Designing Circuits from Imperfect Components in VISI Giga-scale Technologies

Antonio RUBIO, Ferran MARTORELO and Francesc MOLL

Department of Electronic Engineering, Technical University of Catalonia
Campus North UPC, c/ Jordi Girona 1-3, 08034 Barcelona, Spain
E-mail: rubio,ferranm,moll@eel.upc.edu

The evolution of the integrated circuit technology during the last 3 decades has been based on an increasing accuracy of the manufacturing process. With this principle and by using a quality control at the end of the production line (Test Technology) the semiconductor industry has reached very high productivity levels. However, with technology reaching critical sizes (< 65 nm) the manufacturing control is starting to fail and new design principles have to be introduced to be able to produce functional chips from low quality components. In this article two scenarios partially addressing the problem with gradual introduction of redundancy are considered: the scenario of the era at the end of the CMOS Moore's Law and the expected next scenario of nanoelectronic technology using new emergent devices. In the paper techniques for the design of robust electronic systems in spite of the low quality of components are presented for the two scenarios.

Key Words: Integrated circuit technology, manufacturing yield, redundant design, quality of electronic components, error detecting/avoiding codes in digital systems, new emergent devices technology.

1. INTRODUCTION

The semiconductor industry probably represents today the most sophisticated, and accurate manufacturing technology. Very complex circuits made from the aggregation of hundreds of millions of solid-state electronic devices (MOS transistors of edge sizes around 100 nm) on a single silicon crystal are massively produced to give support to multimedia, communication, control and computing market requirements.

The manufacturing process is based on a few dozens of physico-chemical steps applied on a silicon wafer in a controlled fabrication facility. Thousands of circuits are manufactured at the same time, resulting in a spectacularly low cost of each manufactured component. The total amount of transistors manufactured by the whole semiconductor industry has been reaching levels around 10^9 devices in a single year (2002, [1]). This level of sophistication has been reached gradually, from the invention of the integrated circuit by Jack S. Kilby in 1958, based on a miniaturization of the photolithography process causing a continuous scaling down of the component feature sizes (from several microns at the beginning of the integrated circuits era until sub 90 nm in the year 2005). This size reduction trend foretold by Moore's Law (Gordon Moore enounced that the number of devices per IC will double each year in 1975, [2]) has additionally caused an improvement of the circuit performance (such as working frequency or consumption of the circuits), allowing the low cost high performance communication and computing systems we use today.

The implementation process of an integrated circuit has always followed two different and separated stages: Design and Manufacturing. During all these years both stages have had an independent and isolated evolution one of the other [3]. This separation has been an additional benefit for the semiconductor industry. The designer was always sure about the manufacturability of the circuit, and only cared about performance (evaluated from simulation tools) and optimization (minimum number of components). Manufacturing defects were addressed at the end of the manufacturing process, applying a chip by chip final quality control. For this purpose, Integrated Circuit Test techniques have been developed that allow achieving 95% yield levels in mature technologies. In order to achieve such successful test strategies, special design techniques, known as Design for Testability have also been successfully applied.

Present and future CMOS technologies expected at the end of the present decade, (and with more emphasis for future new emergent nanodevices) present new challenges that threaten today's design and test strategies. In these new technologies, the yield is being hit by aggression mechanisms intrinsic with the technology progress. Gigascale circuits (> 10^9 transistors, expected in the era 2007-2010) will be made from transistors with planar features lower that 65 nm and thin oxide layers lower that 1.5 nm the technology. Such small dimensions imply that we are leaving...
the statistical bulk principles and entering in a behaviour related with discrete amounts of atoms. This causes a loss of control in the manufacturing process [3] and a consequent drop in the manufacturing yield becoming a critical situation for the further evolution of semiconductor technology.

This paper discusses the need of introducing new design techniques that assume the low quality of the components and allow keeping the advance of the electronic systems for both eras: the nanoscale CMOS transistors and the new emergent nanoelectronics era. Sections 2 and 3 are dedicated to consider the situation and potential solutions for next CMOS technology and sections 4 and 5 for the future emergent technologies. Finally section 6 summarizes the conclusions of the paper.

