Temperature and Process-Aware Performance Monitoring and Compensation for an ULP Multi-Core Cluster in 28nm UTBB FD-SOI Technology

Alfio Di Mauro†, Davide Rossi*, Antonio Pullini†, Philippe Flatresse‡, Luca Benini*†
*DEI, University of Bologna, Via Risorgimento 2, 40136 Bologna, Italy
†Integrated Systems Laboratory, ETHZ, Gloriastr. 35, 8092 Zurich, Switzerland
‡STMicroelectronics, 850 Rue Jean Monnet, 38920 Crolles, France

***Abstract*—Environmental temperature variations, as well as process variations, have a detrimental effect on performance and reliability of embedded systems implemented with deep-sub micron technologies. This sensitivity significantly increases in ultra-low-power (ULP) devices that operate in near-threshold, due to the magnification of process variations and to the strong thermal inversion that affects advanced technology nodes. Supporting an extended range of reverse and forward body-bias, UTBB FD-SOI technology provides a powerful knob to compensate for such variations. In this work we propose a methodology to efficiently compensate, at run-time, these variations. The proposed method exploits on-line performance measurements by means of Process Monitoring Blocks (PMBs) coupled with on-chip low-power Body Bias Generators. We characterize the response of the PMBs versus the maximum achievable frequency of the system, deriving a predictive model able to estimate such frequency with an error of 3%. We apply this model to compensate Temperature-induced performance variations, estimating the maximum frequency with an error of 7%; we eliminate the error by adding an appropriate body-bias margin resulting in a worst case global power consumption overhead of 5%. As further improvement, we generalize the methodology to compensate also process variations, obtaining an error of 28% on the estimated maximum performance and compensating this error with an overhead of 17% on the global power consumption.**

I. Introduction

Internet of Things, e-Health, Smart Sensors and wearable consumer gadgets are expected to drive the electronic market of the next decades. These applications rely on the capability of the research community to provide devices that couple ultra-low power (ULP) behavior with a reasonable level of performance. Indeed, these applications are characterized not only by an increasingly tighter power budget, but also by an increasing demand of computation capabilities. The pace dictated by the Moore's law has slowed down, and CMOS scaling, which drove semiconductor growth during the past decades, is delivering reduced energy gains [1]. In this Moore's law twilight era, further energy gain can be achieved by moving to the near-threshold computing (NTC) domain [2]. However, electronic devices implemented with most advanced technological nodes feature a strong dependency between the ambient temperature and the operating frequency and leakage power. Unfortunately, this dependency increases in the near-threshold operating region, where ULP devices works to provide high energy efficiency [3][4][5, p. 69].

The effects of process variations require a one-time compensation as soon as the chip is fabricated [1]. Contrarily, thermal variations are dependent on the operating environment, hence, they require a run-time compensation to ensure that chips are able to maximize the efficiency and at the same time guarantee a given target performance.

A well-known approach for achieving post-silicon tuning to compensate variations in circuits is through body biasing [6]. As opposed to adaptation of supply voltage, that requires on chip DC/DC converters or voltage regulators, modulation of body biasing can be implemented with simpler and more efficient circuits, as relatively small transient currents are required to polarize the substrate of the devices [7] [8]. Exploiting forward body biasing (FBB), it is possible to increase the operating frequency of a device avoiding significant growth of dynamic power. This process makes FBB suitable for compensation of variations causing degradation of the operating frequency. On the other hand, reverse body biasing (RBB) allows to significantly reduce leakage power where the process or operating conditions allow the devices to run faster than the target frequency, but with excessive leakage power [9].

Similarly, temperature compensation can be achieved through voltage scaling and body biasing [10]. The majority of the works focusing on temperature effects target high performance devices that are subject to self-heating [11]. These devices necessitate the mitigation of the effects of temperature variation to avoid excessive leakage power dissipation that may lead to thermal runway. However, these works demonstrated the ability to compensate only for relatively small temperature ranges, not representative of the huge amount of scenarios enabled by ULP applications. Indeed, although self-heating does not affect ULP devices due to their extremely small power consumption, compensation of ambient temperature is crucial in this domain as well [12].

In most advanced CMOS technologies, thermal inversion causes an exponential growth of the dependency between

[1]Performance of digital circuits can also be affected by aging. However, since self-heating is minimal for systems operating in ULP domain, this phenomenon can be neglected.

temperature and frequency, especially when operating in near-threshold. Several works have addressed this problem, mainly leveraging adaptation of supply voltage [13][14]. On the contrary, the adoption of body biasing to address variation of ambient temperature has not been extensively explored so far, due to the limited capabilities of most advanced technologies such as bulk and FinFET to provide extended body bias ranges suitable for compensation of wide temperature ranges [15][16].

The methodologies proposed so far are very often highly intrusive and linked to a specific architecture. Solutions like the one proposed in [17] have to be fully integrated with the architecture to provide good results. Less intrusive solutions based on Critical Path Replicas (CPR) have been proposed [18], but also in this case, the architecture's dependency of the performance monitoring system reduces the application field to designs where a complete access to all the SoC's IPs is possible. Contrary to what has been done so far, the performance monitoring system adopted in our solution keeps at minimum the intrusiveness, by exploiting on-chip direct frequency measurements. High generality is given by the fact that the knowledge of architectural details is not required, which is extremely useful when the IPs composing the system are released as encrypted macros, or simply come as hardened macros from different design teams.

