ELECTRONIC SYSTEM DESIGN PARADIGMS IN THE TECHNOLOGIES OF THE YEAR 2020

Carmen García, Francesc Moll, Antonio Rubio

Department of Electronic Design, Faculty of Telecommunication Engineering, Campus Nord
UPC, c/ Gran Jordi Girona 31, 08034 Barcelona, Spain
cgarcia@eel.upc.edu, antonio.rubio@upc,.edu

Abstract: This paper analyses the evolution of late and beyond CMOS technologies showing the relevant effect on the electronic design methods and rules due to the dramatic reduction of quality of the manufactured components. The paper argues that the main impact on such future technologies will be caused by the increase of the manufacturing defect ratio of components and circuits, the reduction of reliability through a huge increase of the error ratio and a drastic drop of the components quality. Consequently, new strategies of design have to be considered, among them the hierarchical fault tolerance that is presented in the paper. This fact will have implications in the design rules as we know them today.

Key Words: Technology evolution, integrated circuits, power consumption, low voltage operation, switching noise, thermal noise, manufacturing yield, fault tolerance.

1. INTRODUCTION

Today´s most advanced integrated technology shows a miniaturizing level characterized by a manufacturing critical dimension of 32 nm and decreasing. Following the scaling down of components predicted by Gordon Moore [1] we have seen a continuous miniaturization of the device sizes that in the last decade was located around the 350 nm generation. This last decade, from 350 to 90 nm nodes, the generations have been called the Pre-Giga transistor era. From 90 to 22 nm (a predicted feasible technology today) is called Giga transistor era and from 22 nm to the ultimate feasible size (presumably 12 or 6 nm) is called the Tera transistor era. The characteristics of the design techniques and technologies are evolving very fast due to the increase of manufacturing constrains. For the Pre-Giga and older technologies, components were of high quality, with a low manufacturing defect rate that produced a high yield. Because of the high voltage (>3.3 volts), the signal to noise figure (S/N) was very high allowing a robust logic and the "zero defects" objective could be reached thanks to an efficient test technology. To summarize, in general we had highly reliable technologies.

For the Giga and Tera eras things are changing. First, manufacturing is exhibiting critical levels of parameter variability due to manufacturing process limitations. These processing deviations are also reducing the theoretical benefits of miniaturization due to the wider distribution of characteristics of the devices. It is expected that in one or two generations the benefits of shrinking devices may be fully cancelled because the manufacturing deviations. Consequently, components are showing limited quality and causing a moderate if not low manufacturing yield. Moreover, because of the continuous reduction of the power supply voltage (<1.1 volts, because of the reliability requirements for the thin oxide of the transistors) switching noise is causing great problems and a significant performance limitation. Finally due to the high speed and complexity of circuits test technology, the efficiency of years past is being lost and components may reach the market with faults.

As we mention in the following section, new nanotechnologies, based on the so-called emerging research devices (ERD) will keep the trend of miniaturization alive and will increase component density. This effort of miniaturization is the key driving force to future computers, because it is the base of highly powerful processing and memory systems, keeping the continuous reduction of size, power consumption and price per component.These extremely sophisticated and powerful technologies, however, will not be rid of all these problems we have mentioned above, just the opposite. Quality will be extremely low, the defect rate will be very high, and voltage will be scaled so low that the S/N figure will easily reach 0 dB causing a high rate of transient, dynamic or soft errors. Yield will be practically zero and if we do not change design rules testing will be operative no longer.

In Section 2, the characteristics of the technology in the year 2020 and beyond are discussed, analyzing the potential manufacturing yield for those technologies. Section 3 deals with the key limiting aspect of these future technologies: reliability; the two main components of such limiting situation will be shown: power dissipation and thermal limits. Section 4 shows and analyzes the main mechanisms to design acceptably robust systems even from low quality components: the use of fault tolerant redundant circuits. Section 5 introduces a new concept, the hierarchical fault tolerance technique, which presumably will be the future design paradigm with a continuous improvement of reliability thanks to the use of fault tolerant mechanisms in all the design flow level. Finally

Section 6 summarizes a general discussion of the design future and draws several conclusions.

