# A Software Framework to Calculate Local Temperatures in CMOS Processors

Alireza Rohani
Testable Design and Testing (CAES-TDT), University of Twente, Netherlands, Enschede
a.rohani@utwente.nl

Hassan Ebrahimi
Testable Design and Testing (CAES-TDT), University of Twente, Netherlands, Enschede
h.ebrahimi@utwente.nl

Hans G. Kerkhoff
Testable Design and Testing (CAES-TDT), University of Twente, Netherlands, Enschede
h.g.kerkhoff@utwente.nl

***Abstract*- A conventional technique to rise temperature in a processor involves the usage of thermal ovens or infrared techniques to heat up and then measure the temperature of the processor. However, local temperatures of each module cannot be controlled by these techniques. This paper presents a software mechanism to heat-up a processor while the temperature of each modules of the processor can precisely be calculated. In order to develop our mechanism, first a mathematical model to correlate dynamic power and local temperature has been developed; next a framework that calculates local temperature for any given workload has been presented.**

**In order to show the details of our model, the proposed framework has been applied to a thirty-two bit full-adder. The applicability of our framework has been demonstrated by using a complex DSP (Digital Signal Processor) as the case study. This paper will show that, despite common belief, there is no linear correlation between dynamic power and local temperatures of a chip.**

**Keywords—Thermal-aware design, CAD, Processors**

## I. INTRODUCTION

The developments of Integrated Circuits (ICs) ranging from Printed-Circuit Boards (PCBs) to VLSI and System-on-Chips (SoCs) has caused complicated faults to emerge in electronic circuits. One of these faults is known as Intermittent-Resistive-Faults (IRFs) which are mostly originating from imperfect interconnections [1] under extreme environmental conditions (high temperature or extensive vibration). So, during a high temperature, a loose interconnections, either soldering in a PCBs or a TSV (Through Silicon Via) in a 3D chip can impact the correct operation of a system. The impact is normally appeared as bursts of signals, with a random duration, in the output of the affected circuit [1]. However, the reproduction of these faults in a simulation-environment is very challenging because of their random behaviour with regard to the time of occurrence. This fact makes the diagnosis of IRFs very challenging since activation of IRFs needs a simulated-controlled environment which is hard to set-up, especially with regard to temperature. That makes the diagnosis of IRFs a hot topic in industry especially if we note that IRF rank among the highest in terms of occurrence (>50%) and the cost of dealing with IRFs is expected to increase in future technology nodes [1, 2, 3].

One of the methods to evoke IRFs is rising the temperature in each certain modules; however, conventional methods to rise the temperature suffer from non-controllability. In other words, it is almost impossible to control the exact temperature of each module separately while the chip is heat up. For example, if a chip is composed of four different modules, traditional methods are not able to set the temperature of the first module to 20°C while other modules are kept cold.

This contribution of this paper is to present a mechanism to conduct temperature-rising of a chip in a controlled way. The first step is to derive a thermal model that can extract detailed granularity of heat distributions in a chip. Such a thermal model enables the designer to extract the temperature of each local block of a design. An important factor is that the granularity of temperature profile should be flexible enough to be determined by the designer. The core contribution of this paper consists of extracting the local temperature of each module base on the executed workload. To be able to do this, first a correlation between local temperature and switching activity (dynamic power) needs to be developed and then based on the imposed activity that a workload imposes on a module, its local temperature can be calculated. Figure 1 shows this concept.

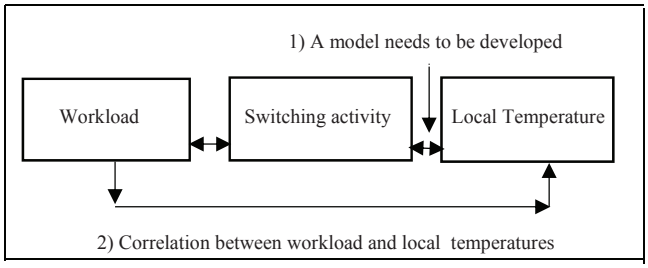


*Figure 1. Our core approach to define a model between the workload and local temperature*

This research was carried out within the FP7 BASTION project (#619871), financed by the European Committee (EC) and the Netherlands Enterprise Agency (RVO).

