TRANSIENT ERRORS IMPACT ANALYSIS FOR SUB-POWERED CMOS CIRCUITS AT MULTIPLE LEVELS OF ABSTRACTION OF A DIGITAL SYSTEM

PHD THESIS ABSTRACT

Chapter 1 – Introduction

According to Moore's Law, the number of transistors on an integrated circuit doubles every 18 months, approximately. This exponential increase of the degree of integration of chips has led to a nearly constant exponential growth of the capabilities of microelectronic devices. Since 2000, the idea that further increase of the integration density of chips may face a physical limit has spread among the members of the scientific community. The miniaturization problems anticipated by those papers have become a reality in the last years, due to the increase of thermal noise voltage, along with the necessity of using lower supply voltages with the purpose of reducing the power consumption.

Due to fundamental physical limitations and increasing requirements of power efficiency, performance and low fabrication cost, current chip designs have started to have reliability issues. With every generation of nanometric devices, the effects of process-voltage-temperature (PVT) variations increase in intensity. Energy and power dissipation problems have become critical especially for battery-powered mobile devices. Low-energy consumption can be sustained through low-powered components and the preferred method consists in the reduction of the supply voltage of the circuits to sub and near-threshold regimes. The behavior of a logic gate supplied at very low voltages has a probabilistic nature, meaning that the output of a gate will be the correct one with a probability p less than one. This behavior can be explained by two factors: the inability of the gate to switch in the desired time window or a single event upset which causes a bit-flip of the output of the gate.

This PhD thesis employs a bottom-up approach in order to analyze the impact of the probabilistic behavior of sub-powered CMOS circuits, at three levels of abstraction: circuit (transistor) level, gate level and register transfer level (RTL). At transistor level, the probabilities of failure of basic logic gates are computed, under different noise assumptions, and the results are used in order to develop gate-level fault models and reliability evaluation methodologies. For circuits composed of thousands of gates or more, analyzing the reliability and fault propagation only at gate level becomes prohibitive due to the huge memory requirements and long simulation time. A hybrid analysis of the system, at both gate level and RTL proves to be very efficient.

Chapter 2 – Sub-Powered CMOS Circuits

Chapter 2 of the thesis, entitled "Sub-powered CMOS circuits" explains the near-threshold and sub-threshold concepts, provides a review of the state-of-the-art implementations of sub-powered circuits and brings into attention the reliability issues.

Complementary metal-oxide-semiconductor (CMOS) is the most widely used technology for developing integrated circuits. The power density of CMOS chips should remain constant with each new and smaller technology node, but real world data shows that since the 90 nm technology this theory doesn't apply anymore. Instead of remaining constant, the power density for technology nodes below 90 nm increases exponentially. The power consumption per unit area of the chip shows an important increase, so many attempts to lower the supply voltage of the circuits are made.

Near-Threshold Computing (NTC) refers to an environment for which the supply voltage is set to a value only slightly higher than the transistors' threshold voltage. The energy per operation is reduced by an order of magnitude in NTC, with respect to the Super-Threshold Computing (STC). The concept of sub-threshold computing implies lowering the supply voltage below the threshold voltage, in a region where load capacitances are charged / discharged by subthreshold leakage currents. Due to this fact, the maximum performance of these circuits is limited, usually to operating frequencies of hundreds of kHz or a few MHz. Sub-threshold operation differs from classical super-threshold operation, because the sub-threshold on-current depends exponentially on threshold voltage and supply voltage, while the typical super-threshold on-current depends linearly on threshold voltage and supply voltage. The experiments performed in this PhD Thesis belong mainly to the sub-threshold regime.

Among the physical implementations of sub-powered CMOS circuits found in the literature, we can bring into attention a near-threshold voltage 32 nm Pentium processor, a 180 mV sub-threshold Fast Fourier Transform (FFT) processor and a sub-threshold microcontroller with integrated SRAM. The 32 nm experimental processor developed by Intel is able to operate over a full voltage range, from nominal to sub-threshold supply. The maximum energy efficiency, almost 10x greater than the one corresponding to the nominal supply voltage, is achieved when the processor operates close to the threshold voltage.

Process variations affecting low-power devices are represented by two main components: systematic and random. The systematic component is usually spatially correlated, so the amount of variation affecting neighboring devices will be equal. Among the random variations, we can mention the varying dopant concentrations.