In this work, we analyze the capabilities of UTBB FD-SOI (Ultra Thin Body and Box Fully Depleted Silicon On Insulator) technology in compensating process variations in the ULP domain by means of on-chip Process Monitor Boxes (PMB) coupled with on-chip Body-Bias Generators (BBG). To demonstrate this potential on real silicon we utilize a test vehicle consisting of an ULP cluster of processors operating in near threshold fabricated in 28nm UTBB FD-SOI technology. The compensation technique exploits one of the unique features of the UTBB FD-SOI technology: the ability to use the wide-range body bias to change the transistors voltage threshold. Body-Bias Generators allows to modulate the V_{bb} in a wide range, theoretically from $-3\,\mathrm{V}$ to $V_{dd}/2$ + $300\,\mathrm{mV}$.

The main contribution introduced by this work is a general methodology, suitable to be implemented both in software and hardware, which enables advanced power management strategies. We start our work by studying the correlation between the output of the PMB and the actual maximum performance of the chip on the whole body-bias range; technical limitations practically reduce the body-bias range (from $-1\,\mathrm{V}$ to $V_{dd}/2$ + $300\,\mathrm{mV}$). As successive step, we derive a general device-dependent linear model to predict the maximum frequency at each operating point. Our exploration covers the following set of OPs: V_{dd} = $\{0.5\,\mathrm{V}$,$0.7\,\mathrm{V}$,$0.9\,\mathrm{V}\}$ and for each supply voltage T = $\{-20\,^{\circ}\mathrm{C}$,$25\,^{\circ}\mathrm{C}$,$80\,^{\circ}\mathrm{C}\}$. Once the model's parameters have been obtained, we estimate the maximum error, finding that it is never higher than the 7%. Then, we propose a methodology exploiting additional body-bias voltage margins to reduce it. Finally, we provide an evaluation of the overhead that this methodology causes on the global power consumption; we quantify it as the ratio between the additional leakage caused by the FBB margin and the global power consumption. The main result of this work is to show that temperature-induced variation compensation can be done by increasing the power consumption budget by less than 10%. We derive, as last step, a more general methodology to compensate also process variations, with a worst-case overhead lower than 17% in practical use case.

Remainder of this paper is organized as follows. Section II describes the chip used as test vehicle and the Process Monitor Boxes. Section III describes the experimental setup, as well as the measurements performed to characterize the devices and extract the empirical models for the Frequency and Leakage current. Section IV provides a description of the analysis performed on the PMB sensors. Section V describes the methodology used to compensate temperature variations. Section VI provides a description of the same methodology applied to the process variations compensation. Section VII introduces a discussion of the results achieved by this methods, as well as a discussion of the overhead introduced by the two different compensations. Finally, section 6 gives some concluding remarks.

II. PULP PLATFORM

Parallel Ultra-Low-Power platform [19] version 3 (PULPv3) [9] is a multi-core SoC for ULP applications that operates in near-threshold to achieve extreme energy efficiency on a wide range of operating points. The SoC is built around a cluster featuring four cores and $64\,\mathrm{kbyte}$ of L2 memory. The cores are based on a highly power optimized micro-architecture implementing the OpenRISC-32bit ISA featuring 1kB of private instruction cache each. The cores do not have private data caches, avoiding memory coherency overhead and increasing area efficiency, while they share a L1 multi-banked Tightly Coupled Data Memory (TCDM) acting as a shared data scratchpad memory. The TCDM features 8 2kB SRAM banks connected to the processors through a single clock latency non-blocking interconnect, implementing a word-level interleaved scheme to minimize banking conflict probability. Off-cluster (L2) memory latency is managed by a tightly coupled DMA featuring private per-core programming channels, ultra-low programming latency and lightweight architecture optimized for low-power and high transfer efficiency. In Fig.1 is shown the architecture of the system.

Advanced power management is enabled by three different power domains: *i)* The "Safe Voltage Domain") hosting the Frequency Locked Loop generators, the two Body-Bias Generators for the SoC and Cluster regions, the PMB Controller and additional infrastructural control logic *ii)* The "SoC Body-Bias Domain" *iii)* The "Cluster Body-Bias Domain". Each domain is monitored by a PMB, which will be described in the next section. In our tests, we will focus only on the Cluster Domain, applying V_{bb} = $0\,\mathrm{V}$ to the other body-bias domains.

A. Process Monitor Boxes

A Process Monitor Box (PMB) can be seen as an on-chip sensor based on ring oscillators, connected to the system through a digital interface, which provides information about the maximum frequency achievable by the device. These sensors are designed and optimized to behave consistently with the other library logic gates, attempting to emulate all the parasitic effects induced by temperature's and process' variations.

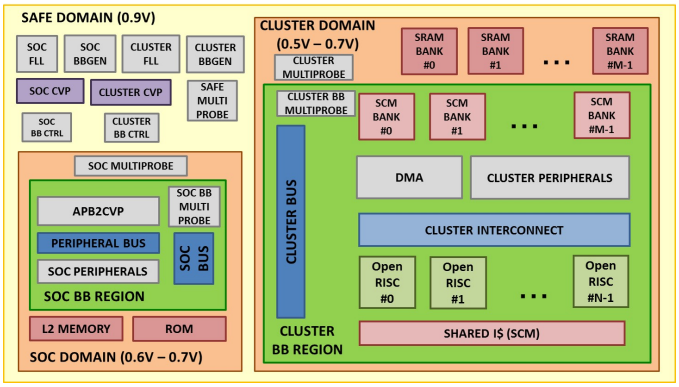

Fig. 1: Architecture of PULP system.

