A High Bandwidth Memory with Remote Temperature Sensors connected a shared TSV

Jong-Pil Son, Ji-Hun Lee, Hong Shim, Jae Hyung Jung, Haesuk Lee, Kyuha Shim, Jae-seung Lee, Kyung-Soo Ha and Young-Soo Sohn
DRAM Design Team
Samsung Electronics
Hwasung, South Korea
jp.son@samsumg.com

***Abstract*— This paper proposes BJT-based remote temperature sensor (RTS) units integrated through a through-silicon via (TSV) to enable the placement of many thermal sensors. This architecture allows the sensors to be positioned close to the cell array with minimal layout constraints. The sensors' output signals are time-multiplexed and routed through a shared TSV to the main temperature-processing unit (TPU) on the base die. Consequently, the area overhead (2.2%) of the TPU on the core die is reduced to 0% by using proposed scheme. Process-voltage-temperature (PVT) variations for each core die are compensated by individual offset-calibration circuits implemented on the base die. With the proposed scheme and optimal RTS placement, power consumption drops by 75 % at IDD5, while refresh-window optimization yields an additional 7 % performance improvement. Moreover, costs are cut because the core-die TPU calibration test is no longer required, thereby eliminating yield loss.**

Keywords—Remote Temperature Sensor, TSV, HBM, BJT, Refresh

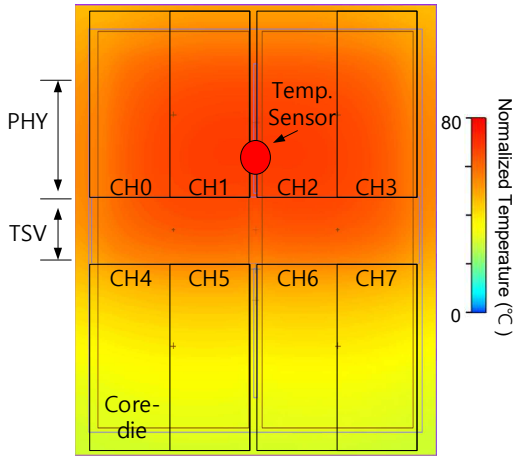
I. Introduction

Increasing the count of stacked semiconductor dies intensifies the electrical load on inter-die signal pathways including TSVs and consequently yields a more intricate temperature distribution in a HBM. In particular, localized hot spots may induce DRAM cell refresh failures and diminish the signal-to-signal margin. To accommodate the multi-hot-spot nature of a HBM, a substantial number of temperature sensors are required. The placement of TPU is constrained by their relatively large die-overhead (≈2.2%). Furthermore, positioning sensors proximate to active cell arrays proves challenging. In this paper, the bulky TPU is eliminated on each core die and substituted with distributed RTS implemented with BJT unit devices. The RTS outputs are time-multiplexed, routed through a shared TSV, and aggregated at a TPU on the base die. Dedicated calibration circuits on the base die enable compensation for PVT variations across stacked dies. Consequently, the system can dynamically monitor hot-spot temperatures, adapt DRAM refresh rates locally, curtail IDD5 refresh current by omitting superfluous refresh cycles, and thereby enhance overall read/write throughput. The approach also eliminates hot- and cold-temperature calibration time and the power consumption of the core die TPUs, while increasing manufacturing yield by removing per-die TPUs.

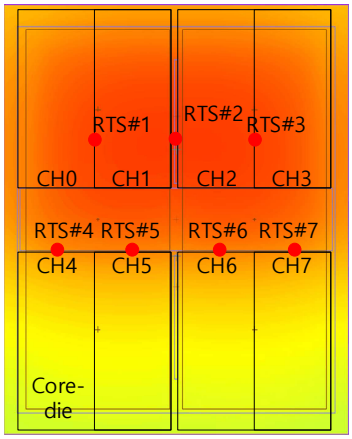
II. Temperature Distributions and Performance

A. Temperature distibutions in a HBM

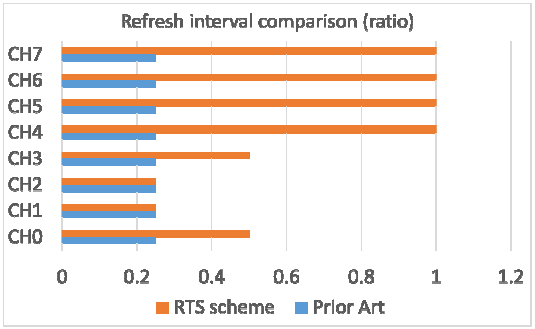
Fig. 1 (a) shows the temperature-distribution simulation results for high-current-consumption HBM4 scenarios.