Our core approach to develop a thermal model based on switching activity of each module is based on the duality between heat transfer and electrical phenomena [4, 5], as shown in Figure 2. Switching activity incurs dynamic power and dynamic power will produce heat in a circuit. The correlation between switching activity and dynamic power is straightforward ($P_{dyn} \propto f$) but the correlation between dynamic power and heat transfer is more challenging. To be able to do the latter, heat is considered as a "*current*" which passes through a thermal resistance and creates temperature differences analogous to a "*voltage*". The power delivered to the block is considered as a current source. Moreover, thermal capacitances are considered to model the transient distribution of heat across the design.

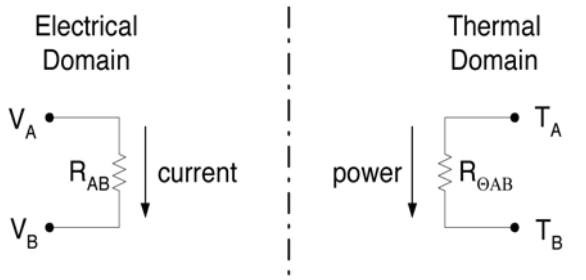


*Figure 2. Fundamental relationships in the electrical and thermal domains*

This model, which is known as the compact model [5, 6], considers the heat conduction between neighbouring modules. The modules can be of any granularity which is selected by a designer. The only criterion for the selection of modules is that the their power consumption should be measurable. The mechanism to extract $R_{\Theta AB}$ will be explained in next section. It is important to note that the model introduced in this work is different from those models that consider a linear dependency between temperature and power consumption. As it will be shown in experimental results, there is no straightforward dependency between power consumption and the temperature of each module.

The remainder of this paper is organized as follows: in section II, some related works are discussed. In Section III the basics of the thermal model is presented. Section IV illustrates the application of our model to a complex processor. Section V concludes the paper.

## II. Related works

The International Technology Roadmap for Semiconductors (ITRS) [7] has projected little change in the supply voltage of ICs that will occur in the future. As a result, power densities in ICs are projected to rise faster for the future technologies. As a result, the importance of temperature-aware designs will be more and more significant. The increased temperature has different consequences; i.e. it can significantly decrease the reliability of integrated circuits with regards to soft-errors, IRFs and aging faults. [8, 9]. Moreover it can deteriorate the performance of digital systems by reducing the carrier mobility of CMOS transistors.

In general, thermal distribution models can be categorized into two classes: numerical and analytical models. The numerical models, use the duality between heat transfer and electrical circuits to articulate the heat transfer to an RC (Resistive-Capacitance) network. The analytical-based methods, derive the local temperature of each part of the chip via a statistical approximation of the temperature of the entire chip.

The analytical-based methods are scarce in the thermal community. As one of the few works, the authors in [10] estimate the local temperature via forming a power grid on the desired granularity and subsequently calculate the consumed power in each grid point from the total power. The authors in [11] presented another analytical approach by using the generalized integral transforms (GIT) to estimate the local temperatures. Although their method is accurate, the granularity of temperatures that can be extracted are fixed and cannot be altered by the designer.

Numerical-based methods are the main focus of the thermal design community in recent years. In [12, 13, 14], the authors present different thermal models. These models all have detailed temperature distribution information across the chip. However, limitation of the above models is that the thermal package is over-simplified. For example, the PCB that greatly affects chip temperature distribution is not included in the models. The authors in [15,16,17], considered a more complex model which is able to model heat transfer via a cooling fan, however, the core thermal model is composed of only thermal resistance which neglect the *temporal* delay of heat transfer between two materials.