Among the three categories of faults experienced by semiconductor devices (permanent, intermittent and transient), transient faults are the most outspread in nature. Studies from IBM and DEC showed that over 85% of all computer failures are due to transient errors. This is why the thesis goal is the analysis of transient errors impact.

Consequently, the reliability assessment of low-powered circuits becomes one of the main problems that must be studied. This can be performed using fault injection techniques,

which are classified in three categories: physical or hardware-implemented fault injection (HWIFI), software-implemented (SWIFI) and simulation-based. Simulated fault injection (SFI) is preferred because it can be used to evaluate the circuit under test during the design phase and it can be applied at multiple levels of abstraction.

Chapter 3 – Transient errors impact analysis for sub-powered CMOS circuits, at transistor level

Chapter 3 relies on SPICE analysis in order to perform Monte-Carlo simulations for sub-powered CMOS gates, under different noise assumptions. The proposed scenario for transistor-level analysis consists of an arbitrary behavioral voltage source intercalated between two serial linked inverters.

The first set of experiments analyzed the effect of the amplitude of the noise, using several Monte Carlo simulations consisting of 50,000 runs each. The supply voltage ranged from 0.3 to 0.8 V, with a resolution of 0.1 V and two different Gaussian distributions have been used for the noise signal: one with sigma 0.2 and one with sigma 0.3. The 65 nm Predictive Technology Model (PTM) transistor model has been used. The probability of correctness of the output of the second gate has been computed.

The second set of experiments aimed to find the minimum pulse width for which the noise will propagate through one or two logic gates. These simulations have been performed using the 45 nm PTM model on a two inverters chain and a two NAND gates chain, respectively. The supply voltage has been ranged between 0.2 V and 0.7 V and unequal PMOS / NMOS stages have been simulated by varying the PMOS transistor width with respect to NMOS transistor width. Both “0” and “1” glitches have been applied and we have observed an exponential increase of the minimum pulse width with the decrease of the supply voltage. For high voltages, almost all common glitches, with durations between 100 ps and 300 ps propagate through one or both gates. But for logic gates functioning in the sub-threshold regime, the duration of pulses must be very high in order through propagate through the gates.

On the other hand, some researchers from University College Cork, Ireland, have performed some additional simulations at circuit-level, considering three values for supply voltage (0.25, 0.30 and 0.35 V), three temperature values (25, 50 and 75 degrees Celsius) and they considered two parameters affected by process variation: threshold voltage and oxide thickness. The authors have extracted the probability of correctness as a function of the gate’s delay: a greater delay results in a higher probability of a correct switch. Also, the supply voltage and the input transition affect the probability of a correct output. Furthermore, the authors have performed delay-dependent reliability evaluation of the NAND-based D flip-flop.

As a conclusion, chapter 3 stands out with the following experiments and contributions:

- The effects of the amplitude and pulse width typical to transient errors in sub-powered CMOS circuits have been simulated
- A decrease in reliability is noticed from the noise amplitude point of view, with the decrease of the supply voltage;
- Regarding the propagation of transient faults, gates operating at low supply voltages show increased resilience to glitches, despite the fact that the noise margins of the circuit are diminishing.

Chapter 4 – Transient errors impact analysis for sub-powered CMOS circuits, at gate / logic level

In this chapter, two methodologies for transient errors impact analysis at gate level are presented. The first one uses the results provided by the circuit level analysis, in order to derive probabilistic fault models with different accuracies and to develop a mutant-based SFI technique. The four fault models developed are:

1. Gate Output Probabilistic model (GOP) – a bit-flip of the output of the gate can occur at any time, regardless of the input pattern, switching activity or previous outputs
2. Gate Output Switching probabilistic model (GOS) – this fault model considers that the probabilistic behavior occurs only when the gate switches, regardless of the switching type
3. Gate Output Switching Type probabilistic model (GOST) – similar to the previous one, but more complex, because it considers different probabilities for charging and discharging
4. Gate Input Switching Probabilistic (GIST) – a different probability is considered for each input combination that determines the switching of the gate

Each of the above fault models has a corresponding mutant architecture. The proposed SFI methodology has been developed in Verilog hardware description language and consists of two main phases: the setup phase and the simulation and results analysis phase. The setup phase comprises the following steps: fault parameter settings, probabilistic gate mutation, input data selection, gold circuit simulation and testbench generation.