(a)



(b)



(c)

Fig. 1. (a) HBM4 Core die temperature distribution simulation @ Burst Read operation (temperature sensor located at hot spot) (b) Proposed distributed RTS Scheme for optimized refresh intervals according to the temperature per each Channel (c) Refresh intervals comparison

TABLE I. ADVANTAGES OF PROPOSED REMOTE TEMPERATURE SENSOR (RTS) USING A SHARED TSV

Advantage	Amount (%)
Performance gain By using RTS/Channel	7.0%
Power Reduction By using RTS/Channel	75% reduction @ IDD5 condition
Area Reduction	2.2%/Core die
Test Time Reduction	0.5% @ EDS(Electrical Die sorting)
Yield Enhancement	0.25%

From the simulation we see that the core die is directly heated by the base die through the TSV region (center) and the PHY interconnect region (top). Because PHY, TSV and the user-defined usage scenarios generate numerous hot spots, a single temperature sensor cannot adequately monitor them. In conventional architectures the TPU block consumes about 2.2 % of die area, which prevents it from being placed inside the cell-array region.

Consequently, earlier core-die designs employed only one sensor, located at the predicted hottest spot outside the cell array, and therefore could not represent the temperature of the cell array in the die. As illustrated in Fig. 1. (b), the temperature difference between the hot spots and the coldest point is approximately ~50℃.

Therefore this paper suggest BJT unit devices for RTSs and deploy them in multiple locations where necessary including cell array region. Using this scheme, users can optimize the refresh intervals for each channel based on its temperature as shown in Fig.1. (c).

B. The goal of proposed scheme

By outputting the temperature for each channel, it is possible to reduce the refresh operation by approximately 70 ℃, thereby reducing power consumption. Performance can be improved by increasing the number of read/write cycles while omitting unnecessary refresh operations at low temperature. Additionally, by routing the data through a shared TSV and linking it to the TPU embedded in the base die, the dedicated temperature-sensor block in each core die can be removed. This yields a 2.2 % area reduction and shortens the hot/cold-calibration time required for the Electrical Die Sorting (EDS) test of the core die. Table 1 succinctly summarizes the benefits of the proposed scheme.

Eliminating refresh cycles accelerates read/write transactions, delivering a 7 % performance gain and a 75 % reduction in power consumption (reported in IDD5). Consequently, overall EDS test time drops by roughly 0.5 % of the total cycle, and the 0.25 % failure rate associated with the legacy TPU is eliminated, improving core-die yield.

III. MULTIPLE RTS CONNECTED WITH A SHARED TSV

Fig. 2 shows that multiple RTS blocks are multiplexed via a time-division MUX, and the combined output is transferred through the TSV to the base-die band-gap reference (BGR) [1]. Each RTS is calibrated by individually trimming its current-mirror.

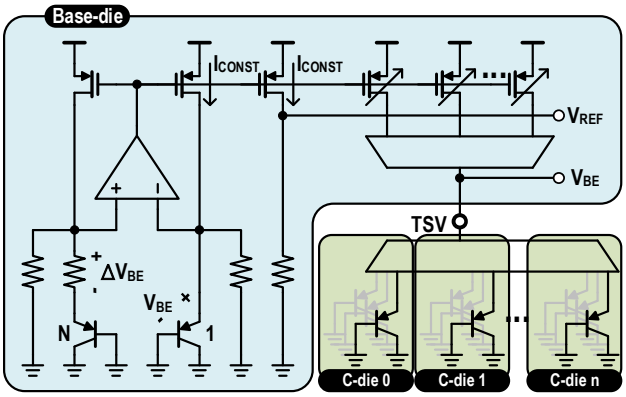


Fig. 2. Multiple RTS are connected to time-division MUX and signal transferred to the Base-die BGR via TSV.