Two well-known numerical models to extract local temperature are called TEMPTEST [18] and HotSpot [19]. TEMPSET uses duality between electrical circuits and heat to model temperature; however, it contains one equivalent RC for the entire chip. As a result, extracting local temperature for each part of the chip is not very accurate. Hotspot which is presented in [19] uses analogy between electrical phenomena and heat transfer in order to model temperature. The main focus of the authors in [19] is to propose a run-time power management framework to avoid local hotspots while our concern is to develop a controllable mechanism to impose temperature in a CMOS processor.

## III. Deriving a compact model for local temperature

The first following subsection describes the basics of our thermal model which is able to extract local temperature of a chip. The second subsection shows the application of this model to a thirty-two bit adder.

### *A. Thermal model*

Our proposed model is based on the analogy between heat transfer and electrical phenomena. The distribution of heat to the neighbouring modules is modelled by thermal resistance. Moreover, a thermal capacitance is being used to

capture the time in which a module resists to change to a new temperature. The values of the thermal resistance and capacitance depends on different parameters such as the distance between modules, specific materials and initial temperature. The power which is delivered to the module is modelled as a "current source". The temperature of each module is analogous to the voltage of each node which is extracted by solving the analogous electrical equation. Considering that the layout of a chip is fixed, changing the workload affects only the value of the "current source". As a result, the temperature of each block for different workloads can be calculated rapidly.

A simple 3D model of an IC with three modules and their thermal resistances have been illustrated in Figure 3. The IC is composed of three modules, B1, B2 and B3.

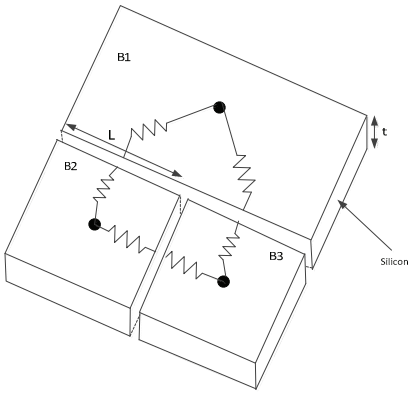


*Figure 3. Thermal resistances in a chip composed of three blocks*

The power is considered to be delivered at the centre of each module (indicated by a dot). This is the dynamic power which is averaged during the whole execution of a workload and can be extracted from available power analyses tools, such as Synopsys power analyser. The exact value of the thermal resistance is calculated based on the work presented in [20]. To calculate the thermal resistance, the following parameters should be considered:

- $A_s$: contact area of the heat source
- $A_p$: contact area of the neighbouring block
- *t:* thickness of the chip (shown in the Figure 3)
- *k*: thermal conductivity of the heat sink

For each block, the first three parameters can be extracted from place-and-route tools. For example, the area between B1 and B2 in Figure 3 can be calculated by the multiplication of L and t (L and t are shown in the picture). The *k* value depends on the material which needs to be extracted from the datasheet of the material. For example the *k* value for silicon is 1.3 W/ cm °C .

The exact equation between the cross-section area of two blocks, *k* and the thermal resistance is extracted from [20], which is as follows:

$$\frac{1}{R} = \frac{\sqrt{(Ap\,.\,As)}}{k\sqrt{(\pi\,.\,Ap\,.\,As)}} \times \frac{\beta\,.\,k\,.\,Ap + tanh(\beta t)}{1 + \beta\,.\,k\,.\,Ap.\tanh(\beta t)} \qquad (1)$$

where:

$$\beta = \frac{\pi^{3/2}}{\sqrt{Ap}} + \frac{1}{\sqrt{AS}} \qquad (2)$$

For example, if two silicon modules with the contact area of 25µm$^2$ and the thickness of 13µm are besides together, the thermal resistance between this two surfaces will be extracted as follows:

$A_s = 25 * 10^{-12}$ m$^2$
$A_p = 25 * 10^{-12}$ m2
$t = 13 * 10^{-6}$ m
$k = 130$ W/m°C (for silicon)

Therefore:

$$\beta = \frac{\pi^{3/2}}{\sqrt{Ap}} + \frac{1}{\sqrt{AS}} = 1312000 \text{ m}^{-1}$$

$$\text{Tanh } (1312000 \times 13 \times 10^{-6}) = 0.99$$

$$\beta \times k \times A_p + \tanh(\beta t) = 994264 * 10^{-6}$$

Which results to:

$$R_{thermal} = 470 \text{ m°C/W}$$

This value for all six thermal resistances as shown in Figure 3 need to be calculated. A MATLAB routine can quickly calculate these values.

As stated before, thermal capacitances are also involved when different materials interact with each other in heat conduction. This happens when the heat conduction between a chip and its package and/or from a package to the air should be modelled. Driving the thermal capacitance is more straightforward as compared to the thermal resistance as it is proportional to the thickness and contact area of interacting materials. From [5], the following formula can be extracted:

$$C_{thermal} = c.T.A \qquad (3)$$

Where c denotes the thermal capacitance per unit volume, being $1.75 * 10^6$ J/m$^3$ K for silicon. The idea of capacitance in this work is to mimic the transient change when heat transfers from silicon IC to the package and from the package to the ambient.

Finally, all these values can form an electrical circuit which needs to be solved by an electrical circuit simulator, such as PSPICE. As stated before, the voltage of each node represents the temperature in the centre of that module. The next section shows the application of this model to a thirty-two bit adder.

*B. Temperature evoking in a thirty-two bit adder*

In order to show the details of the temperature evoking, a thirty-two bit adder has been selected. The general set-up of the adder is composed of thirty-two full- adders (FA). There are two inputs, *A* and *B*, each thirty-two bits wide. The entire system is modelled with VHDL language. The selection of workloads on *A* and *B* are in such a way that control activities on specific parts of the circuit. For example, when A equals

to 00000005 and B equals to 0000000F, the least significant bits of the adders will change their values and consequently will have more activity as compared to the rest of the circuit. Frequent changes of *A* and *B* impose different activities in Least-Significant Bits (S0 to S10), Medium-Significant Bits (S11-S20) and Most-Significant Bits (S21-S32), consequently. The power imposed on the circuit is calculated and will be used to build the analogous electrical network in the circuit.

A power compiler has been used to calculate the power consumption (dynamic power) of each module (full-adder) per workload. The power consumption will be used in generating the analogous electrical workload of the circuit and consequently extract the temperature profile of the system.

Table I shows three different workloads, as well as the dynamic power consumption of three full-adders that are located fairly far from each other. This information was directly obtained by the power analyser of Synopsys tools. As can be seen in this table, each workload incurs a different power consumption on specific blocks. These power information need to be calculated for all full-adders. The next step is to build the analogous electrical network for heat transfer, and using the information of power analyser to complete the analogous electrical network.

*Table I: Dynamic power results*

| Workloads | Inputs | | Power (mW) | | |
|---|---|---|---|---|---|
| | A | B | fA-2 | fA-20 | fA-30 |
| W1 | 0000 0005 | 0000 000F | 8.13 | 1.02 | 1.02 |
| W2 | 0000 FF00 | 0000 5500 | 1.07 | 6.50 | 1.26 |
| W3 | 0F00 0001 | 5F00 0000 | 1.05 | 1.04 | 6.89 |

In order to build the analogous electrical network of the circuit, the layout information is also required. That information is obtained via the Cadence place-and-route tool. However, the package which is placed around the adder is also considered.

