Compact Thermal Sensors for Dense CPU Thermal Monitoring and Regulation: A Review

Joseph Shor, Senior Member, IEEE

Abstract—Nearly all integrated circuits contain power and thermal management circuits. These two functions are intimately connected and help improve the power performance of the system. In recent generation computers, the functionality and complexity of the thermal sensing and management circuits have accelerated in order to improve the performance. Integrated thermal sensors are utilized to monitor the peak temperatures of the CPU’s to ensure that they do not exceed the reliability limits. There are many hot-spots on the CPU which have to be monitored, and these locations are not necessarily obvious during the design stage. As a result, it is required that there be multiple sensors on the CPU to enable accurate coverage. Thermal sensors are needed for cold-spot monitoring as well, in order to determine the required $V_{dd}$ level during low voltage, low power modes. The sensors are also used for fan regulation and power estimation. There can be many tens of sensors in a product, which require them to be small and low power. A variety of mechanisms have been utilized to form the sensing element. These include the parasitic PNP transistor, MOS transistors, resistors, as well as the thermal diffusivity of Si. In the past ten years, a trend towards miniaturization of thermal sensors has occurred. This review paper surveys the recent miniaturized integrated sensors which have been reported in the literature. The design tradeoffs for the CPU application are discussed, as well as the relative advantages of the different sensing mechanisms and their performances.

Index Terms—thermal sensors, thermal management, power management, microprocessors.

I. Introduction

The rapid advance of Moore’s law has enabled higher performance microprocessors and ever-increasing transistor counts, cores and features. The transistor densities have increased such that a processor chip can contains tens of billions of transistors in a relatively small area. This has caused the heat dissipation to be an important factor in computing. Thermal solutions help determine the maximum energy that the integrated circuit (IC) can produce. These heat sinks can be very costly and require significant area. Smaller form factor products can be thermally limited to relatively small thermal design points (TDP). In fact, the TDP point can determine the performance of the chip, such that the same core or chip could be used for laptop, desktop and enterprise processors depending on the thermal budget and heat sink solution which is utilized. The power and thermal management solutions are critical to the IC performance [1–6], and these techniques are co-optimized in order to produce the maximum power-performance. In recent generations more of the thermal and power management circuitry has been integrated on-die. The rise of multi-core processors has necessitated the replication of voltage, frequency and thermal domains per core [5, 7, 8].

Figure 1 shows the relative performance vs. TDP point for the processor code-named Sky-Lake [3]. The performance increases substantially for higher TDP, although this incurs additional cost and area for the cooling solution. The performance is improved by 25-50% compared to the prior-generation processor for various TDP points. Considering the high-cost of process shrink and substantial NRE (non-refundable-expenses) involved in designing a lead microprocessor, the value of a 25% performance increase must be economically substantial. However, Fig. 1 shows that the TDP of the cooling solution has a much greater effect on performance than the shrink, and the performance can increase 3-4X for a better cooling solution. Thus, the accuracy of the measured TDP point is an important factor in the power-performance tradeoff of these ICs.

![Fig. 1. Performance vs Cooling capacity (TDP) for the Sky-Lake processor [3].](image-url)

The TDP is limited by the maximum temperature which the silicon or the system can tolerate before reliability issues limit the lifetime of the device. In addition, there is a critical temperature at which instantaneous catastrophic damage will occur and the chip will become inoperable. This is shown in Fig. 2 for low TDP form-factors [1]. Figure 2a plots the tolerable time in operation vs. TDP. The system limit is shown in the red curve, while the Silicon junction temperature limit is shown in the blue curve. As the TDP power increases, the maximum time at that TDP in decreased. At the maximum power limit, catastrophic damage will occur, so that limit...
cannot be violated. This curve enables the system designers to push the power limit based on the temperature, as shown in Fig. 2b. The IC can enter turbo mode (the highest performance mode) for a relatively short time up to the junction limit, enabling the thermal solution to dissipate the heat before the system limit is exceeded. This turbo mode is a key feature to enable high performance.

Fig 2: (a) Maximum time vs. TDP[1], and (b) Power vs. time for turbo [2].

Fig. 3: Processor thermal map [11].

All of these high-performance, thermally limited modes are dependent on accurate thermal regulation to ensure that the IC does not violate the reliability limits. This requires accurate thermal sensors to measure the hottest portions of the IC. When the chip gets too hot, the thermal sensor can indicate that the maximum permissible temperature has been exceeded, and the thermal management logic can lower the operating frequency to allow the chip to cool down. This procedure is called “throttling” whereby the performance is reduced temporally. The throttle temperature can vary from product to product, based on the cooling solution. In addition to the throttle point, there is also a catastrophic temperature indication, usually ~ 20°C above the throttle temperature, where the entire platform shuts down to prevent irreparable damage.

In older processors, a single thermal sensor was utilized to measure temperature. This sensor could even be off-chip with only a thermal diode on die which was connected to the external thermal sensor / ADC via analog IO pins on the IC [9, 10]. The area and power of an on-die sensor was not a substantial factor in the product, since it was instantiated only once or twice in the die. However, as the power density and transistor count increased, it was observed that multiple hot spots could appear [7, 11-13]. Figure 3 shows a measured thermal map of a microprocessor [11], in which several hot spots are visible. Furthermore, the hot spots are instruction dependent and may not be very predictable during the design phase [12-14]. This has led to vendors placing as many as 10 sensing elements per core [7, 13, 15-18], leading up to many 10’s of cores (40 in [19]), depending on the amount of cores / IC. As a result of this large number of instances, the area and the energy consumption of the sensors became an important parameter, so efforts were made to reduce them substantially [12, 13]. In 2014 a compact sensor was defined as one with an area < 0.02 mm² and thus could be utilized to monitor multiple hot spots [13]. In today’s terms, there have been two more process generations, which would require a 4X shrink to 0.005 mm². Since analog circuits do not shrink very well, new topologies are required in order to maintain this shrink.

In older processors, the IC operated only at relatively high $V_{cc}$ levels, so performance always decreased at higher temperature. More advanced chips utilize a dynamic $V_{cc}$ regulation, know as DVFS (dynamic voltage and frequency scaling) [15, 16, 20]. The temperature measurement is used to determine the $V_{cc}$ level, in order to conserve power. At low $V_{cc}$ levels, the threshold voltage ($V_{th}$) becomes dominant, so the performance is reduced at lower temperatures, an effect known as ITD (Inverse Temperature Dependence). Thus, it is important to determine the coldest areas of the chip as well. Thermal sensors are also used for fan regulation as well as power estimation. The required temperature measurement accuracy of the product is ±3°C at the throttle temperature and ±5°C over the rest of the range. However, there is some inaccuracy associated with the sensor calibration, so the recommended Silicon accuracy is ±1°C at throttle and ±3°C over the range [13].