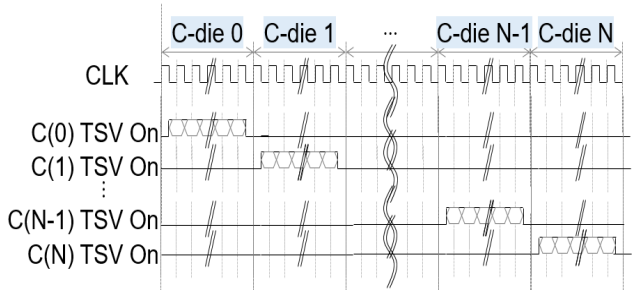


Fig. 3. Timing of the MUX for each RTS of core die. Time-division multiplexer switch each RTS and connects it through a common TSV, synchronized with the CLK signal transmitted from the Base-die to the TSV.

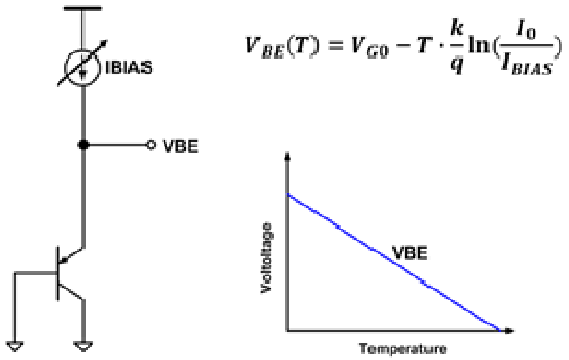


Fig. 4. The RTS connected to the BGR outputs a VBE voltage exhibiting a characteristic of decreasing with temperature (CTAT)

Fig. 3 shows the timing diagram of the multiplexer (MUX) for every RTS on the core die. A time-division MUX sequentially selects each RTS and routes it through a shared TSV, synchronized with the CLK signal that is sent from the base die to the TSV. Using the CLK signal delivered to each core die, the RTS output signal is independently conveyed to the base die through the shared TSV during the several CLK interval allocated to each core die.

Fig. 4 depicts that the RTS, when connected to the BGR, generates a VBE voltage with a complementary to absolute temperature (CTAT) characteristic. [2][3] This CTAT voltage is then fed, in sequence, to a VCO after a

temperature independent reference voltage as shown in Fig. 5.

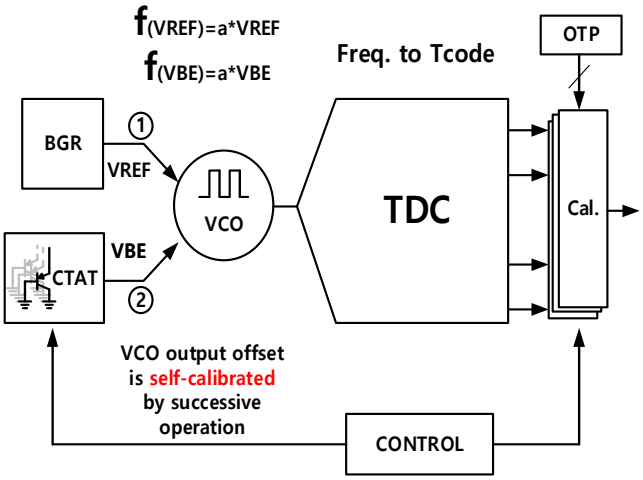


Fig. 5. Proposed self-calibrated VCO in a temperature sensor. The CTAT voltage connected to a VCO. The VCO outputs a temperature dependent frequency signal. The frequency difference converted into temperature code.

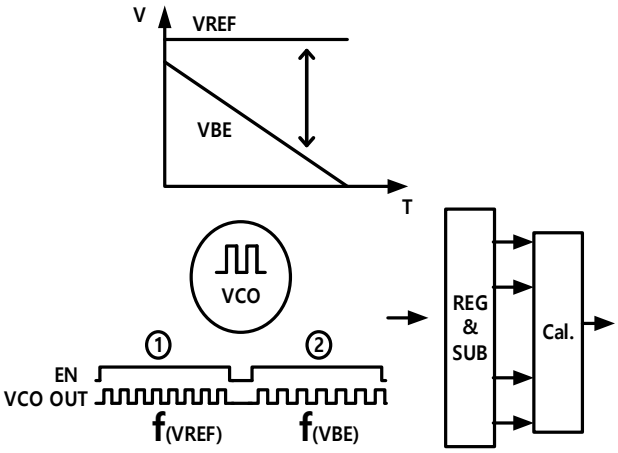


Fig. 6. The sequence for calibrating temperature sensors for acquiring accurate TCODE and temperature linearity using the calibration logics in temperature sensor main body.