The proposed methodology has been applied for 6-bit ripple carry adders (RCA) and carry-select adders (CSeA), implemented using only 2-input NAND gates. 16,000 test vectors have been applied to the inputs of each circuit under test, for each of the fault models GOS, GOST and GISP. The probability of failure of each bit of the result has been computed and, also, the overall probability of failure of the result. Two distinct situations have been considered: (a) all gates of the design have the following delay and (b) the gates on the critical path have the smallest delay, while the gates situated on other paths have larger delays.

The results have shown that the CSeA configuration has better reliability with respect to the RCA configuration and a triple modular redundancy (TMR) set-up does not improve significantly the RCA reliability if all modules are operating at the same supply voltage.

In the second part of chapter 4, a new methodology for reliability assessment is proposed, based on simulator commands and scripts. The new methodology is implemented using two approaches: one that uses a dedicated Verilog module in order to decide the moment when faults are injected in the design and another one based entirely on simulator commands. The probabilities used for these experiments are the previous ones, determined as a function of delay, and the circuits under tests are the same as in the previously described mutant-based methodology, in order to validate the accuracy.

The bit-independent and the overall probabilities of failure obtained using this technique based on simulator commands and scripts are slightly higher than the ones obtained using the mutant-based methodology, for the case of GOS model. This is explained by the fact that the number of faults injected by the second methodology in a certain time window is higher, because it doesn't take into account the switching activity of the gates.

The experiments carried out during chapter 4 use a bottom-up approach: SPICE Monte-Carlo simulations represented the starting point for deriving higher level error models. The simulations of circuits affected by process-voltage-temperature (PVT) variations are used for a mutant-based methodology, which evaluates the reliability parameters of small and medium combinational circuits. The simulation time overhead demanded by the mutant-based methodology is situated between 2x and 5x, with respect to the time required by the fault-free circuit.

The original contributions of the first methodology are:

- The definition of 4 fault-models for basic logic gates, with different accuracies
- Data-dependency feature assured by associating a different probability of failure to each distinct input combination that determines the switching of the gate
- flexible mutant-based SFI architectures for gate-level description of sub-powered circuits, it can be applied for circuits with unbalanced delay paths, multiple voltage islands or asymmetrical heated-up regions
- the flexibility of the proposed methodology for small and medium netlists has been shown by varying the selected parameters (voltage, temperature, delay) according to the topology of the circuit

The contributions of the second methodology are:

- easy to implement reliability technique based on simulator commands and scripts;
- the technique's accuracy has been validated by confronting its results to the ones of the first methodology;

- the simulation time required is reasonable if applied to small and medium complexity netlists; the overhead of this method is 6x – 30x with respect to the fault-free circuit simulation time;
- the technique has been implemented using two different approaches: one of them has the advantage that it brings no overhead to the Verilog code of the design under test.

Chapter 5 – Probabilistic interconnects

Chapter 5 focuses on reliability issues of signals transmitted on low supply voltage interconnects. Reliability degradation of interconnects occurs mainly due to two factors: process variation and crosstalk induced faults.

4 types of saboteurs have been proposed:

1. Standard Signal Probabilistic (SSP) – a simple bit-flip of the logic value of the signal on which it is applied is performed; this model doesn't take into consideration the transition that occurred on that line
2. Switching-Aware Probabilistic (SAP) – the probabilistic behavior occurs only when a transition takes place; the same probability or different probabilities may be considered for charging and discharging processes
3. Full Data Dependent (FDD) – this model considers that the probability of failure for a line is expressed as a function of the data pattern transmitted on the entire bus;
4. Partial Data Dependent (PDD) – this model is a simplified version of the previous one; it considers that the probability of failure for a certain line depends only on the transitions that occurred on a vicinity of 1-wire or 2-wires.

The circuit under test chosen for this type of simulation has been represented by the open-source multi-master Wishbone bus, designed in Verilog HDL. Each simulation campaign consisted in 1000 runs, each data set was chosen randomly and the signal groups which were the subject of SFI were:

- Data write signals
- Data read signals
- Address signals
- Master control and handshaking signals
- Slave handshaking signals

The SFI based reliability evaluation technique for probabilistic interconnects described throughout chapter 5 stands out with the following features:

- four types of saboteurs have been defined
- the crosstalk noise affecting the interconnects is data dependent, so a high accuracy analysis is required, which is implemented as follows: the partial data dependent saboteur takes into account the influence of transitions occurring on the

lines situated in a vicinity of the analyzed wire, while the full data-dependent takes into account the transitions that occur on all wires of the interconnect;
- probabilistic faults have been injected on several groups of signals of the Wishbone bus and the simulations have indicated the most critical signals in the overall reliability;
- the simulation overhead for a SFI campaign is 1.7x higher with respect to the fault-free circuit.

