using the aforementioned IoT[2] techniques (Fig. 9-Fig. 11).

Fig. 9 shows a 5×5×3 mm^3 pressure sensor [12] with sensing capability for 0−10,000 PSI. The 3D stacked system is composed of a MEMS pressure transducer, a battery and 6 IC layers. The system is powered by a 1-mAh lithium rechargeable battery with 2.2−2.8 V output voltage, which the power management unit layer [15] downconverts to 0.6 V and 1.2 V and upconverts to 3.6 V. The energy harvester layer [16] charges the battery using the photovoltaic cell layer [17]. The bridge-to-digital converter chip directly connects to the MEMS sensor and converts the hydraulic pressure signal to digital output. The processor chip includes 8-kB SRAM and Cortex-M0 to coordinate system operation, and additional signal processing is possible when needed. The system is hermetically sealed by black and clear epoxy. The black epoxy covers the IC layers, whereas the clear epoxy allows light access to the photovoltaic cells for recharging and optical communication [6]. The encapsulated system is initially programmed optically and tested in a high-pressure chamber. The processor controls
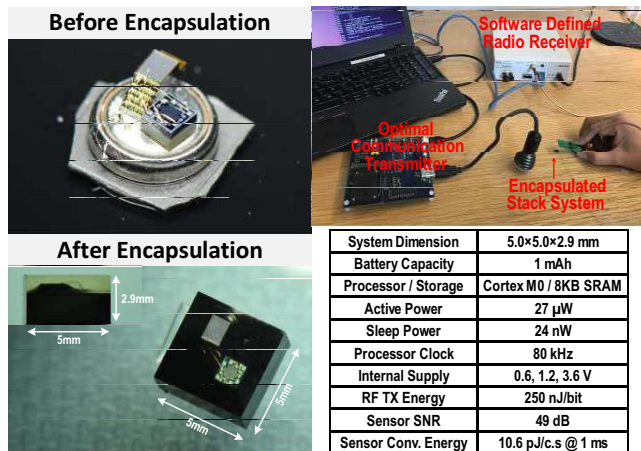
Fig. 9. Photograph of pressure sensor node (left), system demonstration setup (top right), and system performance summary (bottom right).
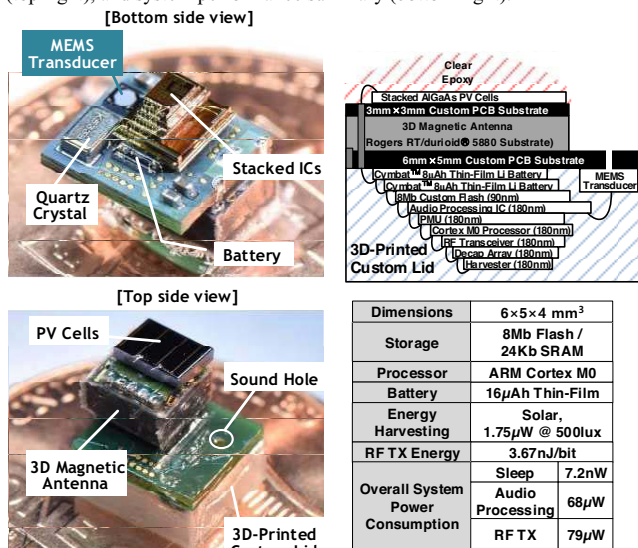
System Dimension	5.0×5.0×2.9 mm
Battery Capacity	1 mAh
Processor / Storage	Cortex M0 / 8KB SRAM
Active Power	27 μW
Sleep Power	24 nW
Processor Clock	80 kHz
Internal Supply	0.6, 1.2, 3.6 V
RF TX Energy	250 nJ/bit
Sensor SNR	49 dB
Sensor Conv. Energy	10.6 pJ/c.s @ 1 ms

Fig. 10. Photograph of audio sensor node (left), cross sectional diagram (top right), and system summary (bottom right).