Figure 4 shows the analogues electrical model for first three full-adders. As can be seen in this picture, the package which accommodates the adder is divided into five parts: one corresponds to the area right under the adder (Rp) and four trapezoids for the periphery (RP1, RP2, RP3, RP4). Each periphery area conducts temperature away from one of the blocks while the Rp area conducts the temperature away from all the blocks of the adder. Finally, the convective heat transfers from the package to the ambient is represented by a single thermal resistance (R6). If the package covers the top of the chip, another ambient thermal resistance on top of the chip is also required.

As can be seen in this picture, the switching power of each module has been represented as a current source in which their values are derived directly from the power synthesis results. The heat conduction from each module to the neighbouring modules are described by five thermal resistance, named R1, R2, R3, R4 and R5, for the block FA1. Notice that R5 is the heat conduction between the FA1 and the area of package which is situated right beneath the IC (the dotted lines are situated beneath the IC). The exact values of all these resistances can be extracted from the layout information and applying them to formulae 2 and 3 in subsection A.

The R6 resistance is the heat conduction between package and ambient temperature. The value of this specific thermal resistance is derived from literature [10]. Moreover, several capacitances are placed to emulate the gradual heat conduction between the IC and the package as well as between the package and the ambient. The value of this thermal capacitance is proportional to the thickness and area of the blocks. These values are also derived from the layout information.

Figure 5 shows the dataflow that has been used to construct this analogous electrical network. The electrical network needs to be solved by an electrical simulator (SPICE) and then the voltage of each node is determined. This voltage represents the temperature of each block during the execution of the corresponding workload. Figure 6 shows the temperature raise of FA1 for three workloads. As it can be seen in the figure, workloads 1 (W1) imposes the highest temperature raise on this adder. Moreover, the exact temperature of this block after execution of workload W1 can be seen in this figure (The temperature raise needs to be added to the ambient temperature).

Figure 6 shows that after approximately three seconds, the temperature of block 1 will be around 32.5°C, considering the ambient temperature of 21°C) if workload W1 is executed. While workload W3 will cause a temperature of 31°C on this block.

Having the exact temperature raise on each module, the designer can quickly apply a workload to the circuit to raise the temperature of a "specific" module to a certain temperature. This will be very useful to "locate" any temperature-induce faults that arise in certain temperature.

