Temperature and Fast Voltage On-Chip Monitoring using Low-Cost Digital Sensors
Lionel Vincent, Philippe Maurine, Edith Beigné, Suzanne Lesecq, Julien Mottin

To cite this version:
Lionel Vincent, Philippe Maurine, Edith Beigné, Suzanne Lesecq, Julien Mottin. Temperature and Fast Voltage On-Chip Monitoring using Low-Cost Digital Sensors. VARI: Workshop on CMOS Variability, Sep 2013, Karlsruhe, Germany. hal-01067982

HAL Id: hal-01067982
https://hal.archives-ouvertes.fr/hal-01067982
Submitted on 24 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Temperature and Fast Voltage On-Chip Monitoring using Low-Cost Digital Sensors

Lionel Vincent∗, Philippe Maurine†, Edith Beigné∗, Suzanne Lesecq∗ and Julien Mottin∗
∗CEA - LETI, MINATEC Campus, 17 rue des martyrs 38054 Grenoble Cedex 9, France
†LIRMM, Université Montpellier 2, 161 rue Ada 34095 Montpellier Cedex 5, France

Abstract—Power efficiency of embedded systems is nowadays a challenge to reach the increasing performance requirements while limiting the power consumption. Adaptive architecture allows to adapt the operating point of each power domain of a MultiProcessor System-on-Chip using independent Dynamic Voltage and Frequency Scaling techniques. Furthermore the optimal operating point of each power domain is dependent on the current variability state, thus it appears essential to monitor locally and on-line the environmental variability. To limit the power consumption and hardware overhead the monitoring system must have the lowest cost and the best possible efficiency. This paper presents a monitoring system based on two complementary methods to estimate dynamically the local conditions using low-cost digital sensors. A fast voltage drop monitoring method is performed every 0.66µs and a low-cost calibration method is also proposed based on the Voltage-Temperature state of the platform at startup.

Index Terms—AVFS, calibration, digital sensors, low-power architecture, monitoring, temperature, variability, voltage drop

I. INTRODUCTION

Today mobile applications require high computational performances under strong power consumption constraints to limit the chip thermal elevations and save battery lifetime. During a long time the power consumption decrease has followed the technological node. But this trend was slow down with the arrival of sub-micrometer technologies and now the leakage power consumption becomes an important part of total power consumption of nanometer MOS transistor. Technology scaling was actually more beneficial for speed performances than for power consumption. Indeed the integration on a single chip of many cores increases the MultiProcessor System-on-Chip (MPSoC) performance because of its parallel computational capabilities. Nevertheless, this architecture is a promising solution to locally manage the power consumption by adapting the "parameters" (e.g. supply voltage, clock frequency) of each block of the SoC using Dynamic Voltage and Frequency Scaling (DVFS) techniques.

The supply voltage decrease and technology scaling have as side effect an increase of the intra-die variability. As a consequence, identical cores in a MPSoC will not behave in the same way, i.e. their characteristics and performances are different. The Process (P) variations, due to production deviations, will affect statically the global and local characteristics of the chip [1]. The variability is also dynamic because of supply voltage (V) and temperature (T) changes. This environmental variability is correlated with the chip activity [2] and its topology [3]. Fastest events can appear in some µs and affect both the timing (i.e. the performance) and the power consumption of the circuit. Therefore, the energy efficiency can only be optimized if the voltage and temperature dynamic variations are monitored and mitigated. These variations are not homogeneously distributed on the chip, thus local mitigating strategies must be developed. Globally Asynchronous Locally Synchronous (GALS) architectures [4] are well adapted for such purpose because of their island structure as a local policy (e.g. a DVFS technique) can be applied in each independent Voltage-Frequency Island (VFI).

Fig. 1: Principle of an Adaptive Voltage and Frequency Scaling architecture. Each black square represents a sensor.

Classical DVFS approaches reduce the power consumption by adapting the VFI parameters according to performance constraints. However, the current variability state is not monitored and design margins are necessary to ensure the VFI functionality (e.g. no timing fault). The trim of the design margins as well as an improvement of the energy efficiency is possible thanks to the implementation of a local feedback control loop.