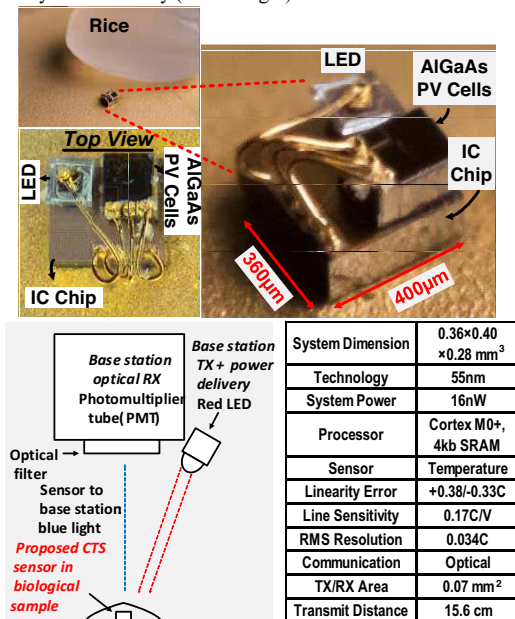| Dimensions	6×5×4 mm³	
Storage	8Mb Flash / 24Kb SRAM	
Processor	ARM Cortex M0	
Battery	16μAh Thin-Film	
Energy Harvesting	Solar, 1.75μW @ 500lux	
RF TX Energy	3.67nJ/bit	
Overall System Power Consumption	Sleep	7.2nW
	Audio Processing	68μW
	RF TX	79μW

Fig. 11. Photograph of cellular temperature sensor (top), CTS and base station diagram (bottom left), and system summary (bottom right).

System Dimension	0.36×0.40 ×0.28 mm³
Technology	55nm
System Power	16nW
Processor	Cortex M0+, 4kb SRAM
Sensor	Temperature
Linearity Error	+0.38/-0.33C
Line Sensitivity	0.17C/V
RMS Resolution	0.034C
Communication	Optical
TX/RX Area	0.07 mm²
Transmit Distance	15.6 cm

periodic storage of the pressure measurements in the SRAM. After pressure testing, the system is triggered to transmit the SRAM data wirelessly.

Fig. 10 shows the complete audio compressing system [13], [18], which measures 6×5×4 mm³ and includes a battery, an MEMS microphone, an RF antenna and six heterogeneous stacked ICs. The audio processor acquires and compresses audio data and streams it to the flash layer through a streaming bus. The 8-Mb flash chip [19] stores compressed audio data at 120 pJ/bit. The RF transceiver [5] communicates with a gateway up to 20 m NLOS. The other parts of the system are also carefully integrated to minimize the system volume. On a 6×5 mm² PCB, two 8-μAh thin-film batteries are stacked with the IC stack. A MEMS microphone is placed next to the IC stack. A 3D-printed custom lid covers the IC layers, the MEMS transducer, 3 capacitors for the RF transceiver, and a 32-kHz crystal. The lid forms a sound chamber and protects the IC layers from ambient light. The other side contains a sound hole for air passage and a 3D magnetic antenna, stacked with photovoltaic cells for energy harvesting. The system shows 54.6 dBA SNR and a 4-32× compressing rate, enabling 38 minutes of recording. It consumes 7 nW in sleep mode, 68 μW for audio processing and flash writing, and 79 μW for radio transmission. It harvests 4.07 μW at 1 klx and takes 10.5 hours to recharge after the maximum recording time.

The ICs in the pressure and audio sensors communicate with each other via MBus [10], [11]. This stacked system is modular in design, providing the flexibility to replace unnecessary features with desirable features depending on the application. This system also enables the use of heterogeneous technology nodes. For instance, the flash layer uses 90-nm technology, whereas the other IC layers use 180-nm technology.

Fig. 11 shows the complete cellular temperature sensor (CTS) [14], which is 360×400×280 μm³ in size. The sensor integrates a Cree LED for optical transmission, 50×50 μm² AlGaAs diode for optical reception, and 180×230 μm² AlGaAs diode for power harvesting on top of an IC chip. The 360×400 μm² IC chip integrates a Cortex-M0+ processor with SRAM, an optical communication receiver and transmitter circuits, and a subthreshold oscillation-based temperature sensor [7]. The fully assembled system is fully functional with wireless optical communication operation and no external connections. The base station photomultiplier detects blue LED flashes from the CTS via an optical filter to remove interferences. The always-on base station sends modulated red LED light to power the battery-less system and provide clock and data. The CTS operates at 3 klx with 16 nW system power consumption including temperature reading and data transmission via the LED. It achieves 0.034°C$_{RMS}$ resolution and +0.11/-0.08°C error.