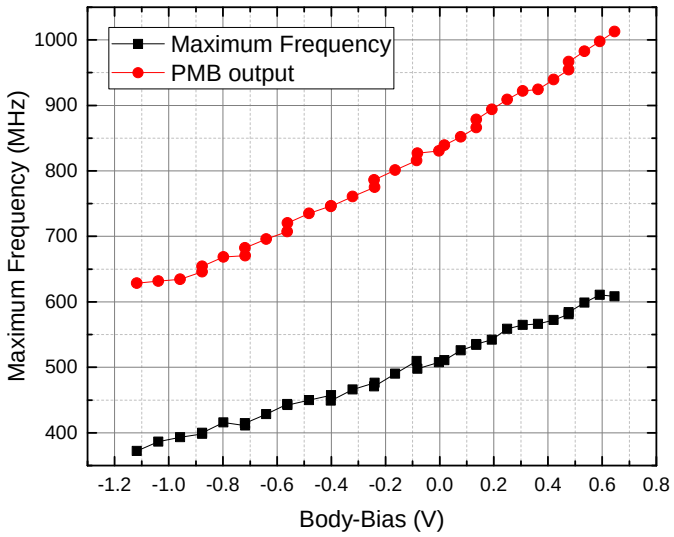

Fig. 2: This plot compares the maximum achievable frequency and the PMB frequency estimation for a typical chip at $0.9\,\mathrm{V}$ and $25\,^{\circ}\mathrm{C}$.

III. Experimental Setup

To compare the ambient temperature and process variations on next generation ULP devices, we conducted a large data acquisition test directly on the PULP prototypes, measuring power and performance metrics at the different corner cases. Based on the measurement results, we have derived a set of empirical models to describe the dependency of the performance metrics to the physical parameters, compensation knobs and environmental conditions. This models will be discussed in the next sections of the paper.

The measurements described are performed with an Advantest SoCV93000 tester system, in connection with a Thermonics 2500E temperature forcing system, able to force an environment temperature ranging from $-80\,^{\circ}\mathrm{C}$ to $220\,^{\circ}\mathrm{C}$. Since we are interested in using the body-biasing voltage as independent variable both for temperature and process variations compensation, we structured the measurements as follows: *i)* We defined the operating point (OP) in terms of {Supply voltage V_{dd} , Temperature}, where V_{dd} = {0.5 V,0.7 V,0.9 V} and T = $\{-20\,^{\circ}\mathrm{C},25\,^{\circ}\mathrm{C},80\,^{\circ}\mathrm{C}\}$; *ii)* For each OP we swept the body-biasing voltage (V_{bb}) in the range $-1\,\mathrm{V}$ to $V_{dd}/2 + 300\,\mathrm{mV}$ using the Body-Bias Generator, and we measured: the leakage power (P_{LKG}), the active dynamic power (P_{DYN}), the total power (P_{TOT}) and maximum frequency (F_{MAX}) achievable by the device. More specifically, we measured the active dynamic power (P_{DYN}) as the difference between the Total Power (P_{TOT}) and the leakage power (P_{LKG}). The Total Power (P_{TOT}) has been measured as the power consumption when the device is executing an arithmetic loop (Matrix Multiplication), and the Leakage Power (P_{LKG}) as the power consumption when the device is not clocked. We extracted the maximum operating frequency (F_{MAX}) by means of a carefully crafted benchmark, able to trigger the most critical paths[2] of the circuit. We verified that the result of the arithmetic operation was returned with the correct timing, and the End-Of-Computation[3] signal was properly asserted. As cross-check, we verified that the arithmetic loop returned a valid check-sum.

To collect, the PMB output data during the test phase, since the result of the PMB's Maximum frequency estimation was temporarily stored in an internal status register of the system, we sent the value to a computer through a standard UART interface.

IV. PMB Characterization

The first part of the methodology proposed in this paper has the aim to exploit the run-time maximum frequency measurement, performed with the test equipment, to determine the actual maximum frequency of the chip. In parallel, by observing the PMBs' output, it is possible to find a correlation between the actual maximum frequency and the one predicted by the PMBs. Once the equation to translate the PMB frequency estimation in F_{Max} has been found, given a target frequency specified by the application and depending on the V_{dd}, the body-bias voltage can be finely regulated to fill the performance's gap.

For the sake of simplicity, in this section we will not consider the variations introduced by the process; accordingly with this assumption, we assume to use a model calibrated on the specific chip. In other words, the correlation between the value provided by the PMB sensor and the maximum achievable frequency of the chip is calculated for the specific device under test. Fig.2 shows on the same plot the maximum frequency measured with the test equipment and the PMB estimation for a typical device at $0.9\,\mathrm{V}$ and $25\,^{\circ}\mathrm{C}$

As shown in Fig.3, the relationship between the values provided by the PMB and the maximum frequency of the device can be modeled with a simple affine function[4] (Equation 1).

$$F_{max} = C_{corr} F_{pmb} + F_0 \tag{1}$$

[2]The critical path, identified by the timing analysis in the communication between the cores and the scm memory, was massively triggered by the algorithm we used as benchmark to measure the maximum frequency.

[3]Physical output pin of the device which certifies that the system completed all the operations and properly entered a known final state.

[4]The C_{corr} parameter represents the ratio between the delay of the critical paths and the minimum delay determining the frequency of the oscillator inside the PMB. The F_0 parameter is an offset related to the critical path of the circuit.