Because the same VCO processes both signals, any VCO offset is cancelled by this sequential operation. Consequently, the frequency ratio between the reference and the CTAT outputs remains constant, independent of the offset. In the first interval, the reference voltage VREF, which does not vary with temperature during a predetermined CLK period, is connected to the input of the VCO, producing a CLK of constant period, and the generated CLK is digitalized through COUNTERs into outputs with multiple bits and recorded by the subsequent stage in registers. In the second interval, during a predetermined CLK period, the output VBE of the RTS, which has a CTAT characteristic, is connected to the input of the VCO, producing a CLK with a period that has a CTAT characteristic; this signal is digitized through the same COUNTERs and then compared with the previously input value. This compared difference is calibrated by following calibration logics. Then, the resulted calibration code is stored at OTP (One-Time Programmable array) as shown in Fig. 5 and Fig. 6. In this paper, 2-point calibration is used.

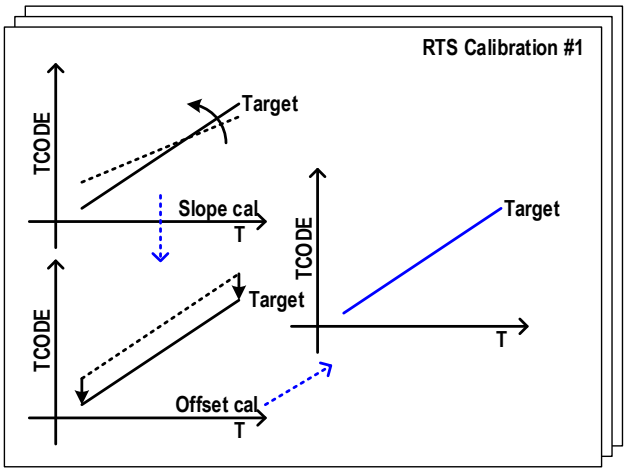


Fig. 7. The sequence for calibrating temperature sensors for acquiring accurate TCODE and temperature linearity using the calibration logics in temperature sensor main body.

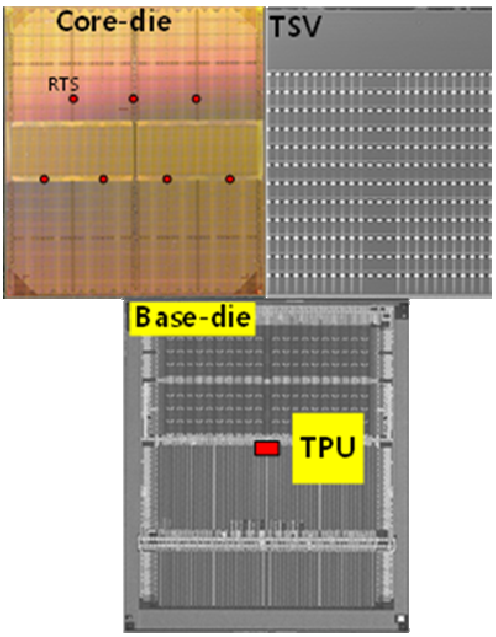


Fig. 8. HBM4 Base-die equipped with a Remote Temperature Sensors.

First, the deviation between the reference code and the RTS output is measured separately at the hot- and cold-temperature points, allowing us to calculate the code-versus-temperature slope. If the measured slope in a RTS is lower than the reference slope, then during slope correction for that RTS, multiply the measured slope by a weighting value to set the slope.

The second calibration step is for offset correction of the output TCODE.

It is a step that subtracts or adds the difference by comparing the temperature code in the operating temperature range with the reference value. Each RTS has an individual calibration code for calibration because it is distributed across different layers as shown in Fig. 7.

Of course, this calibration step is performed after the VBE voltage calibration via current variation in the Mirror of Fig. 2.