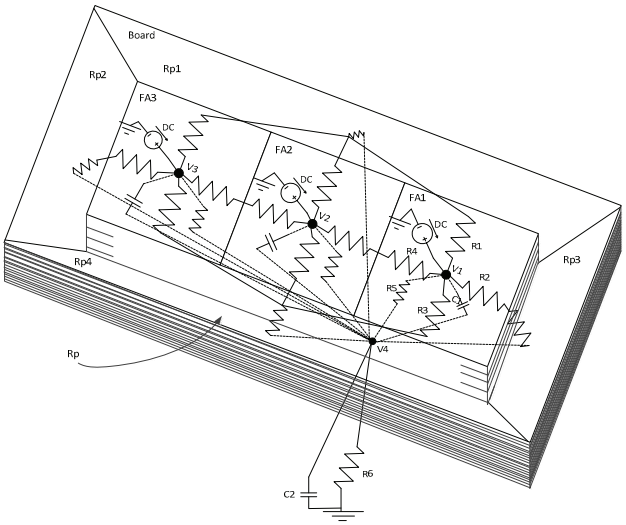


*Figure 4. The analogous electrical network of the adder*

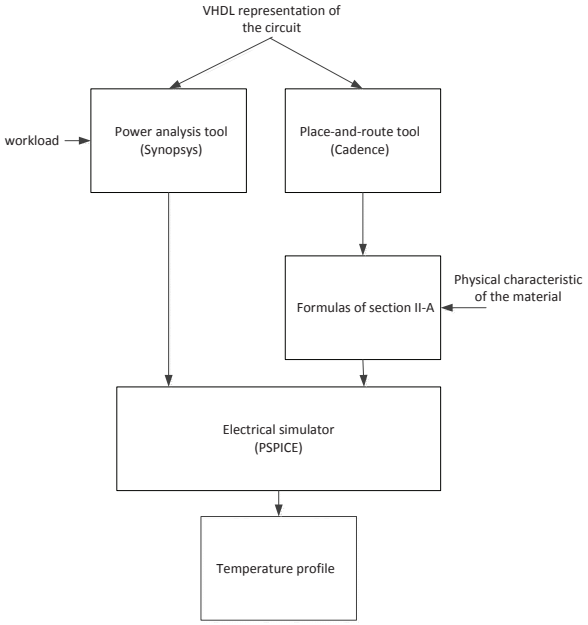


*Figure 5. dataflow to construct the temperature profile*

## IV. CASE STUDY

This section introduces the results which are obtained from applying the framework of Figure 5 to a complex processor. Subsection A briefly explains the processor and subsection B shows the temperature profile of the processor.

### *A. Xentium DSP Processor*

To implement our approach, a high performance Very Large Instruction Word (VLIW) processor, the Xentium® processor has been selected. The Xentium® processor which is designed by Recore Systems [21] is a high performance processor that is being designed for high performance computing. The data-path architecture of this processor is shown in Figure 7. The data-path is based on a VLIW architecture that comprises of ten execution units and five register files. Each execution unit is responsible for a certain class of instructions. E units (E0 and E1) perform load/store instructions; A and S units (A0, A1, S0 and S1) perform arithmetic and logical operations. M units (M0 and M1) are multipliers. C and P units (C0 and P0) perform instructions which program counter is involved. All functional units can access five register files (RFA, RFB, RFC, RFD and RFE) in parallel. This processor was a perfect candidate for us because the dedication of each execution units to certain instruction allows us to adjust a workload quickly in order to impose more/less temperature rise on each execution unit. For example, rising temperature on M units imply a workload that has loads of multiplication instructions. A set of three different workloads has been applied to this processor. The total activity of each execution unit with regard to each workload is shown in Table II. We will show the temperature profile obtained from applying each workload on this processor.

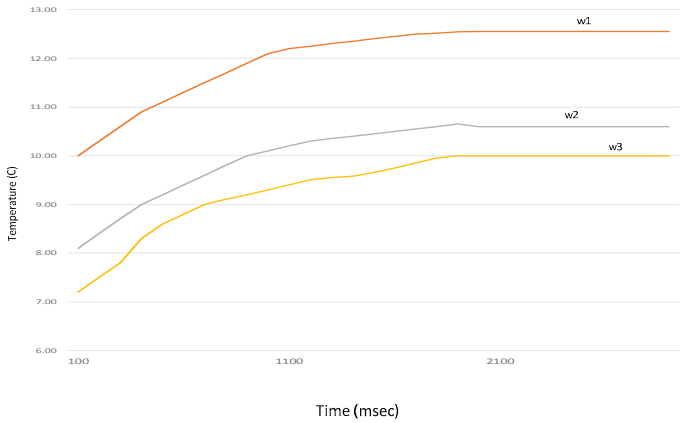


*Figure 6. Temperature raise in full-adder1 for several workloads*

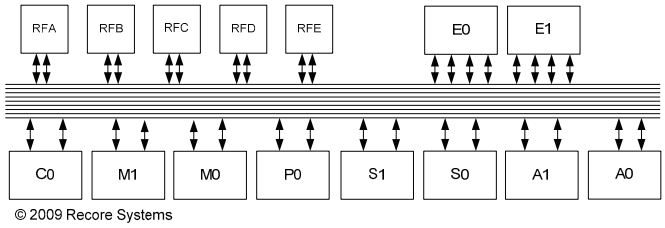


*Fig. 7. The Xentium data-path architecture [21]*

*Table II: Dynamic power results*

| Workload | Switching activity (%) | | | | | |
|---|---|---|---|---|---|---|
| | E0/E1 | M0/M1 | A0/A1 | S0/S1 | C | P |
| W1 | 80 | 10 | 10 | 10 | 50 | 40 |
| W2 | 10 | 10 | 90 | 90 | 40 | 40 |
| W3 | 20 | 20 | 20 | 20 | 90 | 80 |
| W4 | 10 | 90 | 10 | 10 | 40 | 40 |