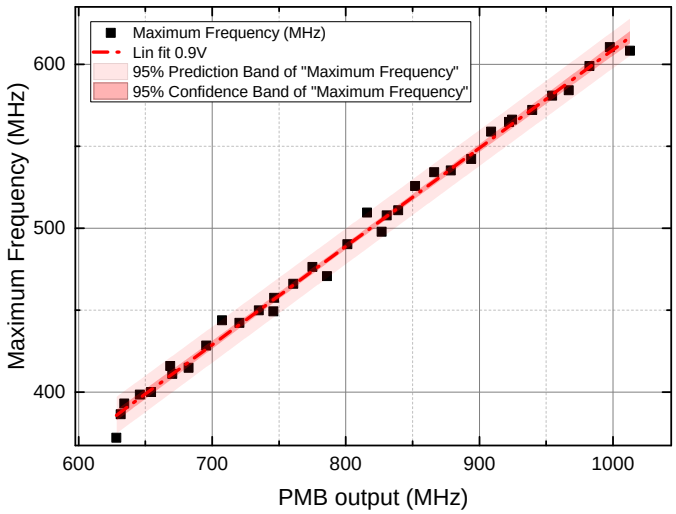

Fig. 3: This plot shows the relation between the maximum frequency and the PMB frequency estimation for a typical chip at 0.9 V and 25 °C. The data have been fitted with a linear model, the parameters of the fit are: y=mx+q where m=0.6 and q=8.7, R-Square (COD) = 0.995.

This means that, once the model is determined, the memory required to store the model's parameters is small, as well as the computational effort required to use run-time the model; the computational operation to convert the output of the PMB sensor is a simple multiplication. TABLE I shows the parameters of the linear fit at the three OPs for a single device. Notice that the C_{corr} is always lower than 1, which means that the maximum frequency of the circuit is always lower than the the one generated inside the PMBs. The R-Square parameter is always equal or higher than 0.995, indicating that the model properly describes the data's trend.

V_{dd}	0.9 V	0.7 V	0.5 V
C_{corr}	0.6	0.59	0.47
F_0	8.72	5.19	3.21
R-Square	0.995	0.998	0.998

TABLE I: Parameters of the model for a typical chip at three different supply voltages and 25 °C.

The equation we use to correlate the maximum frequency with the PMB frequency estimation, as well as other techniques used in other work, can be affected by errors. In our case the error is represented by the residuals of the measurements with respect to the fitting curve. The maximum residual we report for the correlation model, at fixed Operating Point, and for a single device, is the 3%. Comparable errors are reported also in [20], a similar study performed on a DSP architecture implemented with the same FD-SOI technology and exploiting a performance monitor system based on Timing Fault Sensors (TMFLT).

The presence of the model's error, that is a frequency uncertainty affecting the Maximum frequency estimation, can be compensated by means of additional forward body-bias. Notice that the V_{bb} margin added to compensate the model's uncertainty, is conceptually different from the V_{bb} regulation to achieve a target frequency. The approach used to compensate the error will be discussed in the next section.

V. Temperature Variations

The model we propose for the temperature-induced variation compensation is a generalization of the approach we described for a single operating point. To derive a more general fitting function, we performed the same data analysis on the entire data set (multiple OPs) introducing hence an additional degree of freedom. Thanks to this analysis, we obtained a general equation describing all the operating points for a single device. In Fig.4 is reported the entire set of measurements performed at different voltages and temperatures for a typical device. The red solid line, which is described by the equation 1, represents the general model fitting the data. Table II reports the parameters of the model in this case as well as the R-Square.

V_{dd}	0.7 V
C_{corr}	0.59
F_0	5.42
R-Square	0.996

TABLE II: Parameters of the model fitting the data for a single chip, at a given supply voltage, at three different temperatures.

Fig.5 shows the relative error introduced by this generalization. We measured a higher error when the supply voltage decreases; additionally, we also registered a well defined trend in the model residuals, which could mean that the measurement at 0.5 V was disturbed by some other factor.

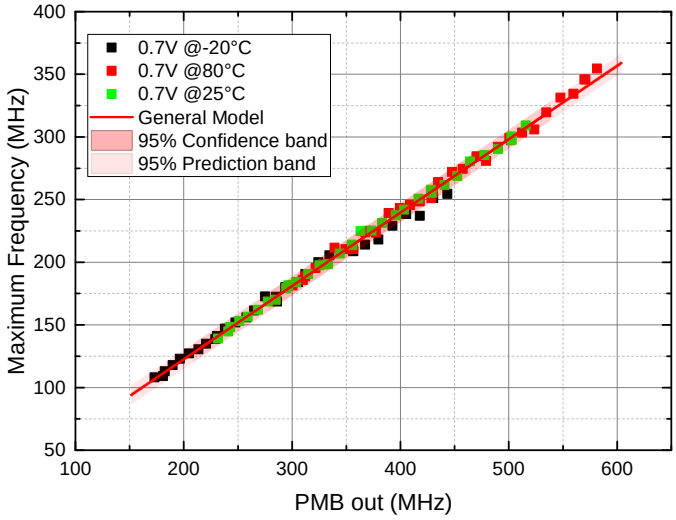

Fig. 4: In this plot are shown all the measurements performed at a single supply voltage V_{dd} = 0.7 V and three different temperatures T = {−20 °C,25 °C,80 °C}. The red solid line represents the general model to correlate the output of the PMB sensor with the maximum frequency of the device, which is the global fitting curve for the entire data-set.