Chapter 6 – Simulated fault injection for reliability analysis of Register Transfer Level circuit description

In this chapter, a novel hierarchical hybrid methodology for reliability assessment of RTL circuit descriptions is proposed. It combines the Gate Level (GL) data dependent SFI for reliability metric extraction of building blocks and the saboteur-based SFI at RTL. This methodology aims to capture the accuracy characteristic to GL SFI, while maintaining the low simulation overhead specific to RTL based evaluation. Performing SFI at higher levels of abstraction, such as RTL, requires orders of magnitude lower simulation time with respect to GL SFI.

The methodology presented in this chapter comprises the following phases:
- correct simulation for a specific set of inputs of the RTL description, in order to capture the inputs of each component
- hierarchical block decomposition, in order to split the RTL design in small simple blocks
- logic synthesis of the modules obtained after the previous step
- data dependent SFI of the GL netlists obtained after the logic synthesis step
- saboteur-based RTL SFI using the probabilities obtained in the GL SFI step

The accuracy of the proposed RTL fault injection methodology has been validated by performing the reliability assessment of a medium sized circuit: a parallel comparator, which is a basic component of the check-node-unit (CNU) processing unit of a Low-Density Parity-Check (LDPC) decoder. The circuit is composed of two types of modules: the sort module, which is used for arranging two pairs of inputs in an ascending manner and the compare-select module, which identifies the first two minimums among a set of four values received as inputs. The probabilities of failure of the circuit obtained using the new methodology are very close to the ones obtained for entire GL SFI of the circuit. Regarding the simulation time, the new approach requires three orders of magnitude less simulation time than GL SFI.

Furthermore, the new hybrid methodology has been applied for assessing the reliability of an AES crypto-core, which occupies 40% of the total number of slice LUTs and 25% of the total number of BRAMs of a Xilinx Spartan-6 FPGA. The AES crypto-core contains more than 1 million of NAND gates and D flip-flops and its entire GL simulation was impossible on the target machine due to the huge memory requirements.

The described hierarchical SFI methodology has the following advantages with respect to the existing solutions:

- it provides high accuracy characteristic to lower abstraction levels due to the GL simulations performed for each block and sub-block of the design;
- the high accuracy is enhanced by embedding the data dependency concept in the GL simulations; this is achieved by exploiting the different accuracy levels provided by the mutant architectures associated to the 4 fault models defined in chapter 3: GOP, GOS, GOST and GISP;
- it maintains a reasonable total simulation time, characteristic to upper abstraction levels, due to its hybrid nature;
- scalability demonstrated for circuits of different complexity.

Chapter 7 – Reliability analysis of Low-Density Parity-Check (LDPC) decoders

Section 7.1 of chapter 7 brings into discussion the theoretical background of Low-Density Parity-Check (LDPC) codes. An LDPC code can be conveniently represented by a bi-partite Graph called Tanner Graph, which contains two types of nodes: check nodes and variable nodes. The most used decoding algorithm relies on message passing, which implies the permanent exchange of messages between the two types of processing units: check-node units (CNUs) and variable node units (VNUs). The hardware implementations of LDPC decoders are usually tailored for FPGAs and consist of processing units (CNUs and VNUs), three types of memories for storing alfa, beta and gamma messages and control units for asserting the sequence of control signals needed for the decoding process.

Section 7.2 presents an RTL saboteur-based SFI for fault-tolerance analysis of a flooded Min-Sum LDPC Decoder. The analysis has been performed at three layers of abstraction as follows:

1. At circuit-level, the authors have performed statistical static timing analysis (SSTA) based on Monte-Carlo SPICE simulation, in order to extract the propagation delay distribution for PVT variations for each standard cell component
2. At gate level, the worst propagation path is determined for each primary output of each combinational block; for each primary output, the delay distribution is derived using a linear composition of PDFs corresponding to standard cell gates
3. At RTL, probabilistic saboteurs are inserted in the RTL description of the circuit, on each primary output of the combinational blocks

For the first step, Monte-Carlo SPICE simulations have been carried out to derive an Inverse Gaussian distribution for standard cell components. During second phase, gate level analysis is used to derive the error probabilities for each primary output of each combinational block. The PDF of each primary output is derived as a linear composition of the components on the worst delay path for that specific output. My contribution during this set of experiments is situated at the RTL, where SFI has been performed using saboteurs, which have been applied at the inputs of memory components.