2. THE TECHNOLOGY IN THE YEAR 2020 AND BEYOND

The International Technology Roadmap for Semiconductors (ITRS, [2]) predicts a potential CMOS technology capable of integrating 1000 Giga-transistors in a single integrated circuit for the year 2020. This will imply the capability to integrate between 22 and 40 CPU cores or more than one thousand components like Silicon On Circuits cores or IPs. Components like Resonant Tunneling Devices [3], Single Electron Transistors [4], Molecular electronics [5], Ferromagnetic logic [6], Spin transistor logic [7] and Memristors-based technologies [8] among others, which will

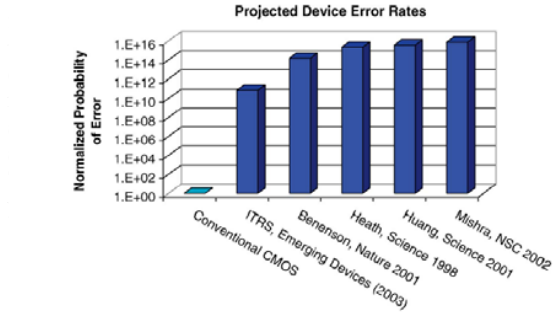

Figure 1. Projected device error rates for future technologies [9].

Probability of component defect	Number of devices			
Error Prob.	10^6	10^8	10^{10}	10^{12}
10^{-6}	0.367	10^{-44}	0	0
10^{-8}	0.99	0.367	10^{-44}	0
10^{-9}	0.999	0.63	10^{-30}	0
10^{-10}	0.9999	0.99	0.367	10^{-44}
10^{-12}	0.99999	0.9999	0.99	0.367

Table 1. Yield prediction for different circuit complexities and component error probabilities.

Nowadays technologies (CMOS-based) are already showing a yield reduction trend. As a matter of example, a mature product yield for a 250 nm technology node was 90%; however, for a 45 nm technology node the effective yield is around 55%. This implies, as it is well accepted, a 5% average yield decrease per node [10].

In order to evaluate this yield (55%) in terms of component defect probability, assuming a component complexity of 10^9 components for the 45 nm technology we calculate a probability of component failure of 10^{-10}.

Table 1 shows the yield prediction for different complexities and component defect probabilities. Observe the sharp trend of yield towards zero when both components increase as predicted in Figure 1 [9]. As a matter of example of potential defect rates for future ERD, we can mention the results published in [11] where a chemically assembled electronic technology with a component density of 10^{10} gate-equivalents/cm^2 exhibits a 0.1 (10%) defect rate, and the molecular electronic memory shown in [12] that with a component density of 10^{11} bits/cm^2 shows a 0.25 (25%) defect rate. If all these technologies would be manufactured massively as we do today with CMOS, the yield would be dramatically inacceptable. However, there is still some hope for future technologies if practically perfect atomic positioning, as have been shown recently by using AFM. In the future it could be compatible with high complexity circuits and high production levels, in this case perhaps we could manufacture perfect components as Gordon Moore predicted in [13] "There is no fundamental obstacle to achieve device yields of 100%". However, in the next section we will discuss a critical problem for future technologies: power consumption.

3. A COMMON DANGER FOR ALL FUTURE TECHNOLOGIES: LACK OF RELIABILITY

In this section we will show that power dissipation is a current limitation for technologies and that there is an unavoidable limit: the thermal limit. Both will affect future circuits with a dramatic reduction of reliability.

3.1 Power dissipation limit.

Power dissipation is a key limiting drawback for modern technologies. As indicated in [14], modern CPUs dissipate around 100 watts getting close to the power density limit of 100 W/cm^2. The trend for the next decade is a huge increase in power density, around 20 times as much as that reached in 2006 for 2018 (see figure 2), and this would cause, even with efficient heat sinks, temperatures in the circuit of around 500 ºC, a value that is not acceptable at all. The only solution for such complex technologies will be to reduce the power supply voltage which has a key impact on power dissipation. However, this effect will also have an impact on the error rate as the following section shows.

3.2 Low Voltage Design

There is a trend of reduction of the power supply voltage in CMOS technologies, due to the stress caused by the electric field in the thin oxide of the transistor. However, it is also a required condition if technologies keep increasing density in order to limit device temperature. Let us now evaluate the operation energy evolution for CMOS technologies until 2020 that is shown in figure 3.