Hot-spot measurements from [19] are shown in Fig. 4 for a steady state processing application. Even for this steady-state application, thermal gradients of 0.5°C/msec are observed. Another study [21] showed much larger thermal transients during dynamic events, as shown in Fig. 5. Although the scale of Fig. 5 is in 0.1 sec units, it appears that transients of at least.
several °C/msec are observed. This indicates that the sensing speed of the compact sensor should be > 1kS/sec. The IC is also required to wake up quickly after deep sleep states, so a fast-initial temperature measurement should be performed 10μs after wakeup, after which a more accurate measurement can be taken [13]. Thus, in addition to being low power and compact, the sensors should have a relative fast conversion speed. Speed can also be traded with power by duty-cycling the sensors between readings.

There are a variety of thermal sensing mechanisms which are potentially appropriate for on-die monitoring. These have been summarized well by [22]. The most prevalent mechanism in industry is the parasitic PNP Bipolar Junction Transistor (BJT), found in the CMOS process [23–41]. The thermal diffusivity (TD) of bulk Si is also a candidate for small sensors [42–48]. The temperature coefficient of resistance (TCR) of resistors in the CMOS process has also been utilized effectively to design sensors [49–64]. Finally, MOSFET devices also have several temperature dependent characteristics, which have been used to derive sensing elements [65–80]. Each of these mechanisms has its relative advantages and disadvantages for the CPU application. This paper will review sensors published in the literature which can be implemented in a compact area, with an emphasis on those whose area is less than 0.005 mm². In section II, BJT sensor architectures will be examined, while sections III–V report on the TD, Resistor based and MOS sensors respectively. The sensors of each category are examined mostly in the chronological order of their appearance in publications. For each mechanism, some early prototypes are shown, which are not necessarily compact.

II. BJT BASED THERMAL SENSOR DESIGNS

A. Early BJT Sensors

Some of the earliest sensors published used the BJT mechanism. These sensors were not necessarily compact enough for the dense CPU monitoring, but since they formed the basis for many subsequent designs, they will be reviewed here.

The parasitic PNP in the CMOS process can generate two possible temperature dependent terms, as shown in Fig. 6, which is a basic bandgap reference circuit. The $V_{be}$ voltage has a negative temperature dependence, referred to as CTAT (complementary to absolute temperature). Two $V_{be}$ voltages at different current densities can also be compared to form a delta $V_{be}$ ($\delta V_{be}$) term which is proportional to absolute temperature (PTAT). Both of these terms are utilized in bandgap references (BGREF) [81, 82] to form a temperature-independent reference voltage. As such there is a strong similarity between bandgap references and BJT-based thermal sensors. In some cases, bandgap references are the basis for the sensing mechanism [19, 28, 29]. The $V_{be}$ voltage has a slightly non-linear CTAT temperature dependence of roughly 1.8mV/°C. This non-linearity can be corrected systematically in production by measuring the 2nd order coefficient over several units and applying this correction in the digital backend over the entire lot, which can number in the millions. The PTAT dependence of $\delta V_{be}$ is typically 0.1-0.2 mV/°C, but can be completely linear, assuming the BJT has an ideality factor which is independent of temperature [83].

Figure 7 shows a switched capacitor sensor from [27] which generates a differential $\delta V_{be}$ voltage. Two scaled current sources are utilized to alternatively charge the emitters of two PNP devices. The $\delta V_{be}$ are then amplified by the switched capacitor amplifier and input to a SAR (Successive Approximation Register) ADC (Analog to Digital Converter) to yield the temperature dependent signal. Dynamic element matching (DEM) techniques were used to reduce the error of the current sources charging the PNP’s. The sensor had a measured inaccuracy of 2°C over a 180°C range, and a resolution of 0.25°C and an area of 1.5 mm² in an 0.6μ process. This was one of the first $\delta V_{be}$ sensors ever published.

Another early sensor design, based on a bandgap core was proposed by Bakker [24]. The BGREF is shown in Fig. 8a, whereby both PTAT and CTAT currents are produced. The PTAT current is inherent to the $\delta V_{be}$ voltage formed across $R_{ptat}$. The CTAT current was formed by replicating the $V_{be}$
voltage across $R_{be}$. These currents were combined to form a temperature independent reference current, $I_{ref}$. The currents were compared in a first order sigma-delta loop in order to digitize the signal, which was decimated by a counter (Fig. 8b). A peak-to-peak inaccuracy of 2°C was achieved over a 160°C range at a resolution of 0.63°C and an area of 1.5 mm².

A high accuracy sensor was presented in 2005 by Pertijs which used a charge balancing approach [23]. This design laid the groundwork for many subsequent sensors. Based on the feedback, either $V_{be}$ or $\delta V_{be}$ would be integrated across the capacitors $C_s$ and $C_{int}$ (Fig. 9), such that the following equation is obeyed:

$$\mu = \frac{\alpha \cdot \delta V_{be}}{V_{be} + \alpha \cdot \delta V_{be}} = \frac{\alpha \cdot \delta V_{be}}{V_{ref}}$$

(1)

where $\mu$ is the output while $\alpha$ is the capacitor ratio between the input and integrating capacitors. The sensor utilized a 2nd order Sigma-Delta ADC to digitize the output. The $V_{be}$ and $\delta V_{be}$ signals were determined by the capacitor ratios which set the parameter $\alpha$. As capacitor matching is very accurate in the CMOS process, this can lead to a high precision design. Dynamic element matching and system level chopping were employed to reduce the transistor mismatch. A 3σ accuracy of ±0.1°C was achieved in an area of 4.5mm².

Fig. 9. Charge balancing sensor from [23].