The methodology proposed in this section has been used to evaluate the error correction capability of an overclocked sub-powered flooded MS LDPC decoder. The simulations indicate that increasing the operating frequency by a factor of 2 with respect to the maximum frequency allowed by the fault-free decoder, will not affect the error correction capability.

Section 7.3 introduces a new LDPC decoder multi-codeword memory-oriented flooded architecture, which has been built with the purpose of increasing the efficiency of BRAM utilization. The proposed mechanism stores multiple messages corresponding to multiple codewords in the same memory word. The synthesis results show a linear increase of the number of LUT flip-flop pairs used, with the increase of the number of processed codewords. The number of BRAMs used remain the same for 1, 4, 6 or 9 codewords processed simultaneously.

Sections 7.4 and 7.5 of this PhD thesis describe two sets of experiments performed for the reliability evaluation of the previously proposed LDPC decoder architecture. The first set of experiments assumes that the memory modules of the LDPC are implemented using only D flip-flops and injects errors in each type of memory (alfa, beta and gamma) using saboteur-based SFI. The probabilities of failure considered for this experiment correspond to LDPC clock frequencies of 400 MHz, 450 MHz and 500 MHz, respectively. The values of the error rates are: 1.25×10^{-3} , 2.4×10^{-3} and 4×10^{-3} per clock cycle, per memory bit. The frame error rate (FER) performance of the decoder has been monitored for each case and a graceful degradation of the decoding performance is obtained with the increase in operating frequency.

The results obtained in section 7.4 show that faults injected in the alfa-memory lead to a slightly lower decoding performance than faults injected in the beta-memory. Furthermore, an LDPC decoder with a faulty LLR input memory (gamma memory) has a insignificant improved error correction capability with respect to a decoder with faulty alfa or beta memories.

For a complete reliability assessment of the decoder, section 7.5 extends the experiments of section 7.4 to the processing units, also. The methodology comprises the following steps:

1. The division of the LDPC processing units (VNUs and CNUs) into combinational and sequential sub-blocks
2. The logic synthesis of the combinational sub-blocks, using Synopsys Design Compiler
3. Critical and non-critical path extraction from synthesis timing reports for each combinational sub-block
4. A delay constraint is applied for the processing units and the appropriate delay is associated for each gate, depending on its appurtenance to a critical path (the path with the maximum number of gates) or a noncritical path (paths with fewer gates than the critical one)
5. Parallel simulation of the fault-free processing unit and the faulty gate level processing unit
6. Extraction of the probability of failure for each bit of each primary output of the sub-blocks of both CNUs and VNUs

7. RTL saboteur-based simulated fault injection of the entire decoder and FER plotting, using the previously determined probabilities of failure for each combinational sub-block

Considering a FER value of 10^{-2} , the error correction capability of the faulty LDPC decoder, for a constraint of 200 ns on CNU, degrades with approximately 0.1 dB, with respect to the FER of the fault-free decoder. The degradation is equal to about 0.3 dB for a constraint of 133 ns on CNU and it increases to over 0.5 dB for a constraint of 100 ns. In the case of VNU injection, an error floor is obtained at a FER of 10^{-1}.

In conclusion, section 7.3 describes a new LDPC decoder flooded architecture, with the following characteristics:
- efficient memory utilization is obtained by packing multiple messages corresponding to multiple codewords into the same BRAM word
- up to 9 codewords can be processed in parallel for 4-bit quantization and up to 12 codewords for 3-bit quantization, without introducing significant memory overhead
- with respect to other LDPC decoders, we use one order of magnitude less BRAM blocks per processed codeword

The experiments depicted in paragraphs 7.4 and 7.5 of this thesis bring the following contributions:
- reliability assessment using saboteur-based SFI is performed for the memory modules of an LDPC decoder, which has a greater complexity than a processor core
- reliability assessment using gate-level mutant-based and RTL saboteur-based SFI is performed for the processing units of an LDPC decoder, in a hierarchical manner
- the FER performance of the LDPC is monitored under different fault injection assumptions;
- a graceful degradation of the error correction capability is noticed, with the decrease of the delay constraint set for each combinational module; the error correction capability depends also on the module which is being injected;
- the VNU units show a greater vulnerability to transient faults than the CNU units;
- an error floor is obtained at a FER of 10^{-1} for the case of VNU injection.