Adaptive Voltage and Frequency Scaling (AVFS) architectures [5] offer an appealing solution to locally monitor and mitigate the dynamic variability. Fig. 1 presents its principle, locally implemented in a VFI. The Local Control box provides settings to the actuators so that the performance and functionality constraint from the Global Control are satisfied. Sensors (black squares) buried in the monitored Processing Element (PE) provide measurements dependent on the current variability state. A VT Estimation Method extracts from the sensors...
readings an estimate of the PE current Voltage-Temperature (VT) state. Finally, the settings of the actuators are tuned according to the Policy of Local Adjustment to reach the most energy efficient VF operating point. This adjustment can be indeed implemented with a re-adaptation loop, the ultimate goal being to decrease the power consumption while maintaining the performances. Note that the AVFS architecture can be implemented with different granularities (e.g. core, cluster) according to the trade-off between the energy efficiency and the hardware (sensor, actuators, control logic) overhead.

The main challenge in the development of such an architecture is the design of a low-cost and non-invasive monitoring system using integrated sensors. Uni-sensitive sensors provide "direct" measurement of one quantity of interest, e.g. voltage [6] or temperature [7], with reasonable accuracy, regarding the readaptation objective. However, their analog design is a drastic limitation because of their integration overhead (large silicon area). Multi-sensitive sensors seem a good alternative that can be duplicated along the chip. Indeed, simple and low-cost digital structures as Ring Oscillators (ROs) provide a frequency dependent on the variability state [8]. Actually, the RO oscillating frequency depends in a complex way on the PVT parameters. Thus, a direct reading of V or T is not possible from a single frequency measurement. Therefore, a set of different ROs together with an VT Estimation Method can be used to estimate the VT state.

In this context, this paper presents a monitoring method to estimate dynamically the Voltage and Temperature state using low-cost digital sensors. Actually two compatible and complementary methods are described to estimate either jointly the VT state or the fast voltage variations.

The rest of the paper is organized as follows. First the VT estimation method, based on the Kolmogorov-Smirnov hypothesis test, is summarized in section II. The VT estimation computational time and its associated maximum reachable throughput are discussed. Then a fast voltage estimation method is proposed in section III to monitor the fast IRdrops. Section IV is dedicated to calibration, which is mandatory to accurately estimate the VT state. Finally a conclusion and some perspectives are given in section V.

II. VT ESTIMATION METHOD OVERVIEW

The estimation of VT from low-cost digital sensors is very appealing to monitor at fine grain the dynamic variability. The use of a set of goodness-of-fit hypothesis tests to estimate VT from the reading of a set of ROs is now summarized [8]. Then throughput considerations are discussed.

A. VT estimation methodology

The Multiprobe consists of 7 different compact ROs and a counter to measure the output frequencies. It can be duplicated along the chip with low design and silicon costs (silicon area of 450µm² in CMOS 32nm STM technology). Each RO has been designed to behave differently from the others with regards to VT variations.

![Fig. 2: Principle of the proposed VT Estimation Method](image)

The monitoring strategy consists in first promptly activating the Multiprobe to avoid auto-warming and limit its power consumption overhead. Then, the VT state is estimated from the 7 oscillating frequencies thanks to a fusion algorithm based on a set of Kolmogorov-Smirnov (KS) goodness-of-fit tests [8].

The KS test [9] is a non-parametric hypothesis test. For two empirical samples of size n, it estimates if both samples come from the same distribution law. The maximum gap between the Cumulative Distribution Functions CDFm and CDFi is first computed:

\[ D_t = \sup_x |CDF_m(x) - CDF_i(x)| \]  

Then, the probability \( p_t \) (p-Value) that \( CDF_i \) and \( CDF_m \) come from the same distribution is given by:

\[ p_t(\lambda) = 2 \sum_{k=1}^{+\infty} (-1)^{k+1} e^{-2k^2\lambda^2} \]  

with \( \lambda = \sqrt{n} \cdot D_t \)