B. Miniaturized BJT Sensors

Sensors utilizing the parasitic PNP in the CMOS process have been popular in industry and can be considered an industry standard [7, 11, 19, 25, 26, 28, 29, 65]. This is because of the predictable physics of the BJT [13]. There have been numerous industrial papers using the BJT mechanism, and not many which utilize other techniques, which indicates this preference. Several designs have suggested placing the BJT remotely from the readout circuitry in order to alleviate the area congestion of the hot spots [7, 11, 28]. This also enables multiple sensing elements per readout circuitry which reduces the area / sensor. However, this entails routing an analog signal across the digital die, which is cumbersome in a digital product and may be somewhat incompatible with digital flows. In addition, the sensing wires may be subject to digital attack noise, which would reduce the sensor’s accuracy. The longer the analog routing, the more severe this effect could potentially be. It is possible to shield the signals with $V_{ss}$ signals alongside, but this will not protect them from signals at higher and lower metals. It may be possible to mitigate the noise using a differential signal and integrating the signal over time [28]. Nevertheless, subsequent sensors attempted to reduce the size of the sensing element and readout circuit so that the entire sensor could be placed at the hot spots.

Fig. 10. BGREF sensor using a current-DAC [25].

Figure 10 shows an industrial sensor which was reported in several product chips and also supported the remote sensing option to save area at the hot-spot [25, 26, 30, 65]. The sensing core uses a BGREF topology, similar to [81], such that the sensing parameter is the $V_{be}$ CTAT voltage. A current digital-to-analog converter (iDAC) is used to digitize the temperature signal. The current is scanned digitally to compare $V_{ref}$ to $V_{be}$. When the two are equal the comparator causes the thermtrip signal to flip, indicating the correct temperature. There is a separate comparator and $V_{ref}$ for the catastrophic temperature indication (cattrip). Chopping was utilized to reduce offsets in both the BGREF branches, as well as the comparators. This sensor was utilized in many products, but as the digital parts of the IC scaled, the analog area of the sensor became too big (0.083 mm² in 32nm [65]).

Another industrial sensor was reported in 32nm in [28], as shown in Fig. 11. This too utilized remote sensing and an overall size was smaller (0.02 mm²). However, this area did not include a voltage regulator, which was also required [19]. The sensing element was based on a $\delta V_{be}$ signal, which required a better ADC than the prior art of this group (Fig.
10). The sensor utilized DEM and chopping to reduce offsets and compensated for the resistance of the remote sensor routing. A 76dB SNR 2nd order Sigma Delta Modulator was used to digitize the temperature signal. This paper also reported some high-volume manufacturing (HVM) results. The accuracy was < ±5°C without calibration and the resolution was 0.5°C. Presumably, this accuracy could be improved significantly with 1 or 2-pt calibration. The sensing speed was > 1 kS/sec, making this sensor very suitable for the CPU application.

A subsequent industrial $V_{be}$ sensor was reported in [19], and the core was based on the BGREF of [81]. The area was 0.006 mm² in 22nm. Miniaturization was achieved by reducing the amount of BJT’s required and combining some of the resistors, since these passive elements occupy significant area in advanced processes. The ADC was a voltage-to-frequency (V/F) converter (Fig. 12) which converted both the $V_{bc}$ and $V_{ref}$ input signals to frequency. A current bias charges one of two small capacitors in a ping-pong manner, until the voltage reached the input signal level, at which point the comparator trips and the capacitors switch. The comparator outputs were input into an SR Flip Flop to yield the frequency signal. These $V_{bc}$ and $V_{ref}$ related frequencies were compared using ripple counters to yield a digital word. The V/F circuit as well as the counters occupied very small areas, which also helped the miniaturization. The $3\sigma$ accuracy reported was ±1.5°C over a 110°C range. The resolution and number of bits were adjustable, such that a higher resolution (0.25°C) signal could be obtained at 5.6kS/sec, or a lower resolution (1°C) higher speed (22kS/sec) could be achieved. A subsequent $V_{be}$ sensor by this group was reported in 14nm using a similar BGREF and a first order Sigma Delta modulator [13]. The area was somewhat larger (0.0087 mm²), but the resolution FOM – (nJ/conv)*res² was improved by 10x [13], since the conversion time was much shorter (20μs). The design also used a highly miniaturize ADC [32] to assist in both the overall area reduction. These compact designs compromised in accuracy and resolution compared to some of the larger designs [23] in order to achieve miniaturization. However, this is justified by the application, which does not require accuracy better than ±3°C, but rather requires a compact, high speed and low-energy sensor. It should be noted that these $V_{be}$ sensors of [13, 19, 25] did not report systematic non-linearity correction, similar to [33, 34, 50, 53] and the errors include the inherent curvature of $V_{be}$.

Fig. 12: Miniaturized $V_{bc}$ sensor, with an area of 0.006 mm²[19].

![Image](image.png)

Fig. 13. (a) Miniaturized NPN sensor and (b) simulated waves

A 28nm NPN sensor with an area of 0.0038 mm² was reported by Eberlein in [36]. The sensor has a simplified structure and exhibits low mismatch since it utilizes a pseudo-differential pair formed by the scaled BJTs (Fig. 13a). A $δ$-$V_{be}$ voltage falls across the resistor, R1, to generate a PTAT current, while a $V_{be}$ voltage is formed across R2, and results in a CTAT current. A SAR algorithm is utilized to adjust R2, such that the CTAT and PTAT currents are compared and equalized, as shown in the simulated waves (Fig. 13b). A negative feedback is provided to the summing node, B, using a back-to-back source follower buffer, which provides feedback current $I_{diff}$. At the zero-crossing of $I_{diff}$, OUT provides a near-digital signal which can be detected to indicate that the proper SAR code for the temperature has been achieved. The PTAT and CTAT currents obey the following equation:

$$\frac{\ln(N)}{R1} \cdot \frac{K}{{q}} \cdot \frac{V_{be}}{R2} = 0 \Rightarrow T = \frac{V_{gs}}{R1} \cdot \frac{R2}{R1} \cdot \frac{R1}{R2} \cdot \frac{1}{TC} \cdot (2)$$

Where $KT/q$ is the thermal constant, $V_{gs}$ is the silicon bandgap (~ 1.2V) and $TC$ is the temperature coefficient of $V_{be}$ (~ -2mV/°C). The sensor achieved a $3\sigma$ error of ±0.8°C over a 150°C range at a sensing speed of 32μs for an 8-bit conversion. The energy was also relatively low at 0.56 nJ/conv for a resolution of 0.5°C.