Starting from the error estimation, since we assume to know the supply voltage at every OP, we used the empirical model that links the frequency and the body-bias voltage to convert the frequency error in a forward body-bias margin

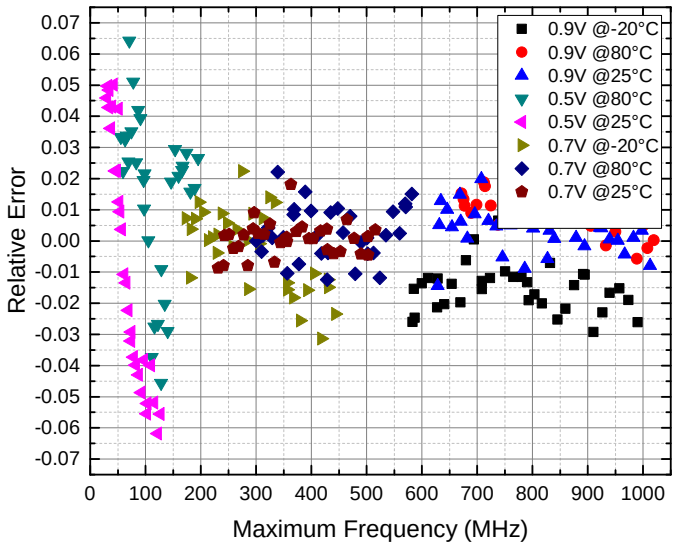

Fig. 5: This plot shows the distribution of the relative error versus the maximum frequency when the general model is used to convert the PMB output value in F_{max}.

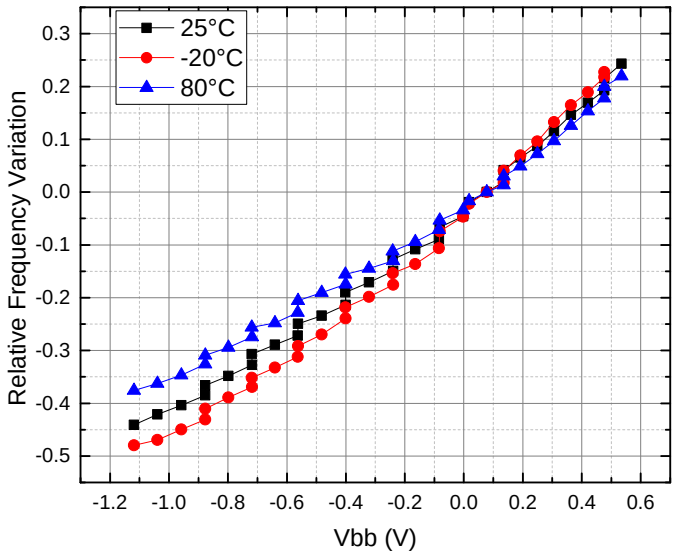

Fig. 7: In this plot is represented the relative frequency variation versus the body-bias voltage, with respect to the V_{bb} = 0 V condition. The operating points are: Vdd = 0.7 V and T = $\{-20\,°C, 25\,°C, 80\,°C\}$.

to add to the V_{bb} regulation. Notice that we are interested in aligning the maximum frequency achievable by the chip to the target one, therefore we need to compensate only for negative frequency errors. In other words, we will compensate those conditions where the general model overestimates the real performance of the device.

As shown in Fig.6, the relationship between the frequency and the body-bias voltage shifts with the temperature, however, as shown by the plot in Fig.7, once the supply voltage has been fixed, the relative frequency variation (in the forward body-bias region) can be considered in first approximation as temperature-independent; because the three curves are superimposed. Starting from this assumption, the amount of additional forward body bias that allows to neglect the effects of temperature variation can be calculated exploiting the relative frequency versus body-bias voltage relation reported in Fig.7, which is the frequency-normalized inverse of Fig.6. To give an example, assuming that we want to compensate the error at 0.9 V shown in Fig.5, that is the 3% in the worst case, we need to consider a V_{bb} margin of 0.1 V (The amount of body-bias is derived by the curves in Fig.7). Notice that the additional amount of V_{bb} causes an increase in the leakage current, hence additional leakage power consumption; the overhead caused by this methodology will be discussed in section VII.

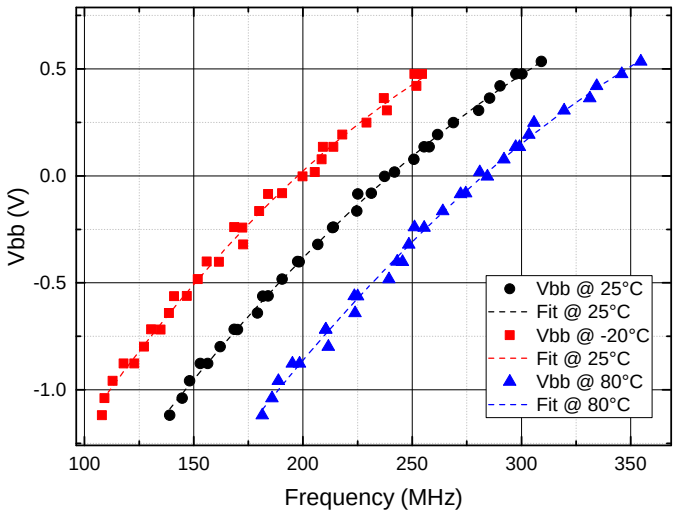

Fig. 6: This plot shows the relation between the maximum frequency of a typical device and the body-bias voltage at three operating points: Vdd = 0.7 V and T = $\{-20\,°C, 25\,°C, 80\,°C\}$.

VI. Process Variations

In this section we generalize the methodology we proposed to reduce the sensitivity to temperature' variations to compensate also process' variations.

As anticipated in the previous section, we considered the variations introduced by the process and the variations introduced by the temperature as separated. As demonstrated in [10], the additional body-bias voltage used to reach the target frequency, plus the margin to compensate the uncertainty of the model, can be summed to the fraction of V_{bb} used to compensate process' variations as independent contributions.