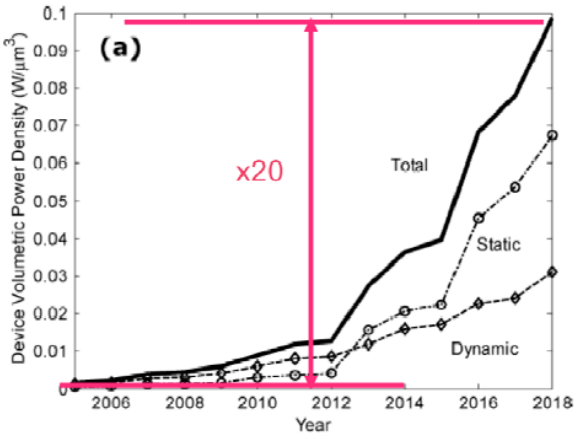

Figure 2. Forecast increase of power density in the next decade [14].

As the figure shows, the energy for a unitary operation (transition in an inverter gate) has been decreasing in the last decades predicting a value of 1 aJ (atto joule) in 2020. This trend agrees also with the switching energy predicted for Emerging Research Devices (Single electron transistors, molecular electronics …) as indicated in ITRS [1].

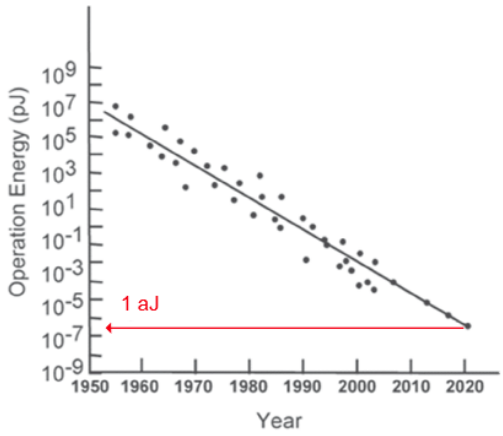

Figure 3. Evolution of the unitary operation energy for CMOS technologies.

The main noise source in modern digital circuits is the one caused by the digital circuit itself when it suddenly switches currents through parasitic components, called switching noise. In [15], a level of switching noise equivalent to 20% of the power supply voltage level in CMOS is predicted for 2020. In the case of digital circuits this voltage fluctuation causes a fluctuation in the propagation delay that is a cause of performance degradation. This could imply an error probability of 10^{-14} in every atomic processing cycle in the year 2020, according to [16]. New design strategies towards the reduction of transitions and even the implementation of clock-less circuits (asynchronous) are being developed. However, even with the advance of such future design strategies there is a factor that cannot be avoided, i.e. the thermal noise caused by molecular agitation of particles in materials due to heat. Even though the level of thermal noise is lower than today's switching noise, it will be a key limit for future technologies (figure 4) as explained in the following section.

3.3 Thermal noise impact

Thermal noise is an unavoidable noise that for low voltage (low power) future technologies could imply a key limiting aspect. In [16], the error rate caused by thermal noise for a given energy operation was investigated. Figure 5 shows this relation, taken from [16], and the vertical arrow indicates the energy corresponding to the CMOS technology in 2020. This implies an error rate of 10^{-28}. If we consider now that the error rate for future ERD technologies is 10^{8} times higher than for the CMOS, we obtain an error rate for processors made with ERD and working at low power of 10^{-20}. This error rate is equivalent to 5 errors per minute, which is too high to be considered a robust technology: it is unacceptable.

Transient faults or errors cause a temporary change in state of the system, without damaging any of the components. Because of their random and non-recurring nature, transient faults are difficult to solve, tolerate or isolate, hence they become a source of major concern, especially in critical real-time application areas.

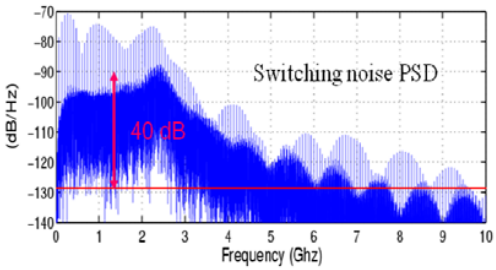

Figure 4. Comparison of switching noise and thermal noise levels (red line).