The NPN BJT used in [36] is available in triple-well CMOS processes but is a cost-adder. In order to make the sensor pure-CMOS compatible, the group reported a modification of this topology in [68]. The NPN differential pair was replaced with a subthreshold NMOS pair, whose $δ-V_{gs}$ have similar
equations as $\delta V_{bc}$. A PNP BJT had its emitter connected to node C (Fig. 13a) to generate the $V_{bc}$ voltage. The sensor had an area of 0.0043 mm$^2$ in 22nm, an accuracy of $\pm 1^\circ$C, a resolution of 0.58$^\circ$C at a conversion time of 30$\mu$s, making this sensor very suitable for CPU applications. Both versions of the sensor [36, 68] exhibited a highly non-linear temperature dependence. However, since these characteristics were very repeatable between units, they could be cancelled out in the digital backend using systematic error correction.

In [34], Bass reported a compact $V_{bc}$ sensor in 65nm with an area of 0.003 mm$^2$, based on a switched-capacitor BGREF (Fig. 14a). Scaled PNP’s are used to form $V_{bc}$ and $\delta V_{bc}$ currents across capacitors C1 and C2 respectively. These currents are summed at node N3 to form a $V_{\text{ref}}$ voltage at the output based on the following equation:

$$V_{\text{ref}} = \frac{(C_2 \cdot \delta V_{bc} + C_1 \cdot V_{bc})}{C_3}.$$  

The $V_{bc}/V_{\text{ref}}$ ratio was digitized using a 1$^\text{st}$ order Sigma-Delta modulator, similar to [32]. Any leakage at the floating node, N3, could cause large inaccuracies, so low-leakage switches were used as well as IO-devices at the input transistor of the integrator amplifier. In addition, both $\delta V_{bc}$ and $V_{\text{ref}}$ could be trimmed by adjusting the capacitors C2 and C3 respectively. The switched-capacitor BGREF has an advantage that $V_{\text{ref}}$ is based on capacitor ratios which can be well matched in the CMOS process. Although the $V_{bc}$ voltage is non-linear, by implementing a $V_{\text{ref}}$ with a slightly positively tilted temperature coefficient, the $V_{bc}/V_{\text{ref}}$ ratio could be linearized [23]. This was implemented in [34] and the curvature error is shown in Fig. 14b. The bottom red curve is the inherent $V_{bc}$ curvature for a nearly flat $V_{\text{ref}}$. The blue curve is nearly linear and indicates a slightly positive $V_{\text{ref}}$. For the green curve, $V_{\text{ref}}$ has a larger temperature tilt, which also results in a larger slope and hence resolution. The sensor could achieve a nearly linear temperature response using this method without backend systematic calibration. The sensor achieved a $\pm 0.4^\circ$C inaccuracy over a 120$^\circ$C range, and an RMS resolution of 0.068$^\circ$C at a conversion time of 4.1 ms. To meet the CPU spec of $> 1$K/sec, a lower resolution conversion could be implemented with this design. As a follow-up to this design, the same group proposed a compact charge balancing sensor [33], with a similar architecture to [23]. A single 50$\mu$m$^2$ BJT sensing element was utilized, which was alternatively charged by two current sources. An additional capacitor was placed in the feedback path to lower the ratio between the smallest and largest capacitors in order to optimize matching. The sensor achieved an area of 0.0014 mm$^2$, a $\delta$$^\circ$C inaccuracy of $\pm 2.5^\circ$C, with a resolution of 0.22$^\circ$C. Seventy percent of the sensors area is occupied by digital circuits, so it should scale well with technology.

![Fig. 14](image1)

![Fig. 15](image2)

Although BJT’s have been widely accepted in the industry, their implementation in FINFET based processes may be somewhat problematic, due to the non-planar structures. To mitigate this a compact sensor based on the bulk diodes was suggested in [37] in 16nm (Fig. 15). These diodes were charge-pumped to negative voltages and the voltage charging of their cathodes (and capacitors C1 and C2) is proportional to the $V_{bc}$ currents, while the difference in the respective charging is proportional to $\delta V_{bc}$. These two can be compared and digitized using SAR logic to adjust one of the capacitors. Custom models were used to model the diodes, since they are usually not modeled in the process. One of the disadvantages of the BJT sensors is that they require a relatively high voltage to accommodate the $V_{bc}$ at low temperatures, which can be 0.8V. Thus, most of the compact BJT sensors required voltages of 1.2V or higher. The charge pumping action of this sensor enables it to function at 0.85V. An accuracy of $\pm 1.5$-$2.0^\circ$C was achieved without trimming at an area of 0.0025 mm$^2$. The resolution of the sensor was 0.3$^\circ$C at a very fast conversion time of 78 kS/sec and an energy consumption of 0.23 nJ.

Although BJT sensors have been placed in production in many products, the non-planar FINFET geometries and shallow junctions may make it harder to control the diode ideality factor. Thus, alternative mechanisms should be investigated for future applications.

III. THERMAL DIFFUSIVITY SENSORS

One of the limitations of the BJT, MOS and resistor-based sensors is that they can be process dependent. BJT sensors depend on the ideality factor of the diodes, and thus require
the process to track and control this parameter. The ideality factor can change with junction depth during manufacturing and a non-ideal case can even be temperature dependent [83]. The MOS sensors are dependent on $V_{th}$, which can be manipulated during the lifetime of the process. The TCR of the resistors are highly dependent on the doping level. In order to make the sensors more process-independent, sensors utilizing the thermal diffusivity of Silicon have been suggested in [42-48]. The diffusivity of silicon does not vary with doping level and is uniform across the die, since the Si crystal quality is very high. The diffusivity of Si varies with temperature as $D=1/T^{1.8}$. The basis for the TD sensors is an Electro-Thermal Filter (ETF) consisting of a resistive heater heated at a specific frequency. A thermopile, at a short distance from the heater, detects the phase shift of the frequency, and this is digitized by a Phase Domain Sigma Delta Modulator (PDSMD) [44].

![First thermal diffusivity sensor](image1)

**Fig. 16** – First thermal diffusivity sensor [42].

One of the first TD sensors is shown in Fig. 16 [42]. The ETF is heated by a resistor heater at a frequency $f_{drive}$ which causes it to generate heat at that frequency. The phase shift is detected by the thermopile and amplified by 30dB. This signal is then input to a gm/C integrator which forms the heart of the PDSMD. The chopper at the output of the gm element is driven by one of two phases ($f_{phi}$ or $f_{phi}$), depending on the Sigma-Delta feedback. Since high accuracy and high resolution were desired in this design, precision analog components and a large capacitor (40pF) were used, which occupied significant area (2.3mm²).