The principle of the proposed algorithm based on the KS Test, is to compare the current measurements with several models stored in memory. Fig. 2 depicts the principle of this VT Estimation method, which can be extended to several Multiprobes in a VFI. The Estimation block computes the estimated \( \{V, T\} \) state. Here, one KS Test evaluates if \( CDF_m \) measurement and \( CDF_i \) (a stored model) are similar and thus corresponds to the same \( \{V_i, T_j\} \) conditions. The CDFs are computed by the CDF Builder respectively from the Multiprobe measurements \( \hat{F}_{V,T} \) and a stored model \( \hat{F}_{V,T} \) corresponding to the condition \( \{V_i, T_j\} \). The Models Database stores \( M \) vectors \( \hat{F}_{V,T} \) acquired during the calibration phase (see section IV) at state \( \{V_i, T_j\} \). The KS Test block runs the test for \( CDF_m \) and each \( CDF_i \). Then the \( p_t \) are collected in \( \hat{P} \in \mathbb{R}^M \). Actually, \( \hat{F}_{V,T} \) (corresponding to the VT state experienced by the Multiprobe) is seldom recorded in the database. Therefore, s CDFs in the subset \( CDF_m \) that best match \( CDF_i \) are kept in the Aggregation box. Then,
The VT estimation computational time and its associated overhead (evaluated to some kB), complexity of the calibration models stored in the database, which influences the memory access, i.e. a variation of 3.7mV or 6.2°C will not influence the frequency of more than 3%. Thus, the estimates seem accurate enough to implement an efficient adaptation loop.

Note that these figures highly depend on the number of models stored in the database, which influences the memory overhead (evaluated to some kB), complexity of the calibration phase and the maximum estimation throughput. This latter is directly linked to the range of dynamics that can be monitored. The VT estimation computational time and its associated maximum reachable throughput are now discussed.

B. Throughput consideration

In the previous section, the results have been obtained with a very fine database granularity. Actually, a coarser granularity of ΔV = 10mV and ΔT = 20°C provides mean estimation errors of 4.3mV and 7°C respectively. These accuracies are sufficient for the monitoring and re-adaptation loop envisioned in the present study. Furthermore, the temperature range has been limited to T ∈ [0, 100]°C, considering mobile applications, leading to 366 models in the database.

The VT estimation method has been implemented in software to evaluate its computational cost. The testbench considered is an AVFS architecture with 4 STMicroelectronics xP70 processors in CMOS 32nm [10]. The additional hardware needed for the AVFS architecture (sensors, actuators) is already available in each processor. The VT estimation method has been written in C language and executed on one of the processors. Experiments have shown that a single KS test requires 2500 cycles in average (included memory access), i.e. the test of a single CDF is performed in 4.2µs at 600MHz.

Thus it appears unrealistic to use a software implementation to monitor either temperature or voltage variations of respectively some tens µs and some µs.

The estimation throughput can be improved with the implementation of a hardware accelerator. The logic block has a complexity equivalent to 9 kgates and allows to perform a single KS test in 42 cycles in average, i.e. an acceleration factor of 60 with respect to software solution. Thus a VT estimate is performed in 25µs at 600MHz.

If this throughput is sufficient to monitor the temperature variations, it solely allows to monitor the slower voltage ones. To overcome this limitation and to be able to track fast voltage variations of some µs, it is mandatory to develop an additional estimation method. In the next section, this fast voltage monitoring method is presented.

III. FAST VOLTAGE MONITORING

The VT estimation method presented in the previous section is able to deliver a VT estimate every some tens of µs. It is a good throughput to monitor temperature variations but it is not sufficient to track fast voltage drops of some µs. The effects of voltage changes on maximum reachable frequency become more and more important for low voltage circuits (up to 0.7%/mV), thus introducing large voltage margins in the adaptive loop if they are not rapidly tracked. In this section a fast voltage monitoring method, complementary with the VT estimation method presented in II, is proposed. It aims to accelerate the V estimation to allow a better IRdrop monitoring and then limit the voltage deviations in the AVFS loop.