2. DESIGN FOR NANOMETRIC CMOS DEVICES

The predicted deterioration of the quality of the components in the sub-100 nm CMOS technologies of the Gigascale integration era threatens the efficiency and application of conventional design methodologies of digital systems. Three aggression sources are mainly considered: the increasing variability mechanisms of the device parameters caused by the critical feature sizes, the intense and practically unpredictable internal noise environment and a moderate rate of physical defects. The Group of Research on High Performance Integrated Circuits of the UPC is investigating and modelling the mechanisms and the impact of the aggressions on performance and reliability, inside a project framework named FUTURIC. From this knowledge, new design methodologies can be proposed in the field of error-correcting circuits and systems. The project focuses on the application of redundant codes in specific places of the data and control path of the systems. A moderate increase of overhead due to this redundancy is accepted in this scenario if the benefits in terms of manufacturing and design stress are clear (ITRS, [4]). The project evaluates a new design scenario where quality of components, performances, redundancy and reliability are all together involved.

The ITRS mentions, among others, the following limiting factors for the circuit manufacturing in this era:

- Non ideal scaling of parasitic components as well as the ratio between threshold voltage and power supply voltage.
- Increase of parasitic coupling between interconnections and devices.
- Serious increase in the parameter variability of manufactured components.
- Predominant effect of lines and interconnections on the delay between stages and consequently on the circuit performance.
- Decrease of the reliability of the components.

As an example Patrick Gelsinger, Senior Vicepresident & CTO of Intel Corporation exposed in the 2004's Design Automation Conference [4] that for the next technologies the impact of the cost of extra redundant components will be practically negligible. Consequently we can summarize:

- The components, MOS transistors, will suffer important random fluctuation in their parameters. Voltage and temperature fluctuations will reinforce this effect. Designs with conventional rules to satisfy the worst case conditions will produce systems with very limited performances.
- The internal noise in the circuit caused by close couplings or noise caused by switching activity and coupling through common means will difficult the signal integrity. Special impact will have the unpredictable behaviour of noise.
- To manufacture an integrated circuit with 100% perfect components will be expensive when not unattainable.

The FUTURIC project has as its main goal to define new methodologies and mechanisms of digital system design when the quality of the components is so low that it is not possible to follow conventional techniques. The project will explore the use of codes to correct or detect errors caused by aggressions (variability, noise and physical defects). The concept is the same used in digital communication links where the quality of the channel cannot be assured but the system inserts correcting mechanisms at an acceptable degree for both service quality and cost. This technique has already been used in memory systems [5]. Codes oriented to reduce transitions in a bus (for power consumption reduction) have also already been studied [6]. For errors caused by delay variations as well as asynchronous systems Delay Insensitive Codes are well established [7].

To illustrate these ideas the next section shows the different errors to consider in gigascale CMOS circuits and how codes could be used to avoid them.

3. THE USE OF CODES IN GIGASCALE CMOS CIRCUITS

Testing next gigascale generation integrated circuits is currently recognized as a major problem and challenge [8]. The huge integration density, the strong and practically unpredictable interaction between components due to the reduced feature sizes, the low and fluctuating power supply voltage, and the increasing variability of component parameters are among the main causes. In order to maintain the improvement in performances offered by technology evolution the use of error detection techniques followed by recovery mechanisms as well the use of error-correcting or error-avoiding codes are nowadays being considered [8]. From the coding point of view three different error sources can be considered:

- Errors generated in the processing and memory blocks
- Errors generated in the large buses transferring data between blocks
- Delay errors between word bits caused by heterogeneous data processing and bus paths.

3.1 Processing and memory blocks

For these blocks three types of faults can be considered: permanent faults, deterministic transient faults and random transient faults. For the first two types, error-correcting codes have to be considered. The block is designed with a redundant section (causing an overhead that could reach more than a 100% in area) generating additional bits so that the receiving block is able to determine whether the received information is valid or not. This technique has been intensively used in fault-tolerant memories [9], where usually Hamming codes are used. Hamming codes are linear error-correcting codes using more or less complex parity errors to detect errors. Although most of the design and analysis done for these codes is oriented to correct single errors in code words the codification exhibits a relevant error-detection capability. The area overhead of the additional processing circuitry required to generat Hamming codes is moderate and the implementation methods are well known. For example, in [9] a
memory section with a single bit error probability of $10^{-6}$ by using a (522, 512) Hamming code gives a probability of an error bit of $10^{-4}$ for a 512-bit word after the error correction scheme.