There were several subsequent versions of this design in which the size was shrunk down to 0.008 mm² [45], 0.0028 mm² [46], and finally 0.00165 mm² [43]. Part of this shrink was accomplished by reducing the accuracy and resolution performance compared to the larger earlier version, which is justified considering the CPU application. This enables the use of smaller analog elements and shorter integration times. In addition, part of the circuitry was digitized, enabling it to be significantly smaller. An example of this is shown in Fig. 17 [46]. Compared to Fig. 16, the 30dB amplification is removed, and the large capacitor is replaced with a current-controlled-oscillator (CCO), which also has the function of integrator. The comparator is replaced by an up-down counter. A short trimming cycle is utilized at the beginning of the conversion, and is included in the conversion time. The digitization of the components allows them to benefit from the process shrink and the circuit can be made much smaller.

The disadvantage of the TD sensors is that the heater requires a significant amount of power, making this mechanism the most power-hungry one. There was an industrial group which published a paper in 2013 with this mechanism [48]. This indicates that there may be some product interest in this mechanism because of its process independence, despite the high power.

![Resistor based sensor](image2)

**Fig. 18**. Resistor based sensor used in an RTC [49].

### IV. RESISTOR-BASED SENSORS

There has been a lot of recent interest in sensors utilizing the temperature coefficient of resistance (TCR) [49-64], with several such sensors appearing in the last few ISSCC conferences, from different groups. The resistor-based sensors tend to be lower power and higher resolution than other mechanisms. The high resolution is useful for real-time clocks (RTC), since high temperature resolution translates directly into frequency resolution. Thermal sensors in this application are used to compensate for the temperature coefficient of the crystal [49]. The sizes and conversion times of these resistor sensors have scaled down recently, making them a potential candidate for the CPU application.

An early resistor sensor for RTC applications is shown in Fig. 18 [49]. The voltage across an NWELL resistor is compared to the supply voltage of a current controlled ring oscillator (CCO). The frequency of the CCO is proportional to $1/(R*C_{B})$, and this can be compared to the reference frequency. A resolution of 0.04°C was achieved at a 0.83ms conversion time and an accuracy of 5°C over a 100°C range, with 1-pt trimming. The area of this sensor was 0.044m², which is close to the compact area definition of [13]. The sensor was integrated with an RTC to provide the temperature compensation.
Another relatively small (0.015 mm²) resistor-based sensor was reported in [51]. This was a 16nm industrial FINFET design. The sensor was both a voltage and temperature sensor and operated at $V_{cc}=0.7V$ (Fig. 19). A known current $I_{inj}$ is dropped across resistors R1 and R2 for temperature sensing. Two voltages along the resistors, V1 and V2 are used to toggle two comparators, leading to a charging or discharging of capacitor $C_{com}$. The output is thus a temperature dependent frequency which can be measured against a frequency reference using a counter. A p-p inaccuracy of ±0.1°C was achieved with a 2-p trim at a conversion times between 10μs-1.6ms. The combination of low size and relatively high-speed make this sensor potentially useful for CPU applications.

Figure 19. Resistor-based temperature and voltage sensor [51].

In principle, the TCR mechanism enables any type of RC filter to become a temperature sensor [50, 52-62, 64]. In many of these sensors, the reference voltage is the supply voltage, which can save significant area. This is a technique which is used in many SAR ADCs recently [63]. This depends on the supply voltage remaining stable throughout the course of the conversion, which may be an issue for noisy supplies. However, mitigation is possible if the signal is integrated over many cycles, thus smoothing any AC noise effects. There have been several groups which have made attempts to miniaturize these sensors, so that they could be used in the CPU application.

Figure 20 shows an example of a Poly-Phase Filter (PPF) resistor-based sensor [52], using silicided poly as the sensing element, due to its large TCR. The PPF is configured as a frequency locked loop (FLL). The integrator function of this loop is done by a CCO driven by a gm element, for area savings. The FLL signals are shown in Fig. 20b. A proportional clock, P is generated, as is a quadrature clock, Q. A comparator detects when $V_{ppf}$, positively crosses $V_{ppf}$. The signal Q is input into a phase-frequency-detector (PFD), and the feedback from the CCO minimizes the phase error, $\phi_{diff}$ between the rising edge of Vo and that of Q. Each time P toggles the amplitude of $V_{ppf}$ jumps by the amplitude of the clocks, P, which is $V_{ppf}$ in this case. Once the lock loops the rise of Vo and Q are in phase and $F_{ppf}$ is four times the PPF’s center frequency, which is temperature dependent, based on the silicided poly resistor. The sensor exhibits a non-linear frequency vs. temperature characteristic of 12°C over a 120°C range, which is associated with the resistors. However, since this non-linearity is highly repeatable between samples, it can be nulled systematically in the digital backend to give an accuracy of ±0.12°C over a 135°C range. It has a resolution of 2.5mK in a 1ms conversion time and consumes 68μW from a 0.85V supply. The sensor occupies an area of 0.007 mm² in the 65nm. An improved version of this sensor was published in [53], where the comparator toggled both the positive and negative crossing of the filter, improving the conversion time and energy by 2X, as well as the 1-p trim accuracy.

Figure 20. Miniaturized Poly Phase Filter (PPF) resistor-based sensor [52].

Figure 21 shows a recent compact resistor-based sensor which utilizes a floating Wheatstone bridge to generate a temperature signal [62]. The sensor is a miniaturized version of the circuits in [63, 64], and is presently the smallest of the resistor sensors, at an area of 1650 μm². It has a relatively good performance compared to sensors of a similar size. The bridge has positive TCR N-type diffusion resistors on one diagonal of the bridge and negative TCR P-type diffusions on the other diagonal. In order to maintain low power in prior-art resistive sensors, large resistors are general utilized, which occupy significant area. In this design, the bridge is turned on only for a short time, during the Track cycle (Fig. 21b) and then left floating, to enable smaller resistors without increasing the power. After this point a SAR conversion takes place, as can be observed in the $V_{ppf}$ and $V_{ppf}$ signal on Fig. 21b. Once the conversion is finished, the sensor is power gated to conserve static power between readings. In addition, the SAR capacitor DAC (CDAC) is placed in the metals above the circuits, to save area. The DAC uses very small unit capacitors, an idea developed in [63], as shown in Fig. 22.