A. Dynamics disparities

The principle of the proposed fast voltage monitoring method is based on the difference of timing dynamics between voltage and temperature variations. Indeed fastest temperature variations are about some hundreds of µs [11] while the fastest voltage ones can reach some ns [12]. Even if it is commonly considered that the fastest voltage variations can not be monitored and mitigated, it seems realistic to track the voltage variations of some µs.

To be able to monitor fast voltage variations using the Multiprobe sensor it is necessary to pose an hypothesis: if the measurement throughput provided by the sensor is
fast enough, then the temperature can be considered constant during the measurement time. Indeed if the temperature is constant during measurement, a variation of the RO frequency is then necessarily induced by a voltage one. As a consequence it appears possible to monitor voltage variation using a single RO of the Multiprobe sensor, if the measurement time is shorter than those of significant temperature variation, i.e. some tens of µs.

In the previous VT Estimation Method the seven ROs of the Multiprobe was required to estimate jointly the voltage and temperature. While the frequencies of each RO was measured one by one because of the sensor architecture, the related measurement time was about 2.7µs. Now, using a single RO, the measurement time can be reduced to 0.66µs. The measurement throughput using the two different methods are depicted on Fig. 3. Accordingly, the key hypothesis presented above, of constant temperature during the time measurement, is validated.

B. Complementarity of the two estimation methods

The frequency of a RO is a complex non-linear function of its PVT state. As seen above the temperature (and process) are supposed constant during the measurement time. Nevertheless, to convert the frequency of a RO into a voltage estimation, it is necessary to know the current temperature of the considered RO. An estimate \( \hat{T} \) of the current temperature can be provided by the VT Estimation Method. Then the complementarity of the two estimation methods appears to be very interesting.

This complementarity allows to monitor two different dynamic variations (V and T) at two different timing granularities. It could be envisioned to run several Fast Voltage monitoring Methods between two VT Estimation Methods. However each run of the Fast Voltage monitoring Method should be executed in a reduce time from the last VT Estimation Method. If this constraint is respected, Fast Voltage monitoring Method executions can be performed by considering that the current temperature is equal to the last \( \hat{T} \) estimated by the VT Estimation Method. With the described measurement sequence the voltage is tracked quickly using the two methods while the temperature is monitored at a coarser grain by the VT Estimation Method. Practically the voltage can be estimated by identifying a current measurement of frequency of a RO \( F_{RO}^{m} \) to the frequencies \( F_{RO}^{V,T} \) stored in the Models Database by considering the temperature \( \hat{T} \) is currently known from an execution of a previous VT Estimation Method. The nearest frequency \( F_{RO}^{V,T} \) to the measured one \( F_{RO}^{m} \) corresponds to the current voltage condition. Then \( \hat{V} \) is deduced from (6), while the frequency of the ROs is a monotone function with respect to \( V \) for a fixed \( \hat{T} \):

\[
\hat{V} = \min_{V} |F_{RO}^{V,T} - F_{RO}^{m}| \tag{6}
\]

C. Validation

To evaluate the accuracy of the presented Fast Voltage monitoring Method, various voltage estimations have been performed, with Matlab simulations, for many different VT states along the voltage and temperature ranges considered section II. From this experiment are obtained respectively, a mean absolute voltage estimation error and a mean voltage estimation error standard deviation error of :

\[
|\epsilon_{V}| = 3.9mV, \quad \sigma_{\epsilon_{V}} = 3.7mV \tag{7}
\]

This accuracy is very close to the one described section II using the VT Estimation Method. However the accuracy of the Fast Voltage monitoring Method is very sensitive to temperature uncertainties because of the constant temperature hypothesis.

While the Fast Voltage monitoring Method proposed is a simple search of minimum in a set of frequencies, its implementation is very simple. To evaluate the throughput of a such method it has been implemented in C language on the same AVFS demonstrator test-bench as described section II-B and by using the same Models Database with 366 models.

A comparison of two frequencies \( F_{RO}^{V,T} \) and \( F_{RO}^{m} \) can be
performed in 318 cycles on average, including the memory access. Then at 600MHz, a \( \hat{V} \) estimate can be obtained every 32\( \mu \)s using a software version of the Fast Voltage monitoring Method.