IV. CONCLUSIONS

Continuous technological innovations are opening new possibilities in the IoT: mm-scale sensors, also known as IoT² devices. The design of IoT² systems has faced many challenges due to their unique size constraints. This paper discussed these

design challenges. We reviewed key low power circuit techniques for SRAM, miniature neural network accelerators, radio and optical communication, and temperature sensor. We also described efforts to minimize size and achieve a modular design. Furthermore, we presented three proto-type IoT[2] implementations using a minimal-volume 3D die-stacking method. The pressure and audio sensors integrated photovoltaic cells and a MEMS transducer, harvester, battery, processor, PMU, radio, optical receiver, and application-specific IC (audio processing chip, bridge interface chip) in mm-scale. The cellular temperature sensor featured a 0.04-mm^3 system volume, integrating an LED, photovoltaic cells, and an IC.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[2] D. Kim, G. Chen, M. Fojtik, M. Seok, D. Blaauw, and D. Sylvester, "A 1.85fW/bit ultra low leakage 10T SRAM with speed compensation scheme," in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, 2011, pp. 69–72.

[3] S. Bang *et al.*, "14.7 A 288μW programmable deep-learning processor with 270KB on-chip weight storage using non-uniform memory hierarchy for mobile intelligence," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 250–251.

[4] Y. Chen *et al.*, "Energy-Autonomous Wireless Communication for Millimeter-Scale Internet-of-Things Sensor Nodes," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3962–3977, Dec. 2016.

[5] L. Chuo *et al.*, "7.4 A 915MHz asymmetric radio using Q-enhanced amplifier for a fully integrated 3×3×3mm3wireless sensor node with 20m non-line-of-sight communication," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 132–133.

[6] W. Lim, T. Jang, I. Lee, H.-S. Kim, D. Sylvester, and D. Blaauw, "A 380pW dual mode optical wake-up receiver with ambient noise cancellation," in *2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits)*, 2016, pp. 1–2.

[7] K. Yang *et al.*, "9.2 A 0.6nJ −0.22/+0.19°C inaccuracy temperature sensor using exponential subthreshold oscillation dependence," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 160–161.

[8] I. Lee, D. Sylvester, and D. Blaauw, "A Subthreshold Voltage Reference With Scalable Output Voltage for Low-Power IoT

[9] M. Seok, G. Kim, D. Blaauw, and D. Sylvester, "A Portable 2-Transistor Picowatt Temperature-Compensated Voltage Reference Operating at 0.5 V," *IEEE J. Solid-State Circuits*, vol. 47, no. 10, pp. 2534–2545, Oct. 2012.

[10] P. Pannuto *et al.*, "MBus: An ultra-low power interconnect bus for next generation nanopower systems," in *2015 ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA)*, 2015, pp. 629–641.

[11] Y. s Kuo *et al.*, "MBus: A 17.5 pJ/bit/chip portable interconnect bus for millimeter-scale sensor systems with 8 nW standby power," in *Proceedings of the IEEE 2014 Custom Integrated Circuits Conference*, 2014, pp. 1–4.

[12] S. Oh *et al.*, "A 2.5nJ duty-cycled bridge-to-digital converter integrated in a 13mm3pressure-sensing system," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 328–330.

[13] M. Cho *et al.*, "A 6×5×4mm3general purpose audio sensor node with a 4.7μW audio processing IC," in *2017 Symposium on VLSI Circuits*, 2017, pp. C312–C313.

[14] X. Wu *et al.*, "A 0.04mm3 16nW Wireless and Batteryless Sensor System with Integrated Cortex-M0+ Processor and Optical Communication for Cellular Temperature Measurement," in *2018 Symposium on VLSI Circuits*, 2018, pp. C191–C192.

[15] W. Jung *et al.*, "8.5 A 60%-efficiency 20nW-500μW tri-output fully integrated power management unit with environmental adaptation and load-proportional biasing for IoT systems," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, 2016, pp. 154–155.

[16] W. Jung *et al.*, "An Ultra-Low Power Fully Integrated Energy Harvester Based on Self-Oscillating Switched-Capacitor Voltage Doubler," *IEEE J. Solid-State Circuits*, vol. 49, no. 12, pp. 2800–2811, Dec. 2014.

[17] A. S. Teran *et al.*, "AlGaAs Photovoltaics for Indoor Energy Harvesting in mm-Scale Wireless Sensor Nodes," *IEEE Trans. Electron Devices*, vol. 62, no. 7, pp. 2170–2175, Jul. 2015.

[18] S. Oh, T. Jang, K. D. Choo, D. Blaauw, and D. Sylvester, "A 4.7μW switched-bias MEMS microphone preamplifier for ultra-low-power voice interfaces," in *2017 Symposium on VLSI Circuits*, 2017, pp. C314–C315.

[19] Q. Dong *et al.*, "11.2 A 1Mb embedded NOR flash memory with 39μW program power for mm-scale high-temperature sensor nodes," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 198–199.