Figure 21. Miniaturized floating Wheatstone Bridge sensor[62].
The C* and C+ signals are inverted, such that the three LSB’s (Least Significant Bits) are based on unit layout jogs. This enables a much smaller capacitor implementation for the LSB’s. The MSB’s (Most Significant Bits) are scaled up based on the ratio to the LSB’s, so the entire DAC is much smaller. The sensor also has a relatively low RMS resolution (0.53°C) to conserve area and enable a faster conversion, as is appropriate for the CPU application. It can achieve a p-p inaccuracy of 1.2°C/2.9°C with/without curvature correction, at a conversion rate of 50kS/sec and a power of 108nW. This sensor has one of the best specifications for accuracy and power performance for the sub-0.005m² sensors.

The temperature signal is converted to frequency by charging and discharging capacitors using this current. This frequency can be compared to a reference frequency to yield a digital output. The sensor yielded a ±1°C accuracy at an area of 0.028 mm², which is very compact, especially considering that it was designed in a 1μm process.

Fig. 22. Miniature SAR LSB technique using layout jogs[63].

Fig. 23. Early miniature MOS sensor [73].

V. MOS SENSORS

One of the earliest MOS based sensors was reported by Szekely in [73]. The MOSFET device has two temperature dependent characteristics, threshold voltage, $V_{th}$, and the gain factor, $\beta$ ($= \mu * C_{ox}$, where $\mu$ is the mobility and $C_{ox}$ is the oxide capacitance). The $V_{th}$ has a temperature dependence of $\approx -1.8mV/°C$, while $\beta$ has a temperature coefficient of $\approx +0.5%/°C$. The two parameters are summed together to yield a temperature dependent current as shown in Fig. 23. The two leftmost NMOS branches in Fig. 23 have scaled currents such that node C is roughly equal to $2 * V_{th}$, while node D is roughly equal to the Vdsat of the leftmost NMOS. This leads to a situation where $V_{th} \approx V_{dsat}$, such that the temperature dependence is:

$$I = \frac{1}{2} \cdot \frac{W}{l} \cdot \mu C_{ox} \cdot V_{th}^2$$  (5)

The current is integrated across two capacitors in a voltage-current to frequency converter, in a ping-pong manner, with the $V_{ref}$ as the reference. The frequency output was thus proportional to $\mu * V_{th}$. The sensor had an area of 0.0038 mm² in 90nm and 0.0051 mm² in 32nm. The 32nm version was designed with a very narrow pitch so that it could fit into the repeater channels of the processor, which are generally relatively empty of devices, so that the Silicon area is basically free. This was a 15X size reduction compared to the traditional BJT based sensor used by the group [30].

Fig. 24. Thermal probe from [66].

Fig. 25. High volume manufacturing dependence of $V_{th}$ and mobility in the thermal probe (Fig. 24) [65].

An industrial MOS based sensor was reported in [65, 66] with the application of dense CPU hot spot monitoring in mind. Ten of these sensors were positioned in a processor core at suspected hot-spots. A sparse analog supply was routed across the core to provide a 1.2V analog $V_{cc}$. The sensor (Fig. 24) utilized the temperature dependence of $V_{th}$ and $\mu$ (mobility). A voltage reference, $V_{ref}$, approximately equal to $V_{th}$ was output by the voltage reference block. The current reference has M4 sized much larger than M5, such that M5 is at the edge of the linear region, with its overdrive, $V_{gs} - V_{th} = V_{ref} = V_{th}$. The current in M5 obeys the following equation:

$$I = \frac{1}{2} \cdot \frac{W}{l} \cdot \mu C_{ox} \cdot V_{th}^2$$  (5)

Authorized licensed use limited to: Bar Ilan University. Downloaded on October 11,2020 at 05:32:33 UTC from IEEE Xplore. Restrictions apply.
achieved a p-p inaccuracy of 1.2°C over a 110°C range at a resolution of 0.2°C in a 0.1ms conversion time. Since this sensor operated in a current mode, the power supply sensitivity (PSS) was < 0.4°C/100mV. Some high-volume manufacturing (HVM) results of the output count (~frequency) and their correlation to mobility (~ $I_{dd}$) and $V_{th}$ are shown in Fig. 25. These results are from thousands of units from several production lots. There is a strong, nearly linear correlation of the frequency to both parameters. These sensors were used to characterize the hot-spots of Intel’s Sandy-bridge processor in [19].

Another sensor which utilized ring oscillators as the sensing element was proposed in [67] (Fig. 26). The frequency of an RO is dependent on $V_{th}$ and mobility. However, these two effects cancel each other out. For elevated temperatures, $V_{th}$ is lowered, thus increasing the transistor’s overdrive which raises the RO frequency, while mobility and hence conductivity are reduced, which lower the frequency. Two RO’s with different $V_{th}$ will have differing frequency dependencies as a result of this. This is shown in Fig. 26, and the ratio of the frequencies can be used to indicate the temperature. This sensor does not require a reference clock, as do most other sensors. The frequency ratio obeys the following equation:

$$\frac{f_{vco1}}{f_{vco2}} = \frac{(V_{dd}-V_{th1})C_{L2}}{(V_{dd}-V_{th2})C_{L1}} = \left(1 - \frac{\Delta V_{th}}{V_{th}}\right) \frac{C_{L2}}{C_{L1}}$$  

(6)

where $C_{L1}$ and $C_{L2}$ are the load capacitances of the ROs. These capacitances are designed to have the opposite supply dependences as the $V_{th}$ in order to cancel the effect of $V_{dd}$. The differing $V_{th}$ were implemented using the reverse short channel effect (RSCE). At longer $L$, the $V_{th}$ in this process is reduced. The load capacitors were designed to have a large contribution from the diffusion capacitors at the drain $C_{db}$. This capacitor decreases for higher $V_{db}$, which compensates for the $V_{dd}$ effect on the overdrive of the RO. The supply dependency was mostly removed using this technique and the PSS was 34 °C/V. The authors suggest a small voltage regulator (0.001 mm$^2$) to further improve the accuracy. Some off-chip circuitry was utilized to correct the non-linearity of the sensor, which is similar to the systematic error correction proposed in other works. The sensor had an area of 0.004 mm$^2$ in 65nm, an inaccuracy of ±0.9°C over a 100°C range, and a resolution of 0.3°C in a 22μs conversion. The energy consumption was 3.4nJ/conversion.

