Thermal Challenges to Building Reliable Embedded Systems

Zebo Peng
Embedded Systems Lab
Linköping University, Sweden

ABSTRACT

More and more embedded systems are used in safety-critical areas such as automotive electronics and medical applications. These safety-critical applications impose stringent requirements on reliability, performance, low-power and testability of the underlying VLSI circuits. With silicon technology scaling, however, VLSI circuits operate very often at high temperature, which has negative impact on reliability, performance, power-efficiency and testability. This paper discusses several thermal impacts on VLSI circuits and their related challenges. It presents also a few emerging techniques that take temperature into account in the design and test processes.

I. INTRODUCTION

There is an exponentially increasing number of embedded computers controlling virtually all devices and systems in a huge spectrum of application areas including aerospace, manufacturing, chemical processes, healthcare care, automotive, transportation, telecommunication, and consumer appliances. Many of these systems are safety-critical, such as automotive electronics and medical equipment, with stringent reliability and real-time requirements. At the same time, with silicon technology scaling, VLSI circuits used to implement the computational components of these systems are built with smaller transistors, operate at higher clock frequency, run at lower voltage levels, and operate very often at higher temperature. Consequently, they are subject to more faults and interferences. We are therefore facing the challenge of how to build reliable and predictable embedded systems with unreliable and unpredictable components.

The most visible technology scaling issue is the continuous shrinking of feature sizes, which has led to the rapid increase of power density. High power densities have, in turn, caused not only large energy consumption but also increased chip temperatures. And high temperatures have negative impact on reliability, performance, power-efficiency, and testability. These thermal impacts should therefore be carefully analyzed and taken into account in the design process in order to build reliably embedded systems.

This paper will discuss the different thermal impacts on VLSI circuits implemented with deep submicron technology, and their related challenges, especially for system-level design. We will also present several emerging techniques that take temperature into account in the design and test processes for embedded systems.

II. THERMAL IMPACT ON RELIABILITY

The impact of temperature on the lifetime of integrated circuits has been extensively studied [1]. Many of these studies have considered failure mechanisms such as electro-migration and time-dependent dielectric breakdown, which has led to the conclusion that device lifetime decreases exponentially with junction temperature. They have also shown that thermal cycling─not only average and maximum temperature, but also the amplitude and frequency of temperature oscillation─has a huge impact on the overall lifetime of a chip.

For embedded systems, the thermal cycling impact can be exploited in the design process. In particular, many modern embedded systems consist of a set of periodic tasks, implementing several control functions, which run on a multiprocessor hardware platform with several microprocessors connected together by a communication infrastructure. Once the hardware platform is chosen, the periodic control tasks should be mapped onto the microprocessors. The tasks mapped on the same microprocessor should be scheduled as well as the messages to be transported over the shared communication infrastructure. Traditionally, mapping and scheduling are performed mainly to satisfy the real-time requirements and/or to optimize the control quality, which leads often to many thermal cycles with large temperature ranges. It is therefore important to develop temperature-aware task mapping and scheduling techniques that take thermal cycling into account in an optimization loop.

In order to perform design optimization with respect to not only timing constraints, control quality, power consumption, and cost, but also the thermal impact, we need a technique to characterize accurately and quickly the steady-state dynamic temperatures of a multiprocessor platform executing periodic tasks. Traditional system-level thermal analysis techniques based on simulation, such HotSpot [2], are too time consuming to be used inside such a temperature-aware system-level optimization loop. Therefore, fast system-level temperature analysis techniques should be developed, and one of such techniques has recently been reported in [3].

III. THERMAL IMPACT ON SPEED AND POWER

For embedded systems running safety-critical applications, circuit speed and power consumption are of great importance. Circuit speed influences the predictability of a system in terms of timing, while power consumption determines the battery lifetime of many mobile devices. In both these areas, chip temperature has, unfortunately, very large negative impact.

In general, high temperature leads to performance degradation, which is due to reduced carrier mobility and driving current as well as increased interconnect delay. A recent study has shown that the maximal clock frequency can be reduced by 23% when the temperature of a circuit is raised from 50°C to 110°C, when the circuit is operated at 1.1V [4]. This clock frequency and temperature dependency relationship should be carefully analyze in order to build embedded systems with predictable behavior in terms of timing.

The impact of temperature on power consumption is firstly reflected by that leakage power increases rapidly with the increase of temperature [4]. As leakage power grows significantly in relation to dynamic power with the deep submicron technology, due to the reduction of threshold voltage, channel length, and gate oxide thickness, the total power consumption of a circuit grows rapidly also with the increase of temperature, which in turn will further increase the temperature. This positive feedback may even lead to thermal runaway and burn the chip. Therefore, it is important to reduce the temperature of a chip in order to decrease the leakage power.

To reduce dynamic power, we can make use of dynamic voltage/frequency scaling (DVFS), which is especially suitable for embedded system applications with a set of tasks that run repeatedly with their respective periods. DVFS exploits the available slack times in the periods by reducing the voltage and/or frequency at which the processors operate to achieve energy efficiency. We have two types of slacks: the static slack, due to that, when executing at the nominal voltage level, a task finishes before its deadline even

978-1-4799-2217-8/14/$31.00 ©2014 IEEE

when executing its worst number of cycles, and the dynamic slack, due to that most of the time a task executes less cycles than the worst case. Offline DVFS techniques can only exploit the static slacks while online approaches can be used to further reduce energy consumption by exploiting the variation of the execution time of the tasks. However most of the existing offline and online approaches don't take into account the impact from the temperature of the chip.