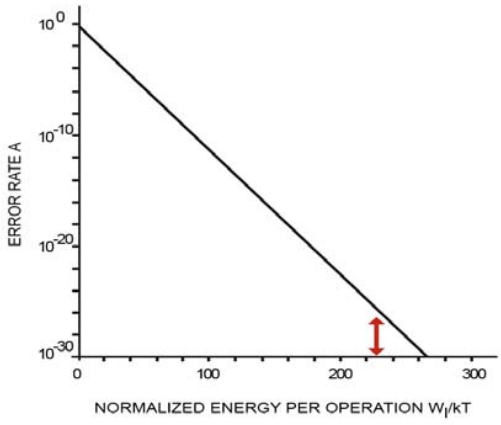

Figure 5. Error rate versus normalized energy per operation. Red arrow marks CMOS technology in 2020 operating at room temperature [16].

3.4 High degradation effects

Gradual circuit performance degradation, as a result of the device aging, is the main limiting factor of structural reliability together with the initial manufacturing defects. In modern integrated circuits and future devices due to the extremely reduced dimensions, electron and holes are gaining high kinetic energies and that carriers may be injected into the isolating oxides causing permanent effects and degrading the

performance of the devices. The degradation phenomena is expected to increase with the shrinking of devices, so it is a important source of failures that requires dynamic fault tolerance [23].

3.5 Unacceptable possible scenario

From the results shown above we can predict a low quality, unacceptable technological scenario for the future. We have seen that in terms of manufacturing yield and degradation, with the techniques used until now we can expect a zero manufacturing yield. Perhaps in the future we will learn to positioning atom by atom in the right place. This would imply a 100% error-free manufacturing yield, but we do not know the efficiency of such a technology in terms of number of manufactured components per unit of time. Power supply is another limiting aspect; with the CMOS voltage scaling down trend we will reach temperatures higher than 500 ºC. In case of drastic voltage reduction this would imply a dramatic reduction of performance in CMOS designs and an unacceptable error rate for ERD technologies in one or two decades. Until now integrated circuits have been designed considering a high quality component scenario, in such a way that with an efficient quality test procedure it was possible to have both a zero defect delivery to market and a high manufacturing yield. In the future, yield will decrease and test techniques could be inefficient (detecting all the components as defective). However, in 1956 [17], the solution to improve reliability of complex systems was already introduced by von Neumann: the use of fault tolerant systems. Fault tolerance implies redundancy of resources and this redundancy can be of hardware, of codes and of time. All three are equivalent and because future technologies will exhibit a huge number of component density we will consider hardware redundancy in the rest of the paper.

4. FAULT TOLERANCE TECHNIQUES

4.1 Successful fault tolerance precedents

One of the most referenced examples related to fault tolerance of systems with a medium/high ratio of defective components is the reconfigurable architecture of TERAMAC designed by HP researchers [18]. In this example, a computer is organized from a reconfigurable structure made up of around 8 million components (220 thousand programmable logic cells, 145 thousand interchip elements and millions of crossbars). The system showed to be efficient to tolerate 10% of defective programmable cells. The principle of tolerance was based on reconfigurability; first a whole testing phase determined the set of defective components, later the systems were reconfigured using healthy components only, avoiding the use of defective ones. This is one of the principles of the reconfiguration technique, test-find-replace. The same technique is used in reconfigurable memories, which integrate sparse components, or even in multi-core systems where in every startup the systems test the cores and cancel the defective or marginal ones.

In spite of the apparent success of this technique, it cannot be used in computers that exhibit transient errors as are those caused by noise, the unavoidable source. The test of the whole systems takes time decades higher than the period of appearance of transient faults. Another drawback is that the method is not efficient if defect rate is very high, which is the case we expect in nanodevice architectures. In the scenario of nanosystems, we will find a significant number of defects at any level, section or subsection.

4.2 Other fault tolerant techniques adequate for transient errors

Hardware fault tolerant structures implemented through the use of massive redundancy have been reported previously [19,20,21]. The more relevant are: NAND multiplexing, N-modular redundancy (NMR) and averaging cells.

The NAND MULTIPLEXING technique was introduced by Von Neumann in 1956 [19]. The design is based on the implementation of a functional gate into three stages: executive stage, restorative stage and output stage. In each one of these stages the redundancy level is given by N (with typical values of 100 or 1000). Figure 6, shows the results of this technique for three different levels of redundancy (N= 10, 100 and 1000). Observe that for high values of N the error probability is limited even for a significant individual gate error rate. However, many publications [20] have shown that this technique has a limitation, because the system is not able to tolerate global errors when the individual error rate is higher than 1% approximately.