For random transient faults much simpler error-detecting codes can be used, so the error can be recovered after an appropriated re-processing of data. Parity and k-out-of-n codes are adequate for this purpose.

3.2 Large buses transferring data

Large VLSI buses are considered IC's sections with high probability of transient errors, due to the large coupling capacitances between lines and the high probability of interference from substrate. Self-checking detection codes have been reported to be adequate for the diagnosis of transient and crosstalk faults affecting bus lines [10]. Such faults are detected on-line identifying the affected lines. In [10], it is shown that it is possible to generate a detecting scheme with self-checking capabilities considering a set of realistic internal faults such as node stuck-at, transistor stuck-open and stuck-on, bridges and crosstalk.

3.3 Delay error in word bits

The complex processing logic and transferring blocks cause and important deviation of the path delay for the set of bits of the word, this effect is enlarged because the important process parameter deviations. This is a drawback in the conventional concept of synchronous systems. For large deviations of synchronous systems and especially in the case of asynchronous communications (GALS strategy, globally asynchronous, locally synchronous) Delay Insensitive (DI) Codes are applied [11]. One-Hot, Double-Rail, Knuth, Berger and Sperner codes are delay insensitive. One-Hot codes are among the most trivial DI codes but they are very inefficient for large code lengths. Double-Rail and Berger codes are separable and their encoding is simple. Berger codes look especially promising for the use in DI circuits [7]. Sperner codes are optimal DI codes but no easy encoding scheme is known. Finally, Knuth codes are subcodes of Sperner codes allowing an acceptable encoding scheme.

The FUTURIC project objectives deal with the investigation on error-correcting codes and the design of efficient recovery mechanisms.

4. EMERGENT NANOTECHNOLOGY ERA

As discussed in section 2 electronic technology is facing a new design environment as silicon based electronics is reaching its limits. According to current predictions the silicon MOSFET are going to lead the technology until the 10 nm node [11] but due to its physical limitations different devices and/or technologies are needed to extend the technology further. Alternatives have been researched to provide working devices on the scale of 1-2 nm or below. Some of them are single electron tunnelling (SET) devices [12], quantum devices, carbon nanotubes [13], molecular devices [14] or DNA based technologies [15]. None of these technologies is still sufficient to compete with silicon designs so it is still unsure which of them is going to be the replacement of the MOSFET for deep nanometer scale designs (below 10 nm).

However, independently of the exact implementing technology, there are several difficulties that need to be considered before the electronic industry is able to exploit the nanoscale. There are three main difficulties:

- High number of defects
- Large variability of device parameters
- Very low SNR (around 1-0 dB)

This new situation is inherent to the nanometric scale. Once we approach dimensions of 1-2 nm, statistical properties of matter are no longer valid and instead the manufacturing process are going to deal with finite amounts of atoms where the lack or excess of one atom may produce a significant change in the dimensions or characteristics of the device [16, 17]. As - due to cost, both in time and money - it is not possible to manipulate the matter atom by atom for building electronic ICs, new approaches are needed to manufacture the nanometric electronic designs. Most promising approaches are based on bottom-up fabrication techniques mostly based on self-assembly properties provided by chemical reactions. These methods are very cheap and may be used in deep submicron design. However, chemical reactions do not provide a 100% yield [14]. These are the main causes of defects and parameter variability on nanotechnologies. Obviously, as the accuracy of nanoscale manufacturing improves these problems will be lessened but not completely avoided.