When ignoring the chip temperature, the maximally allowed clock frequency for a given supply voltage *Vdd* is computed by implicitly assuming that the chip is operated by the maximum temperature, *Tmax*, at which it is allowed to run. Most of the time, the chip will not operate at *Tmax*, which means that it can actually run at a higher clock frequency. Or, we can set *Vdd* at a lower level and still deliver the same clock frequency as originally required. Due to the quadratic relation between dynamic power and *Vdd* of a CMOS circuit, we can achieve considerable power saving, when *Vdd* is reduced. For example, a dynamic power saving of approximately 52% can be achieved if *Vdd* is reduced from 1.3V to 0.9V, which is enabled by knowing that the chip is operated at 60°C instead of the maximum level of 110°C [4].

Temperature-aware DVFS techniques will be needed in order to exploit both static and dynamic slacks of an embedded application. It needs efficient methods to make quick decisions dynamically based on online temperature information. An interesting technique has been reported in [5], which consists of an offline temperature aware optimization step and an online voltage/frequency setting method based on temperature sensor reading. During the offline step, voltage/frequency settings for all tasks are pre-computed, using a temperature-aware DVFS algorithm, based on possible start times of the tasks and possible temperatures at the start times. The resulting voltage/frequency settings are stored in look-up tables. At runtime, each time a task terminates and a new voltage/frequency level will be set, the online mechanism looks up the appropriate setting from the tables, depending on the actual time and temperature reading. In this way, quick and accurate DVFS decisions are made at runtime to significantly reduce the dynamic power consumption [5].

IV. THERMAL IMPACT ON TEST QUALITY

It is well-known that during test, more power is consumed than in the normal functional mode because of a substantial increase of switching activities in the circuit under test. This high power consumption will lead to high temperature, which will cause test failures and yield loss, and in the worst case damage the chips with overheating. We need therefore temperate-aware test techniques, such as thermal-aware test scheduling technique to generate the shortest test schedule such that the temperature constraints are satisfied.

However, the thermal implication here doesn't mean that we simply keep the temperature below an upper limit during test. Many silicon defects are sensitive to a particular temperature level. For example, metal interconnect defects may pass a delay test at nominal temperature but fail the same test at a higher temperature. This means that speed tests, such as maximum-frequency test and transition delay test, should be applied at a higher temperature level in order to detect these temperature-dependent defects. Therefore, higher temperature is not always bad.

On the other hand, parametric failures induced by subtle defects, such as resistive vias/contacts and weak opens, are hard to detect even when the circuit operates under the worst environmental condition, such as maximum temperature. In these cases, a speed test needs to be applied at two temperatures (hot/cold) and at a particular frequency. The defective chips can then be screened as outliers by comparing the test results at the two different temperatures. This analysis means that we should test a chip at multiple temperature levels in order to get the best test quality, which leads to the challenges of how to perform multi-temperature testing efficiently, and how to set the chips under test in the desirable temperature ranges.

Furthermore, elevated temperature (usually together with higher voltage) can be used in a burn-in process to accelerate and detect early-life failures in order to produce reliable chips. The elevated temperature and voltage speed up the aging and wear mechanisms so that the chips experience their early life before testing. The wear mechanisms that are accelerated include metal stress voiding and electro-migration, metal slivers bridging shorts, and gate-oxide wear-out and breakdown. Recently several studies have, however, shown that some wear mechanisms are speeded up more efficiently by large temperature gradients rather than the high temperature itself. Therefore, a burn-in process that has not created the appropriate thermal scenarios do not sufficiently speed up the formation of the defects that depend on large temperature gradients and consequently such early-life defects will go undetected. The challenge here is how to introduce a burn-in process that enforces such appropriate temperature scenarios on the IC, which is an extremely difficult problem and can't be solved with traditional burn-in methods using a hot chamber. An interesting approach has recently been developed to apply high power stimuli selectively to the cores of an IC under burn-in through the test access mechanism to create the required temperate gradients [6]. In this way, no external heating equipment is required.

V. CONCLUSIONS

Embedded systems for safety-critical applications have put stringent requirements on reliability, performance, power-efficiency and test-ability of the underlying VLSI circuits. These different requirements are all impacted by the temperature of the chip. This paper has discussed several of these thermal impacts and their related challenges. It has also presented briefly several emerging techniques that take temperature into account in the design and test processes of embedded systems, especially at the system-level.

Note that many issues discussed in this paper, such as the influence of temperature on reliability, power consumption, and testability, are not new, taken individually. However, the interplay of these issues and their increased impacts have led to many great challenges. In particular, there are still many open problems in how to develop efficient global optimization techniques to consider the different thermal impacts and other design requirements at the same time, so that we can build highly reliable and predictable embedded systems in an efficient manner.

ACKNOWLEDGMENTS

Thanks to Petru Eles, Nima Aghaee, Ivan Ukhov, and Main Bao for the cooperation and help. Many ideas presented here are inspired by several research projects with their participations.

REFERENCES

[1] J. Srinivasan et al., "The Impact of Technology Scaling on Lifetime Reliability," *Proc. DSN*, 2004.

[2] W. Huang, et al., "HotSpot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design," *IEEE Trans. on VLSI Systems*, 2006.

[3] I Ukhov et al, "Steady-State Dynamic Temperature Analysis and Reliability Optimization for Embedded Multiprocessor Systems," *Proc. DAC*, 2012

[4] W. Liao, et al. "Temperature-Aware Performance and Power Modeling," Tech. Report, UCLA, Engr. 04-250, 2004.

[5] M. Bao et al, "Temperature-Aware Idle Time Distribution for Leakage Energy Optimization," *IEEE Trans. on VLSI Systems*, 2012.

[6] N. Aghaee, et al. "An Efficient Temperature-Gradient Based Burn-In Technique for 3D Stacked ICs," *Proc. DATE*, 2014.