Here, we discuss the results of the analysis performed on a population of 8 different chips. In figure Fig.9 are reported the correlation data of all the chips in exam, for a given OP. Notice that the variations introduced by the process can be classified in two types: *i)* The Inter-chip variations, that can be observed in terms of performance gaps between different devices, as shown in Fig.8 *ii)* Intra-chip variations, as demonstrated by [21], resulting in different first N critical path, that can affect the consistency between the behaviour of the circuit and an on-chip performance monitor, confirmed also by our analysis.

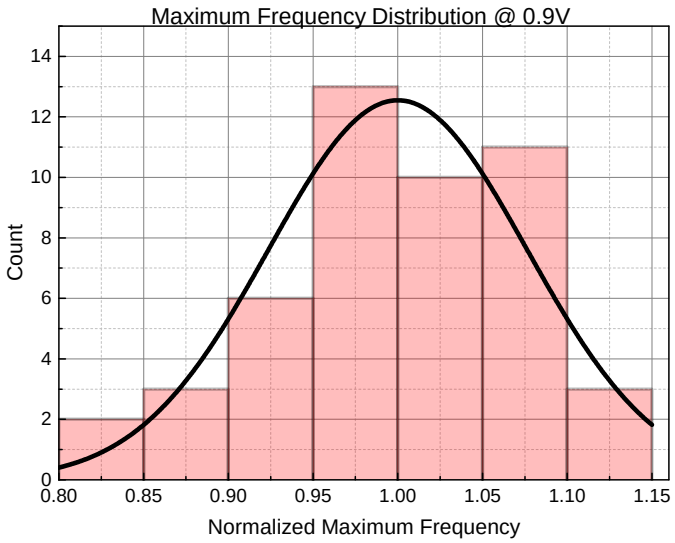

Fig. 8: Normalized distribution of the maximum frequency for the entire population of 48 chips at $0.9\,\mathrm{V}$, $25\,^{\circ}\mathrm{C}$.

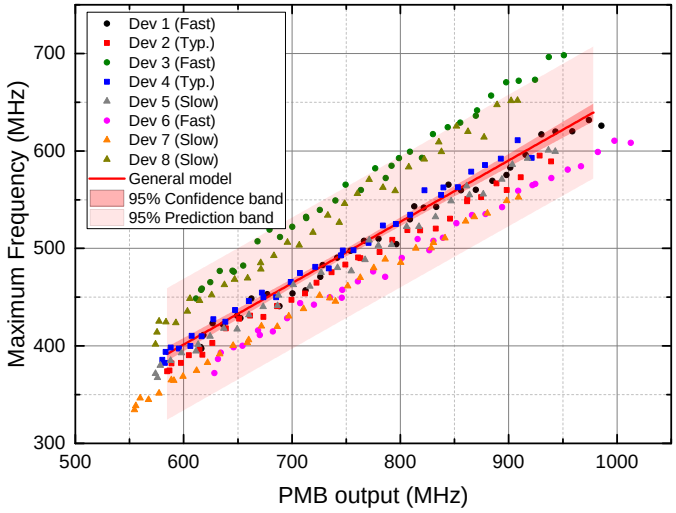

Fig. 9: In this plot are represented the data of 8 different chosen devices, the red solid line represents the fitting curve we assume as process-independent model.

As for the temperature variation compensation, we derived a general correlation model between all the chips. Fig.10 shows the error when this model is used to convert the output of the performance monitor sensor. Also in this case the model is the one described by equation 1, table III reports the values of the parameters and the R-Square; notice that in this case the fitted data have a larger variability, resulting in a larger Prediction band and lower R-square.

V_{dd}	$0.7\,\mathrm{V}$
C_{corr}	0.63
F_0	23.67
R-Square	0.814

TABLE III: Parameters of the model fitting the data for 8 chips, at a given supply voltage, at three different temperatures.

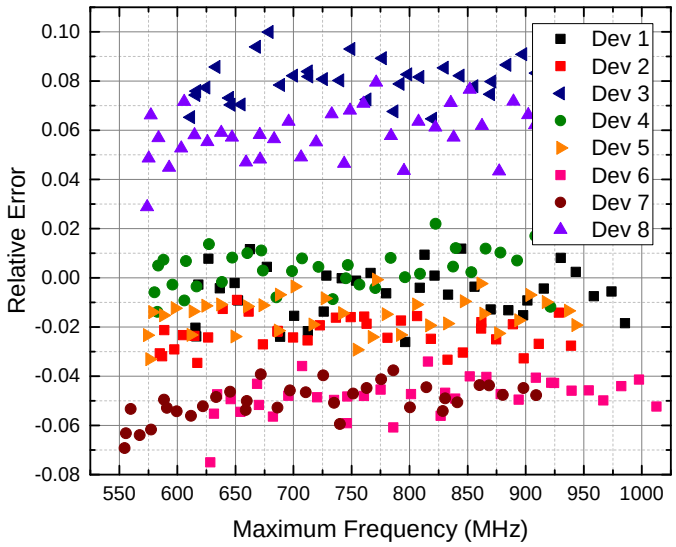

Fig. 10: This plot shows the distribution of the relative error versus the maximum frequency when the process-independent model is used to convert the PMB output value.

To generalize the model and compensate the error, as in the previous section, we derived the function that links the v_{bb} with the relative frequency change at different supply voltages. As shown in figure Fig.11, the sensitivity of the frequency with respect to the body-bias, in the forward body bias region, decreases when the supply voltage increases. Starting from this assumption, we can assess that, using a V_{bb} margin to compensate a frequency gap at $0.9\,\mathrm{V}$, at minimum the same frequency gap will be compensated also for lower supply voltages. More specifically, we can state that the frequency sensitivity to body-bias voltage is always higher than $10\%/100\,\mathrm{mV}$. Using this more general model, as expected, we obtained higher frequency errors (Fig.10).