### *A. Temperature profile of Xentium Processor*

The four different workloads have been applied to the Xentium processor and based on the procedure depicted in Figure 5, the temperature rise in each block has been extracted. The extracted numbers have been shown in table III. The hottest block in each workload is also shown in Bold. A more detailed expansion of heat in different blocks of the processor have been shown in figure 8. The dark red means hottest temperature.

The first observation from Figure 8 is that the temperature rise of each module does not follow a linear dependency with delivered power (which is shown in table II), but other factors such as the temperature of adjacent blocks or the position of the block on the package play important roles. For instance, workload W1 imposes more power on C/P units compared to A/S unit but the temperature rise on A/S units are higher. The reason for this observation is that A/S units are placed closer to the hottest units (E units). In case of other workloads, there is also no linear correlations between dynamic power and the temperature of that block.

*Table III. temperature raise for each workload*

| Unit | Workload1 | Workload2 | Workload3 | Workload4 |
|---|---|---|---|---|
| **C** | 7 | 9 | 5 | **9** |
| **M** | 4 | 9 | 11 | 4 |
| **P** | 6 | 11 | **15** | **9** |
| **S** | 8 | **14** | 9 | 5 |
| **A** | 8 | **14** | 7 | 5 |
| **E** | **13** | 4 | 4 | 4 |

Another observation from Figure 8 is that ambient conduction has an important impact on the temperature rise of each block. For example, in the case of W2 and W3, the E1 unit has the lowest temperature, while the switching activity of this block is in the same range of other blocks. Our detailed investigation showed that E1 unit loses a lot of heat to the ambient because it has two adjacent sides with the ambient, while it is not the case for E0 unit.

The extracted temperature profile allows test engineer to have a more clear view about the behaviour of a workload on heat generation of a processor. This will help to identify and locate temperature-induced faults more precisely in a processor.

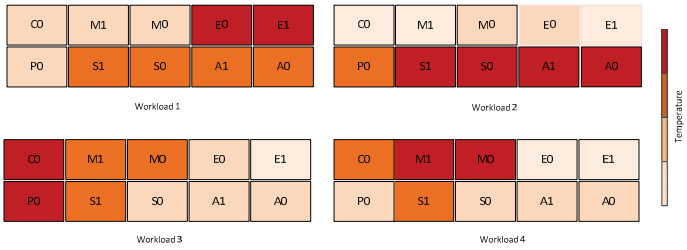


*Figure 8. Distribution of heat for different workloads (The darker a block is, the hotter it becomes for that workload)*

## V. CONCLUSIONS

In this paper, a framework to extract the temperature profile of a system has been presented. First, the power consumption of local blocks are extracted, and using the layout information of the chip, an analogous electrical network of heat transfer is produced. Solving the analogous heat transfer can reveal the temperature profile of the system.

The proposed approach in this paper has been applied to a Xentium processor and temperature profile of the processor has been extracted for different workloads. Moreover, the proposed approach can define the temperature profile for any granularity, as long as the power consumption of the desired modules is measurable. It is very important to note that there is no straightforward dependency between the power consumption and the temperature of each module.

## VI. FUTURE WORKS

This paper showed a mechanism to measure the local temperature of CMOS processors by means of workload. Our next step is to adjust the workload to impose certain temperature on each module. As a result, a mechanism will be developed that can define a thermal workload that can impose exact temperature on each module.