Fig. 26. Ring oscillator sensor from [67].

The delay of CMOS circuits is a strong function of $V_{th}$ and mobility, and thus ring oscillators (RO) or delay chains at CMOS levels can be utilized as thermal sensors. These sensors can offer relatively small sizes, fast conversion times, and low power. However, the basic inverter delay is more dependent on the voltage supply than it is on temperature, which is a problem that must be addressed in these sensors. Several versions of a delay-based sensor were suggested in [74, 76] which could achieve low area (0.008 mm$^2$) and fast conversion rates (469 kS/sec). The authors suggested utilizing a voltage regulator to mitigate the PSS problem, although the area of the regulator was not included in the sensor’s area. Another author used delay lines in addition to an accurate PTAT-based absolute thermal sensor (ATS) to generate a relative temperature sensor (RTS) [75]. The RTS was not necessarily accurate, but could indicate whether the temperature was higher, lower or equal to the ATS. A thermal map of the chip was generated using this approach. In [69], an analog supply voltage was used to power ROs, which were distributed across a system-on-chip (SOC). The regulated voltage minimized the $V_{cc}$ dependence and the process spread was controlled by adjusting the body bias of the ROs in the 28nm FD-SOI process. The exponential dependence of the ROs was linearized using a look-up-table in the readout circuit. The sensor achieved a 3σ inaccuracy of -2.2/+2.6°C at a 1-pt trim, for a sensing element (a.k.a. probe) size of 225 μm$^2$. The readout circuit occupied an area of 0.011 mm$^2$, with a power consumption of 616 μW and had a conversion rate of 27.8 kS/sec for 11 sensing probes. Thus, the average area and energy per probe was 0.001 mm$^2$ and 2nJ/S respectively.

MOS devices in subthreshold exhibit an exponential dependence which is very similar to BJTs’. They can be utilized instead of BJTs in BGREFs [82] and thermal sensors. For example, in [79, 80] a subthreshold PMOS configuration is suggested whereby the gate, drain and bulk are shorted to $V_{ss}$ and the source is connected in the BGREF configuration instead of BJTs (see Fig. 6 for the BJT version). This is
referred to as a DTMOS (dynamic threshold MOS), since the threshold voltage is dependent on the $V_{gs}$. A miniaturized sensor with a related concept was proposed in [70], as shown in Fig. 27. In this case, instead of placing a DTMOS in the BGREF, scaled ring oscillators are included instead. The feedback determines the supply voltages of the two ring oscillators, in a similar manner as a standard BGREF or constant-gm circuit [84] would determine the $V_{be}$ or $V_{gs}$ voltages. The RO’s are highly temperature dependent and $V_{cc}$-independent due to the current control. Assuming the ROs are in subthreshold, the frequency will obey the following equation:

$$F = \frac{Kt \ln(N)}{qR C V}$$

where $Kt/q$ is the thermal voltage, $N$ is the RO ratio, $R$ is the resistance, $V$ is approximately $V_{th}$, and $C$ is the total capacitance in the RO. The frequency signals need to be level shifted up to CMOS levels for the counter operation [85]. Since $F$ is strongly dependent on the capacitance, the input devices of the level shifter could potentially cause a $V_{cc}$ dependence as well. In order to prevent this, the sensor thus included a buffered supply for the level shifter. The sensor had an area of 1850 $\mu$m$^2$ in 65nm, a p-p inaccuracy of ±2°C over a 120°C range, a resolution of 0.32°C, and an energy consumption of 0.94nJ for a 10µs conversion.

![Fig. 28. (a) 2T reference [85] and (b) leakage-based ring oscillator sensor [72].](Image)

Taking the subthreshold concept to an extreme, transistors in cutoff mode will leak current according to the exponential temperature dependence. These currents can be somewhat difficult to control across process corners, but several authors have successfully demonstrated compact thermal sensors using this mechanism. A good example of this is shown in Fig. 28 and is based on a similar concept as the 2T-reference [86]. This circuit generates a voltage reference proportional to the $V_{th}$ difference between the native-$V_{th}$ IO device (which has $V_{th}=0$) and the diode connected standard $V_{th}$ (SVT) device below it (Fig. 28a). This effect can be utilized in a thermal sensor by replacing the SVT device with a ring oscillator (Fig. 28b) [72]. The RO supply, VVDD is approximately 300mV, which is sufficient to allow the RO to oscillate in the subthreshold mode. A 2-pt calibration is used to fit the output frequency to the subthreshold model. A differential delay cell, which looks 2-stages behind before turn-on, is utilized to reduce / eliminate crowbar currents. This enables the frequency to be highly linear with current. The sensor has an area of 0.0088 mm$^2$ in the 180nm process. The sensor should scale well with technology, since it consists of mainly digital circuits. The accuracy is ±0.2°C over a 120°C range. The energy is 0.6nJ/conv at a resolution of 0.073°C and a conversion time of 8ms. Considering the relatively high resolution, the conversion time can likely be reduced to meet the CPU specification as a tradeoff to resolution.

![Fig. 29. Leakage based delay elements from [78].](Image)

![Fig. 30: Relative Inaccuracy (RA) vs. Area. RA= (p-p error)* range/100.](Image)

![Fig. 31: RFOM vs. Area. RFOM = (Energy/conv.)*res$^2$.](Image)

Figure 29 shows a leakage-based sensor which uses a transistor in cutoff as either the NMOS footer (Fig. 29a) or a PMOS header (Fig. 29b). The leakage has an exponential dependence, causing the output, A, to have a long delay in one direction. This is sensed by a Schmidt-trigger buffer. A ring oscillator is formed by several of these stages and the frequency, $F$, exhibits an exponential temperature dependence, and obeys the following equation:

$$\ln(F) = a \cdot \frac{1}{T} + b$$

where $a$ and $b$ are fitting coefficients which can be extracted from a two-point calibration. The frequency is fed into a frequency-to-digital block, which measures $F$ against a reference frequency and linearizes it. The oscillator can be configured to have the cutoff transistor be either NMOS or PMOS for two modes of oscillation. After linearization there is still roughly ±1.5°C curvature, which can be compensated in the backend using systematic correction. An 3σ inaccuracy of
### Table 1. Comparison of the compact sensors.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Year</th>
<th>Technology (nm)</th>
<th>Type</th>
<th>Range (°C)</th>
<th>Area (μm²)</th>
<th>Supply Voltage (V)</th>
<th>Current (μA)</th>
<th>Conversion time (ms)</th>
<th>Resolution (μK)</th>
<th>Trim</th>
<th>Resolution FOM (μK/nJ)</th>
<th>PSS (°C)</th>
<th>BE Curvature Correction required?</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>2020</td>
<td>65</td>
<td>PNP</td>
<td>-10/-125</td>
<td>1400</td>
<td>3.43 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>No</td>
</tr>
<tr>
<td>33</td>
<td>2019</td>
<td>65</td>
<td>PNP</td>
<td>-15/-110</td>
<td>1500</td>
<td>3.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>37</td>
<td>2018</td>
<td>65</td>
<td>MOS</td>
<td>-20/-75</td>
<td>1700</td>
<td>4.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>36</td>
<td>2017</td>
<td>65</td>
<td>MOS</td>
<td>-20/-50</td>
<td>1900</td>
<td>9.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>69</td>
<td>2016</td>
<td>65</td>
<td>MOS</td>
<td>-10/-110</td>
<td>2100</td>
<td>9.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>68</td>
<td>2015</td>
<td>65</td>
<td>MOS</td>
<td>-20/-75</td>
<td>2300</td>
<td>9.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>67</td>
<td>2014</td>
<td>65</td>
<td>MOS</td>
<td>-20/-50</td>
<td>2500</td>
<td>9.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>66</td>
<td>2013</td>
<td>65</td>
<td>MOS</td>
<td>-10/-110</td>
<td>2700</td>
<td>9.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>65</td>
<td>2012</td>
<td>65</td>
<td>MOS</td>
<td>-20/-75</td>
<td>2900</td>
<td>9.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>64</td>
<td>2011</td>
<td>65</td>
<td>MOS</td>
<td>-20/-50</td>
<td>3100</td>
<td>9.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>63</td>
<td>2010</td>
<td>65</td>
<td>MOS</td>
<td>-10/-110</td>
<td>3300</td>
<td>9.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
<tr>
<td>62</td>
<td>2009</td>
<td>65</td>
<td>MOS</td>
<td>-20/-75</td>
<td>3500</td>
<td>9.85 ± 0.05</td>
<td>86</td>
<td>0.82</td>
<td>0.03</td>
<td>0.0</td>
<td>1.43 ± 0.03</td>
<td>6.14</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Notes:**
- Only sensors that have a complete sensor (sensing element + readout) were included.
- (a) Relative Inaccuracy defined as peak-to-peak-inaccuracy * 100/Range
- (b) FOM has an area of 17.46 μm² and a current of 68μA, including readout + 11 sensing elements. The area of 1000 um² and current of 68μA were obtained by averaging the overall area/current in terms of the number of sensing elements.

±0.7°C is achieved at a resolution of 16 mK and for a 1.3ms conversion. The leakage currents are not very $V_C$-dependent, so the PSS is 5°C/V. The energy consumption is 12.2 nJ/conv. The area of the sensor is 0.0017 mm² in 55nm CMOS and should scale well with technology, since the circuits are digital.

### VI. Discussion and Conclusions

There has been an accelerated interest in small sensors in the last ten years, due to the need to monitor multiple hot spots in the CPUs. Both industrial and academic groups have published numerous papers utilizing all the sensing mechanisms. Part of the design tradeoff in these sensors is that some of the high precision analog methods needed to be removed or scaled down in these sensors, such that often their performance is inferior to the larger sensors. However, this is justified in the CPU application which only requires a Si accuracy of ±3°C over the range and ±1°C at the high temperature throttle point [13]. Thus, the accuracy and resolution of the sensor can be somewhat relaxed to enable scaling. Part of the advantage of a small sensor is that it can be placed close to the hot-spots. This also enables an accuracy advantage, since placing it further away would result in some errors associated with the thermal gradients.

Figure 30 shows the relative inaccuracy vs. Area for Sensor’s found in Prof. Makinwa’s survey [22]. It is observed that most sensors with a sub 0.08 mm² area have relative inaccuracies worse than 1°C. Some of the recent resistor-based sensors in the 0.005-0.01 mm² category, as well as one MOS sensor are significantly better than this, but still trail the state-of-the-art. In the sub-0.005 mm² areas, only 3 sensors have inaccuracy better than 1°C. A similar trend can be observed in the resolution figure of merit (RFOM) vs. area (Fig. 31), where RFOM = (Energy/conv.)*$res²$. This FOM reflects that resolution is limited by noise, which can be reduced by linearly increasing the power, or quadratically increasing the current/energy. Thus, the RFOM exhibits the tradeoff between energy consumption and resolution. There are only four sensors with area < 0.005 mm² which have RFOMs < 0.1 nJ*K -2. Even the best of the compact sensors still trail the overall state of the art by 100X. Of the different mechanisms, the resistor-based sensors seem to have the best overall performance after miniaturization. These design compromises are part of the cost of miniaturization. However, as there are more groups working on this miniaturization now, one would expect this to improve with time. Table 1 shows a comparison of sub-0.005 mm² sensors, and also includes some of the best performers in the 0.005-0.01 mm² category. The best specification in different metrics are highlighted in green. Most of these sensors require curvature correction in the digital backend. As mentioned earlier, the dominant mechanism in products is the parasitic PNP BJT. This is because of the predictable and reliable physics of the BJT device [13]. However, in advanced processes, such as FINFET, the non-planar structure and shallow junctions may cause the ideality factor to degrade, making the other mechanisms more attractive for products. There has been a limited amount of papers published on FINFET-based designs, although many of the new product generations utilize this technology [12, 13, 37, 51]. It remains to be seen if the PNP sensors will remain dominant going forwards. The emergence of IoT products will accentuate the need for low energy, so it would be expected that the ultra-low energy architectures, including the resistor-based sensors, could be implemented in products in the future. As far as performance is concerned, this author does not expect required accuracy specification to tighten up; the accuracy specification of ±1°C to ±3°C should be good enough. It is notable in Table 1 that all but one of the sensors have appeared since 2015 and nearly half after 2019. This indicates that this miniaturization is important and relevant today and is being pursued by multiple groups.

K. Souri, Y. Chae, F. Thus and K. Makinwa, “12.7 A 0.85V 600nW all-CMOS temperature sensor with an inaccuracy of ±0.4°C (3σ) from −40 to 125°C,” 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, 2014, pp. 222-223.