The NMR or TMR TECHNIQUES. NMR and TMR are based on the use of a N-tuple or Triple modular redundancy of the basic system. The system evaluates answers from the N or 3 modules and assumes as good the one voted in majority [19]. The system is simple and efficient but only for very low error rates as shown in figure 7.

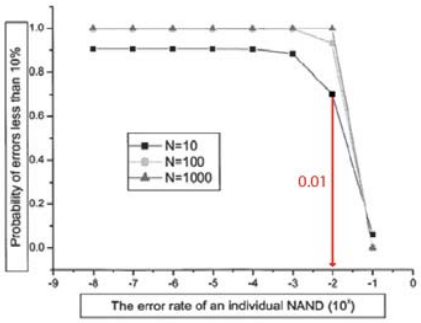

Figure 6. Error tolerance caused by NAND multiplexing versus individual error rate [21].

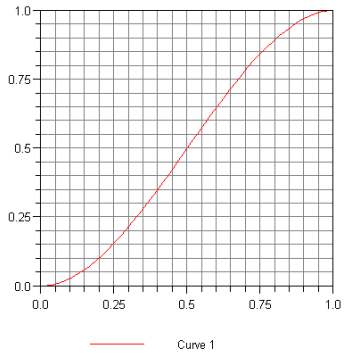

Figure 7. Error tolerance caused by TMR. The vertical axis is the error rate of the TMR system and the horizontal the error probability of each one of the three components.

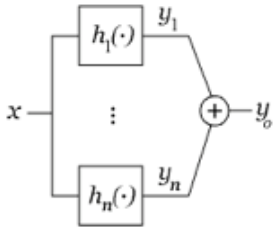

Figure 8. Schematic structure of the tolerant averaging cell

The AVERAGING CELL. This technique is based on the compensation of imperfect components. Published first in [21], it is a very efficient technique for even very high error rates, and it is adequate for molecular electronics where each component is implemented from a bunch of parallel molecules. The system exhibits a very high applicability for defect tolerance, deviation tolerance and noise tolerance. Figure 8 shows the basic structure of the cell and figure 9 shows the comparison of the error tolerance between the averaging cell with two different levels of redundancy and a NAND multiplexing with N=100. Observe that the NAND multiplexing is not efficient for individual error rates close to 0.01 and that the averaging cell satisfies the tolerance principle.

Figure 10 shows the comparison between the different fault tolerance mechanisms, including reconfigurability, showing how they cover different individual error rates with different efficiencies. This fact will be used in the following section to propose a new design technique.

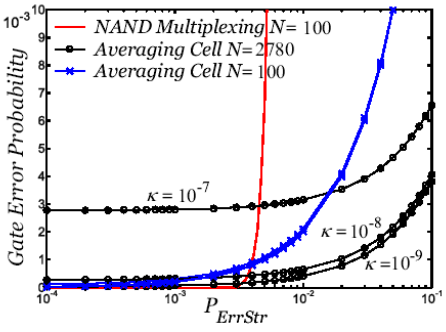

Figure 9. Comparison between Averaging cell and NAND multiplexing [21].

The main points we can conclude from these previous sections are: Fault tolerant techniques are not magical solutions for low quality components. The main effect of these mechanisms is that they usually reduce the error rate of the global block in comparison with the error rate of the individual components. The efficiency of this reliability improvement exhibits a limit for high error rate, at different levels for different mechanisms. The only one that is able to deal with very high error rate is the Averaging cell based on a compensating mechanism. Figure 10 shows that the required redundancy is very dependent of the individual error rate and the fault tolerance mechanism. All these concepts point to the proposal put forward in the following section.

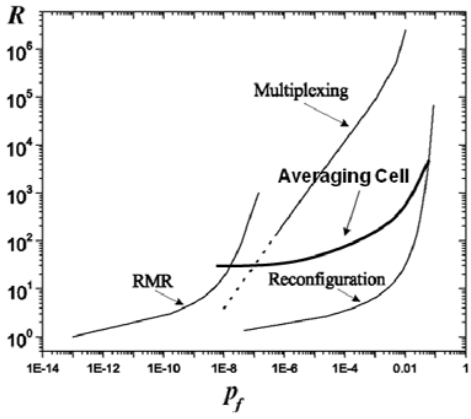

Figure 10. Comparison between the different tolerant mechanisms. Vertical axis is the redundancy factor, R, and the horizontal the component error probability [20].