This throughput is not sufficient to validate the method but a hardware implementation could achieve the required timing performance. It has been evaluated that a fast voltage estimation could be performed in 0.25\( \mu \)s at 600MHz using a hardware accelerator. The maximum throughput of the Fast Voltage monitoring Method is thus limited by the measurement of the RO. The throughput of the two estimation methods can be compared on Fig. 3. Note the Fast Voltage monitoring Method is almost 30 times faster than the VT Estimation Method that allows the complementarity of the two methods.

To validate the complementarity of the two proposed methods, several Fast Voltage monitoring Methods have been executed between two successive VT Estimation Methods. The Fig. 4 depicts the voltage (Fig. 4a) and temperature (Fig. 4b) tracking using the two methods and considering the respective estimation throughput described on Fig. 3. The blue dots represents the current voltage and temperature state of the sensor. The red dots are the voltage and temperature estimations using the VT Estimation Methods run 11 times. The red circles are the results of the Fast Voltage monitoring Methods run 9 times between two VT Estimation Methods. The tracking is quite good with voltage and temperature mean absolute errors of respectively 4.4\( mV \) and 3.1 \( ^\circ \)C.

The performance proposed by the Fast Voltage monitoring Method allows the mitigation of fast voltage drops in the AVFS loop presented in I. However, the two presented methods needs the Models Database containing the frequencies of the ROs stored as models during a calibration phase. This essential calibration step will be described in the next section.

IV. CALIBRATION

The calibration phase aims to build the Models Database needed to perform the two proposed estimation methods. The calibration phase is needed to take into account intra-die process variations in order to be able to estimate VT conditions accurately and with absolute quantitative values. In this section a Calibration Method based on process characterization of the Multiprobes is presented and validated. It is adapted to the proposed estimation methods and presents a good accuracy/complexity trade-off.

A. Accuracy/complexity trade-off

A Calibration Method has to deal with a trade-off between accuracy and complexity of post-silicon test platform. To obtain a very accurate Models Database, complex and expensive methods are required. The theoretical ideal Calibration Method would be to impose many various voltage and temperature values to the circuit. By measuring and storing the frequencies of the ROs of the Multiprobe for each different \( \{V, T\} \) calibrated situations, the obtained Models Database will perfectly match the current process corner. However it is very costly to build a such fine grain Models Database because voltage and temperature can not be easily adapted and controlled without costly test hardware.

At the opposite, a no-calibration method can be envisioned. As seen above, the calibration phase is needed to obtain absolute quantitative values of estimation, but it can be envisioned to use a Models Database obtained from Spice simulation of the sensor for a typical process corner. While the real process state of the circuit is surely different that those of the obtained Spice Models Database (typical), the estimation will not be accurate. The related accuracy loss induced in the VT Estimation Method is very important for the voltage estimation but it is quite limited regarding the temperature estimation because of the linear dependency of the ROs behavior with respect to the temperature.

The proposed Calibration Method presented in the next section is a trade-off between the two previously presented ones.

B. Proposed Calibration Method

The proposed Calibration Method builds a Models Database by correcting Spice simulations using a unique VT calibration measurement. The principle is to adjust the Spice Models Database obtained in a theoretical Typical (TT) corner to reach the current process corner. The Spice Models Database is obtained during the chip design phase from simulation of the sensor. This calibration is performed offline and only at the chip start-up. Then the corrected Models Database is stored in memory to be used by the two presented methods at run-time.

It firstly requires to characterize the current process state of the considered Multiprobe. Usually some post-fabrication process sensors are implemented in complex circuit to monitor their current real process state. They provide an information about the global process corner of the circuit, but they are not sufficient to monitor the intra-die process variations. The current process corner of a Multiprobe can be determined by measuring the frequencies of its ROs at start-up. The calibration state corresponding to these measurements is
\( \{V_{\text{calib}}, T_{\text{calib}}\} = \{V_{\text{nom}}, T_{\text{ambient}}\} \) and it is known while \( V_{\text{nom}} \) is the initial voltage applied to the chip and \( T_{\text{ambient}} \) is the current temperature of the room. A process characterization can be performed for each RO of the considered Multi-probe by comparing the calibration measurements \( F_{\text{calib}} = \frac{F_{\text{current}}^{V,T}}{F_{\text{calib}}^{V,T}} \) at the corresponding Typical models of the RO \( F_{\text{calib}}^{V,T} \) stored in the Spice Models Database. Fig. 5 presents the principle of the proposed Calibration Method for a single RO at \( T = T_{\text{calib}} \). The ratio \( r \) represents the process characterization of a considered RO of the Multi-probe as described above.