Noise is a more persistent problem. Due to the reduction of device's dimensions the density of dissipated power is rapidly increasing. The maximum dissipated energy density cannot overcome the limit for the materials that build the devices (100 W/cm² for silicon [18, 19]). To reduce the dissipated power the signal levels used in the devices must be reduced as much as possible. Classically, the signal working levels are calculated to allow a safety margin that avoids the effects of noise (e.g. interferences, thermal noise, crosstalk couplings...). If these levels are reduced, errors due to noise will increase. It is not possible to eliminate noise as it is a physical phenomenon due to temperature and capacitive couplings (that are increased with reducing distances). Some authors indicate that the signal to noise ratio in nanodevices may be as low as 0-1 dB [20]. So it is necessary to devise noise tolerant architectures to provide a safe way to produce electronic circuits.

All three problems cannot be completely eliminated thus design techniques that permit to tolerate them are becoming more and more necessary. The topic of fault and defect tolerance has been widely studied since the beginning of the electronic technology. Starting with von Neumann's work on NAND multiplexing and majority voting cells [21] lots of research has been done in this area. Some theoretical studies showing the limits have been also realized [22]. However for implementation most works are based on von Neumann's ideas. In triple modular redundancy each circuit is triplicated and its output is given by a majority gate. Two voted outputs are considered correct. R-modular redundancy is the extension of this technique by replicating each circuit R times or cascaded R-modular redundancy in which the modular architecture is cascaded to higher levels forming redundant functional units. NAND multiplexing and parallel restoration produce redundant elements and add restitution circuits to increase the robustness against the noise [23]. An alternative fault tolerant technique is based on information theory. As shown in section 3, systems can use codifications to detect and correct errors produced by faults and defects [24]. These techniques are well suited for transmission mediums, but their implementation for computing circuits is not straightforward. One of the main difficulties is that codification requires circuits to code/encode information, if these functions are not fault free the whole system tolerance is
compromised. For this reason these techniques can be used in systems with low defect rates, but they are nearly useless in systems with high numbers of defective elements.

Since considering nanotechnology for electronic design all fault and defect tolerant techniques have been revisited to provide architectural solutions for nanocomputing [15, 25-27]. All these techniques, although different, are based on the same principle: redundancy. Either redundant cells or redundant information are necessary to provide a way to tolerate errors. The difference among these techniques is how to use the spare cells or information. From recent studies it has been proved that not all of them are valid for systems where the number of faults and defects are very high [Nikolic02]. It seems that most promising techniques are reconfigurability [28] and averaging cells [27, 29].

These techniques consider noise as a nuisance to the system performance. So even if they provide a way to tolerate occasional faults, signal levels still require some restrictive safety margin to reduce noisy transitions. In contrast to these ideas we observe that biological systems (in particular sensory neurons) process information with SNR close to 0 dB [30]. This exceptional noise tolerance - in fact, noise even improve the system response - is due to a phenomenon called stochastic resonance (SR). We can define it as the improvement of a non-linear system performance by the presence of noise. By it, signals clearly under the activation level can be detected. Figure 1 left shows the system response in SR and in noiseless conditions. This phenomenon was discovered in the 80’s [31] and an extensive theoretical research has been realized since then (for an overview see [32]). This phenomenon has been observed in biological systems [33], but has also been used for engineering applications [34, 35]. A second phenomenon, more recently observed, is supra-threshold stochastic resonance (SSR) [36, 37]. It permits the improvement of the information transmitted by a non-linear array system by noise even for supra-threshold signals. Figure 1 right depicts the array response in SSR and in noiseless conditions. This figure clearly shows that the amount of information at the output is larger in the stochastic resonant regime.

5. APPLICATION OF MASSIVE REDUNDANCY ARCHITECTURES

Architectures valid for nanotechnology applications must be able to provide tolerance to the three problems. Furthermore it is convenient to use regular structures with elements as simple as possible to simplify the manufacturing process [14]. The Group of Research on High Performance Integrated Circuits of the UPC is investigating new architectures tolerant to defect and parameter variations of devices that can use noise to improve its performance. Our group is working with a structure that combines both tolerance to defects and variations and stochastic resonance [29, 38]. The structure is composed by an array of simple redundant elements and produces its output by averaging all the individual outputs. It can be used as a building block for nanotechnology functions.