## REFERENCES


[1] F. J. Mesa-Martinez, M. Brown, J. Nayfach-Battilana, J. Renau, "Measuring Power and Temperature from Real Processors," Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on, pp. 1-5, 2008

[2] H. G. Kerkhoff and H Ebrahimi, "Intermittent resistive faults in digital cmos circuits," IEEE International Symposium on Design and Diagnostics of Electronic Circuits and Systems, pp. 211-206, 2015

[3] J.T Chang,E. J. McCluskey, "Detecting bridging faults in dynamic CMOS circuits," IEEE International Workshop on IDDQ Testing, pp. 106-109, 1997.

[4] A. . Krum. Thermal management. In F. Kreith, editor, The CRC handbook of thermal engineering, pp. 2.1–2.92. CRC Press, Boca Raton, FL, 2000.

[5] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, "Temperature-Aware Microarchitecture, " International Symposium on Computer Architecture, 2003.

[6] K. Skadron, M. R. Stan, K.Sankaranarayanan, W. Huang, S. Velusamy, D. Tarjan, "Temperature-aware microarchitecture: Modelling and implementation," ACM Transactions on Architecture and Code Optimization (TACO) , vol. 1, no. 1, pp. 95-125, 2004

[7] SIA. International Technology Roadmap for Semiconductors, 2001.

[8] D. Rossi, M. Omana, C. Metra, A. Paccagnella, "Impact of Bias Temperature Instability on Soft Error Susceptibility," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 23, no. 4, pp. 743-751, 2015.

[9] A. P. Shah, V. Neema, "Effect of process, voltage and temperature (PVT) variations In LECTOR-B (leakage reduction technique) at 70 nm technology node," International Conference on Computer, Communication and Control (IC4), pp. 1-6, 2015

[10] K. Skadron, M. R. Stan, K.Sankaranarayanan, W. Huang, S. Velusamy, D. Tarjan, "Temperature-aware microarchitecture: Modelling and implementation," ACM Transactions on Architecture and Code Optimization (TACO) , vol. 1, no. 1, pp. 95-125, 2004

[11] P. Huang and Y. Lee, "Full-Chip Thermal Analysis for the Early Design Stage via Generalized Integral Transforms," IEEE Transactions on Very Large Scale Integrated (VLSI) Systems, vol. 17, no. 5, 2009, pp. 613-626, 2009.

[12] T. Wang, C. Chen. "3-D thermal-ADI: A linear-time chip level transient thermal simulator," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 21, no. 12, pp. 1434–1445, 2002.

[13] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif. "Full chip estimation considering power supply and temperature variations," In Proc. of Intl. Symp. On Low Power Electronics and Design (ISLPED'03), pp. 78–83, 2003.

[14] P. Li, L. Pileggi, M. Asheghi, and R. Chandra. "Efficient full-chip thermal modeling and analysis," In Proc. of Intl. Conference on Computer-Aided Design (ICCAD), November 2004.

[15] C. J. M. Lasance. "Two benchmarks to facilitate the study of compact thermal modelling phenomena," IEEE Transactions on Components and Packaging Technologies, vol. 24, no. 4, pp. 559–565, 2001.

[16] M. N. Sabry. "Compact thermal models for electronic systems" IEEE Transactions on Components and Packaging Technologies, vol. 26, no. 1, pp. 179–185, 2003.

[17] E. G. T. Bosch. "Thermal compact models: An alternative approach," IEEE Transaction on Components and Packaging Technologies, vol. 26, no. 1, pp. 173-178, 2003.

[18] A. Krum. "Thermal management". In F. Kreith, editor, The CRC handbook of thermal engineering, pp. 2.1–2.92. CRC Press, Boca Raton, FL, 2000.

[19] K. Skadron, T. Abdelzaher, and M. R. Stan. "Control theoretic techniques and thermal-RC modeling for accurate and localized dynamic thermal management," In Proc. HPCA-8, pp. 17–28, 2002.

[20] N. Rinaldi, "Thermal analysis of solid-state devices and circuits: an analytical approach," Solid-State Electronics, vol. 44, pp. 1789-1798, 2000

[21] Recore systems. www.recoresystems.com.