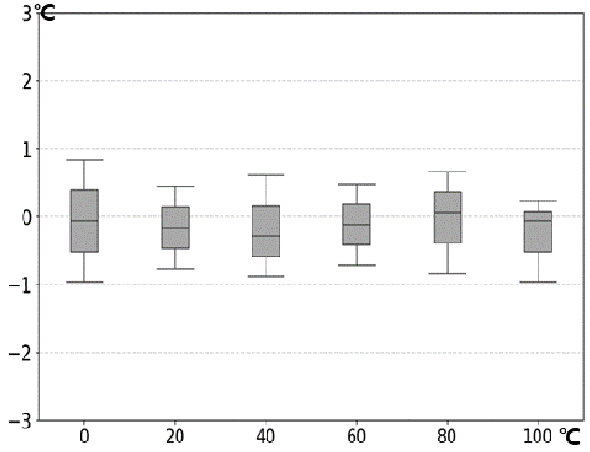


Fig. 9. RTS Inaccuracy (℃) plot according to the temperature.(measured 2249 samples)

IV. Experimental Result

The proposed scheme is implemented in HBM4 devices by using Samsung's TSV (Through-Silicon-Via) technology. Fig. 8 presents cross-sectional views of the TSVs, where each RTS (Rapid-Thermal-Sensing) output is time-multiplexed and routed to a shared TSV. Microscope images illustrate that both the core die and the base die adopt the proposed architecture. demand by 2.2 % and improves read/write performance by roughly 7 %. By applying selective refresh only to non-hot regions of the cell array, the IDD5 current is lowered by 75 %. Moreover, the overall test time is shortened by 0.5 % of the total EDS (Electrical-Design-Simulation) duration. These enhancements markedly boost HBM competitiveness, and the approach can be broadly applied to monitor hotspot temperatures in future 3-D products that are heat-sensitive. Physical evaluations confirm that the proposed scheme delivers the outstanding characteristics listed below. As shown in Fig. 9, the RTS temperature-code error remains within 1 ℃ across the measured temperature range (2,249 samples). Although the stacked dies are fabricated using different process technologies—memory-optimized for the core die and logic-optimized for the base die—this work is the first to demonstrate RTS operation in heterogeneous stacked dies. The silicon-level validation also proves the scheme's feasibility for real-world product deployment. Figures and Tables

TABLE II. Summary of Measurement Result and Comparison with State-of-the-Art Thermal Sensors

Items	This Work	ISJ23 [4]	TCAS24 [5]
Process	Samsung Foundry 4nm	65nm	180nm
Sensor	BJT	MOS	BJT
Area(mm^2)	0.12	9.15	0.42
Power (uW)	240	4	300
SampleNo.	2249	8	10
Inaccuracy	0 ~ -1℃	±0.6℃	±1.1℃
Stack with Diff. die	O(12H)	X	X

References

[1] H. Banba, H. Shiga, A. Umezawa, T. Miyaba, T. Tanzawa, and S. Atsumi, "A CMOS bandgap reference circuit with sub-1-V operation," IEEE Journal of Solid-State Circuits, vol. 34, Issue 5, pp. 670–674, May 1999

[2] Jee-Ho Park, Jung-Hye Hwang, Changyong Shin, and Seong-Jin Kim, "A BJT-Based Temperature-to-Frequency Converter With ± 1°C (3\Σ) Inaccuracy From – 40 °C to 140 °C for On-Chip Thermal Monitoring," IEEE Journal of Solid-State Circuits, vol. 57, Issue 10, pp. 2907–2918, Oct. 2022

[3] Seong-Jin Kim, Simon Sheung Yan Ng, David Wee, Yoon Hwee Leow, Fan-Yung Ma, and Sie Boo Chiang, "High Accuracy Remote Temperature Sensor Based on BJT Devices in 0.13-µm CMOS," IEEE International Symposium on Integrated Circuits, pp. 408–411, 2014

[4] Yuhang Shui and Aili Wang, "A 14.17 pJ · K2 FoM CMOS Temperature Sensor With 173 µm2 Sensing Core for Remote Sensing in 65-nm CMOS," IEEE SENSORS JOURNAL, VOL. 23, NO. 22, 15, pp. 27059–27067, Nov. 2023

[5] K. Elissa, "An Ultra-High Linear Digitization Temperature Sensor Based on SAR ADC With Common-Mode Temperature Drift Suppression, " IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, Vol. 71, NO. 3, pp. 1047-1051, Mar. 2024.