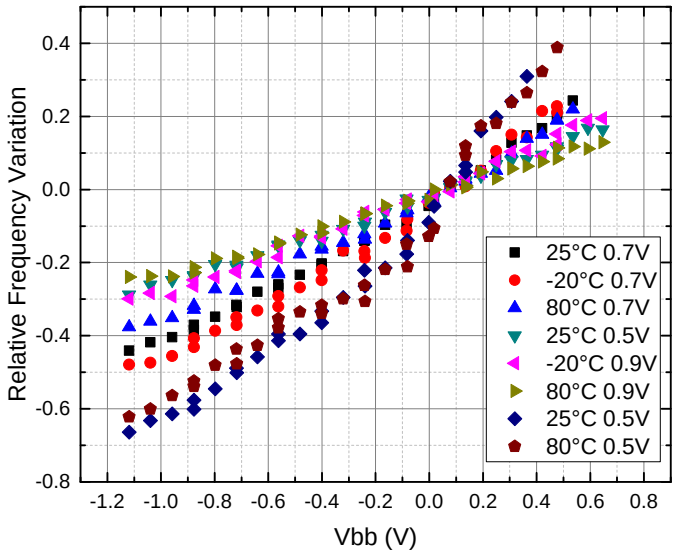

Fig. 11: The graph shows the Relative frequency sensitivity to the V_{bb} at different operating points.

VII. Methodology's Overhead

The methodology we propose uses the forward body-biasing to compensate the uncertainties in process monitor readings. Since the V_{bb} has a straightforward effect on the leakage current, hence on the leakage power consumption, it is important to quantify the overhead introduced by the body-bias margins. As it is well known, the higher is the frequency of the system, the more dominant is the dynamic power consumption with respect to the leakage power. Hence, the overhead decreases as the target frequency increases, since only the leakage power changes when the FBB is applied. Notice that we define the overhead as the additional leakage power with respect to the global power consumption, assuming to calculate it in the most common operating condition for a near-threshold processor, that is when the system is running at the maximum achievable frequency, given a supply voltage.

The following tables report the worst case overheads at three different conditions: *i)* when the Process-unaware model is used *ii)* when it is possible to calibrate the model on the single chip *iii)* when we can fully characterize the operating point knowing also the temperature.

When only the supply voltage is known, it is necessary to compensate temperature and process variations, as a consequence, the frequency error in the estimation of the maximum one is high and also the power consumption overhead. Table IV summarize the results at three different operating voltages.

V_{dd}	0.9 V	0.7 V	0.5 V
F_{err}	11%	12%	28%
$overhead$	14%	10%	17%

TABLE IV: Frequency error and power consumption overhead for temperature and process variation compensation.

When the model can be calibrated on the specific device, it is possible to reduce the error in the maximum frequency estimation approximately of a factor of 3. Table V shows the results in terms of frequency errors and overhead in this condition.

V_{dd}	0.9 V	0.7 V	0.5 V
F_{err}	3%	4%	7%
$overhead$	4%	3%	5%

TABLE V: Frequency error and power consumption overhead for temperature effect compensation.

Assuming to probe also the temperature, the only error affecting the model is the one given by the fitting curve, which is the 5% in the worst case (0.5 V), resulting in a worst case overhead of the 3%.

VIII. Conclusion

In this work we demonstrated the body-biasing capabilities of the UTBB FD-SOI technology in compensating temperature and process variations. We developed a methodology that provides run-time Maximum Frequency estimation by means of Process Monitor Boxes, enabling performance optimization policies exploiting body-bias regulation. We generalized the model to operate across all the allowed operating points in terms of supply voltage and temperature, for a single device, by means of a finely tuned body-bias-based temperature compensation. We proved that the error can be lower than the 3% in the best case, and it can be kept below 7% in the worst case. In this condition we measured an overhead in the global power consumption always lower than 5%. We generalized the methodology to compensate also process variations proving that the overhead is lower than the 17% in the worst case and lower than 10% in the best case. We observed that having the possibility to probe also the temperature, the error can be reduced to 5%, causing an overhead on the global power consumption lower than 3%.

Acknowledgment

We thank STMicroelectronics for chip fabrication. This work was supported by European project ExaNoDe (H2020-671578), European FP7 ERC Advanced project MULTITHERMAN (291125) and European project OPRECOMP (732631).

References

[1] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. N. Mudge, "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, 2010.

[2] D. Markovic, C. C. Wang, L. P. Alarcón, T. Liu, and J. M. Rabaey, "Ultralow-power design in near-threshold region," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 237–252, 2010.

[3] M. Alioto, "Ultra-low power design approaches for IoT," in *2014 IEEE Hot Chips 26 Symposium (HCS)*. IEEE, aug 2014.

[4] A. Pahlevan, J. Picorel, A. P. Zarandi, D. Rossi, M. Zapater, A. Bartolini, P. G. D. Valle, D. Atienza, L. Benini, and B. Falsafi, "Towards near-threshold server processors," in *2016 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2016, pp. 7–12.

[5] M. Alioto, Ed., *Enabling the Internet of Things*. Springer International Publishing, 2017.