5.1 Cell structure

The averaging cell architecture is based on an array of identical elements with a common input, $x$, and an averaging circuit that combines all the individual array outputs, $y_i$, to provide a global output, $y$, to. The cell response depends on the function the elements realize and the noise at the cell output on the number of elements in the array. Figure 2 right shows the ideal architecture of the cell. In it the array elements are represented by their function $h(\cdot)$ and an ideal averaging circuit is represented by an adder and an averaging factor by $n$. The cell can be operated in analogue mode if its output is taken just after the averaging ($y$) or in digital mode if a threshold device is added ($y'$). Figure 2 left shows the model for each array element. In them their internal noise is modelled as an independent identically distributed (i.i.d.) additive noise source, $\eta_i$ (we consider Gaussian white noise with zero mean and a given standard deviation (std), $\sigma_i$). The transfer function, $h(\cdot)$, is modelled by a soft limiter centred on its threshold value, $T$. The element output goes from $a_i$ to $a_h$ with a gain $g$.

5.2 Noise tolerance

As discussed previously, SNR are expected to be close to 0 dB for nanodevices. Under these conditions it is very interesting to use SR phenomena to tolerate noise and take advantage of it. The averaging cell is able to exploit this phenomenon. Figure 3 (left) shows how noise may improve the system performance. The figure of merit we use is a modified SNR [38] in which we consider both the noise and the error to the desired output, instead of only noise. This measure is used instead of the usual SNR because it provides more information about the performance of the cell. The peak in this measure appears for non-zero noise amplitude near the region in which the input signal to internal noise ratio is close to 0 dB -- where noise amplitude is equal to the maximum input signal [40]. The output SNR depends on the number of elements in the array. Using a structure with $n$ gates having the same threshold appears to be suboptimal. However, when noise is large it is the optimal setup [38]. The digital output takes advantage of the extra information transferred in the analogue mode to provide a better noise tolerance. Figure 3 right shows a plot comparing the
output error probabilities of the cell and a classical single element gate. The averaging cell clearly outperforms the classical gate when it operates in SR regime.

5.3 Parameter variation tolerance

One of the effects of noise is the linearization or smoothing of functions [36]. This effect plus the redundancy of the cell and the averaging function permits the linearization of the individual characteristics of the gates. Each gate has a transfer function (TF) that depends on the noise amplitudes and its actual parameters. For medium/large noise amplitudes it becomes independent of the actual gate parameters [38, 39]. Figure 4 shows the transfer function for different noise standard deviations (Gaussian noise is considered) for different limiter circuits with infinite gain (left) and gain 2 V/V (right). As noise grows, the TF of both circuits evolves to the same TF independently of the different initial parameters. It can be seen that for noise amplitudes in the SSR regime \( \sigma_n = 0.5 \text{V rms} \) the TF is independent from its actual gain. Thus the cell provides a high tolerance to parameter variations.

5.4 Defect tolerance

Defects in the system can only be addressed either by reconfiguration, redundancy or a combination of both techniques. The array structure with averaging provides an elegant way to deal with large numbers of defects. In this structure, any stuck-at, short or open defect only produce either an offset, gain variation or a combination of both errors at the output by the effect of the averaging function. This makes the cell TF highly robust to defects. Critical defects are transformed to non-critical degradations of the system performance. System immunity to defects can be controlled in the design stage by selecting the number of elements in the array.

6 CONCLUSIONS

Because of the quality reduction in components, due to the inherent nature of sub nanometric devices, conventional design rules of complex digital systems will be no longer efficient. New design concepts and rules are required for both the CMOS technology in the next decade and the presumable new emergent nanotechno devices.

For the CMOS technology, where unpredictable noise and other aggressive mechanisms will be present the use of error-detecting or correcting codes on the silicon circuits are proposed. Similarly to what happen in digital communications an error may be corrected by a recovering strategy.

For future nanometric devices authors present the averaging cell as a promising circuit for future technologies. The averaging cell is not only tolerant to noise and parameter variations but is also optimal for transmitting information in noisy conditions. The cell is tolerant to noise, devices parameter variations and defects. For these characteristics the averaging cell seems a good building block for technologies with poor quality, defective and noisy components as nanotechnology devices are expected to be.

REFERENCES