It is possible to adjust the Spice typical Models Database using the process characterization \( r_{k} \), \( k \in 1,7 \) of each RO of the Multi-probe. To build the Models Database \( F_{\text{current}}^{V,T} \) corresponding to the current process state, \( F_{\text{calib}}^{V,T} \) is adjusted by a correcter function \( f_{\text{cor}} \) (dotted arrow) as described in Fig. 5. This function represents the way to pass from the Typical corner (in green in Fig. 5) to another extreme one (Slow for the presented example in blue in Fig. 5). This function has to be characterized during the chip design phase, from sensor Spice simulation, for the two extreme (Slow-Slow and Fast-Fast) process corner. The correcter function \( f_{\text{cor}} \) has been identified by a least squares errors minimization as a polynomial of order 8 in V and 4 in T. On Fig. 5 only the frequencies corresponding to the Slow \( (r^{SS})_{V,T} \) in blue) and Typical \( (r^{TT})_{V,T} \) in green) corners, obtained during chip design characterization, are represented. The correction applied to \( F_{V,T} \) at each \( \{V_{i}, T_{j}\} \) point is \( r_{i,j} f_{\text{cor}}(\{V_{i}, T_{j}\}) \) such that for a considered RO:

\[
F_{\text{current}}^{V,T}(V_{i},T_{j}) = F_{\text{calib}}^{V,T}(V_{i},T_{j}) \cdot r_{i,j} f_{\text{cor}}(\{V_{i}, T_{j}\}) \quad (8)
\]

This computation is performed for each RO of the considered Multi-probe. This Calibration Method allows to take into consideration the real local process state of each Multi-probe in the Models Database building.

C. Validation

The proposed Calibration Method has been implemented on the hardware platform presented section II-B. The calibration sequence is executed as follows for each Multi-probe:

- Measurements of ROs frequencies at the chip start-up
- Computation of ratio \( r \) (process characterization) for each RO of the Multi-probe
- Loading of the Typical models from Spice simulation \( F_{\text{calib}}^{V,T} \) and the correcter function \( f_{\text{cor}} \)
- Building of the \( F_{\text{current}}^{V,T} \) Models Database using (8)
- Uploading in memory the calibrated Models Database

Two different data files are needed \( (F_{\text{calib}}^{V,T} \) and \( f_{\text{cor}} \) to preform the proposed Calibration Method. They are stored in a non-volatile memory (in the considered test-bench it is an external memory) and these files have to be loaded in local memory for the processing phase of the Calibration Method. For the considered example of 366 models the total memory size of the files is about 24kB and those of the generated Models Database is about 21kB.

Furthermore for the considered platform and the desired Models Database granularity, this Calibration Method has been performed in software, at the chip start-up, in some 225\( \mu \text{s} \) per Multi-probe at 600MHz, including measurements of the frequencies of the ROs and the loadings of the files.

V. CONCLUSION

A complete monitoring framework based on two main complementary methods has been presented to track dynamic environmental variations. A simple digital sensor based on a set of different ring-oscillators provides the variability measurements. An efficient and low-cost calibration method of this sensor has been proposed. The monitoring framework permits to cover a large range of different orders of dynamic variations in order to mitigate their effects efficiently in future works.

On-going works focuses on the development of re-adaptation strategies to tune the settings of the actuators with respect to the variability monitoring and the performance constraint, in order to reach an optimal energy efficiency.

VI. ACKNOWLEDGMENTS

This work has been done in the context of the BENEFIC project labeled within the framework of CATRENE (CA505).

REFERENCES