[6] J. Tschanz, N. Kim, S. Dighe, J. Howard, G. Ruhl, S. R. Vangal, S. Narendra, Y. Hoskote, H. Wilson, C. Lam, M. Shuman, C. Tokunaga, D. Somasekhar, S. Tang, D. Finan, T. Karnik, N. Borkar, N. A. Kurd, and V. De, "Adaptive frequency and biasing techniques for tolerance to dynamic temperature-voltage variations and aging," in *2007 IEEE International Solid-State Circuits Conference, ISSCC 2007, Digest of Technical Papers, San Francisco, CA, USA, February 11-15, 2007*. IEEE, 2007, pp. 292–604.

[7] K. Sundaresan, K. Brouse, K. U-Yen, F. Ayazi, and P. Allen, "A 7-MHz process, temperature and supply compensated clock oscillator in 0.25 μm CMOS," in *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03.* IEEE, 2003.

[8] G. Gammie, A. Wang, M. Chau, S. Gururajarao, R. Pitts, F. Jumel, S. Engel, P. Royannez, R. Lagerquist, H. Mair, J. Vaccani, G. Baldwin, K. Heragu, R. Mandal, M. Clinton, D. Arden, and U. Ko, "A 45nm 3.5g baseband-and-multimedia application processor using adaptive body-bias and ultra-low-power techniques," in *2008 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*. IEEE, feb 2008.

[9] D. Rossi, A. Pullini, M. Gautschi, I. Loi, F. K. Gurkaynak, P. Flatresse, and L. Benini, "A 60 gops/w, 1.8 v to 0.9 v body bias ulp cluster in 28nm utbb fd-soi technology," in *2015 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*. IEEE, oct 2015.

[10] S. Kumar, C. Kim, and S. Sapatnekar, "Body bias voltage computations for process and temperature compensation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, no. 3, pp. 249–262, mar 2008.

[11] C. Oh, H.-O. Kim, J. Seomun, W. Kim, J. Jeon, K.-T. Do, H.-S. Won, and K. S. Kim, "Thermal-aware body bias modulation for high performance mobile core," in *2012 International SoC Design Conference (ISOCC)*. IEEE, nov 2012.

[12] M. Alioto, "Ultra-low power VLSI circuit design demystified and explained: A tutorial," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 1, pp. 3–29, jan 2012.

[13] W. Lee, Y. Wang, T. Cui, S. Nazarian, and M. Pedram, "Dynamic thermal management for FinFET-based circuits exploiting the temperature effect inversion phenomenon," in *Proceedings of the 2014 International Symposium on Low Power Electronics and Design*, ser. ISLPED '14. New York, NY, USA: ACM, 2014, pp. 105–110. [Online]. Available: http://doi.acm.org/10.1145/2627369.2627608

[14] G. Ono, M. Miyazaki, H. Tanaka, N. Ohkubo, and T. Kawahara, "Temperature referenced supply voltage and forward-body-bias control (TSFC) architecture for minimum power consumption [ubiquitous computing processors]," in *Proceedings of the 30th European Solid-State Circuits Conference*. IEEE, 2004.

[15] D. Jacquet, F. Hasbani, P. Flatresse, R. Wilson, F. Arnaud, G. Cesana, T. D. Gilio, C. Lecocq, T. Roy, A. Chhabra, C. Grover, O. Minez, J. Uginet, G. Durieu, C. Adobati, D. Casalotto, F. Nyer, P. Menut, A. Cathelin, I. Vongsavady, and P. Magarshack, "A 3 GHz dual core processor ARM cortex TM -a9 in 28 nm UTBB FD-SOI CMOS with ultra-wide voltage range and energy efficiency optimization," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 4, pp. 812–826, apr 2014.

[16] G. Yeap, "Smart mobile SoCs driving the semiconductor industry: Technology trend, challenges and opportunities," in *2013 IEEE International Electron Devices Meeting*. IEEE, dec 2013.

[17] D. Blaauw, S. Kalaiselvan, K. Lai, W.-H. Ma, S. Pant, C. Tokunaga, S. Das, and D. Bull, "Razor II: In situ error detection and correction for PVT and SER tolerance," in *2008 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*. IEEE, feb 2008.

[18] S. Clerc, M. Saligane, F. Abouzeid, M. Cochet, J.-M. Daveau, C. Bottoni, D. Bol, J. De-Vos, D. Zamora, B. Coeffic *et al.*, "8.4 a 0.33 v/-40 c process/temperature closed-loop compensation soc embedding all-digital clock multiplier and dc-dc converter exploiting fdsoi 28nm back-gate biasing," in *Solid-State Circuits Conference-(ISSCC), 2015 IEEE International*. IEEE, 2015, pp. 1–3.

[19] F. Conti, D. Rossi, A. Pullini, I. Loi, and L. Benini, "PULP: A ultra-low power parallel accelerator for energy-efficient and flexible embedded vision," *Journal of Signal Processing Systems*, vol. 84, no. 3, pp. 339–354, nov 2015.

[20] E. Beigne, A. Valentian, I. Miro-Panades, R. Wilson, P. Flatresse, F. Abouzeid, T. Benoist, C. Bernard, S. Bernard, O. Billoint, S. Clerc, B. Giraud, A. Grover, J. L. Coz, J.-P. Noel, O. Thomas, and Y. Thonnart, "A 460 MHz at 397 mV, 2.6 GHz at 1.3 v, 32 bits VLIW DSP embedding f MAX tracking," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 125–136, jan 2015.

[21] M. Zandrahimi, Z. Al-Ars, P. Debaudand, and A. Castillejo, "Challenges of using on-chip performance monitors for process and environmental variation compensation," in *Proceedings of the 2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. Research Publishing Services, 2016.