5. HIERARCHICAL FAULT TOLERANCE STRATEGY

Because of the dramatic impact of very low quality components offered by future nanotechnologies, design methodologies have to take into account design for reliability improvement techniques in the same way than what happened in the past with the introduction of design for testability rules in design flow. In the future, designers will have to set the design of robust designs (adequately robust for the specific application) as their objective and this hardware will have to be implemented from very low quality and low reliable components.

As a practical solution, we propose the use of the hierarchical structure of the electronic design flow as an iterative mechanism of reliability improvement. This idea was first presented in [22]. Figure 11 shows an example of this strategy, where the different design flow levels (technology, transistors, circuit, gates, logic blocks …) are used to improve the reliability (the arbitrary numbers inside the column give idea of a continuous reduction of the error rate). Also in order to illustrate the concept, different fault tolerant mechanisms are applied to every stage of the design.

For many years the mission of testing has been described as screening defects and to a lesser extent speed classification. Perhaps the common mission of tests and design has to be extended to screening for reliability.

Each design layer will base its structure on “components” coming from the previous layer with a given reliability (never perfect components). This layer not only will implement new components of a higher design hierarchy, but it will introduce fault-tolerant mechanisms (those adequate for the error rate level of the components) that will improve (not solve) the reliability offered to the next layer.

This will allow different orders of reliability improvement until the final level is reached with an adequate quality. At the higher level it is expected to find the techniques based on reconfiguration, based on the test and replace principle [11] while in the lower levels, mechanisms as Averaging, NAND Multiplexing and equivalents will tolerate and compensate the dynamic erratic behavior of future devices

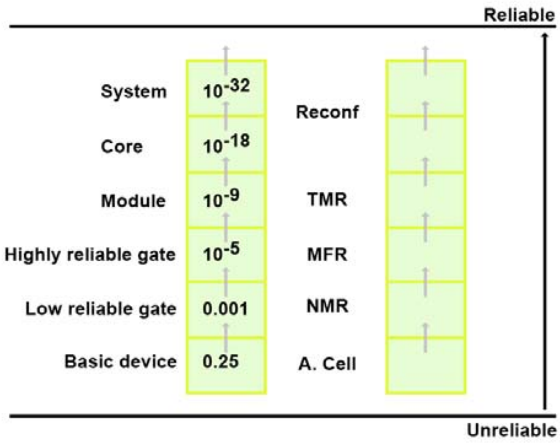

Figure 11. Example of hierarchical multiple coordinated fault tolerance strategy taking benefit from the well known hierarchical design flow.

6. CONCLUSIONS

We can accept the prediction that future technologies will exhibit components with a high defect rate and a high rate of transient errors. The sophisticated manufacturing process and the high level noise caused by the chip itself (switching noise) will be the main limiting point for CMOS technologies in the coming years. These drawbacks or limitations will be very likely solved with the evolution of technologies and new design styles. However there is one source of error that is unavoidable, the one caused by thermal noise. Thermal noise is the cause of a very high transient error rate, due to the equality of the noise and signals levels (the latter because of the necessary power dissipation limitation in order to operate within a feasible temperature range). In order to tolerate these errors the use of hardware redundant circuits like the ones introduced by von Neumann half a century ago will be required. All these mechanisms require massive redundancy, but this will not be a problem in such future technologies where we expect decades of higher component density. However, a detailed analysis of the mechanisms show there are not a panacea. NAND multiplexing as a matter of example with a level of redundancy N=100 is very efficient to tolerate transient error up to an individual error rate of around 0.01. Averaging cell needs a much higher redundancy factor, but it is adequate for very high error rate (even for levels close to 1). TMR is very efficient with a redundancy requirement of 3, but is fails when individual error rate takes moderate values. From all this the authors suggest a hierarchical fault tolerance strategy, where different fault tolerance mechanisms are used at different levels of component complexity. This implies a strategy of reliability-aware technique that can be implemented together with the hierarchical flow of electronic system design. Consequently, the solution for robust systems is not only in the hands of manufacturers, as in the past, but in each and every one of the basic steps of the design flow.

Acknowledgements

These results are part of research project TEC2008/01856 of the Science and Innovation Ministry in Spain and the TRAMS FP7 project of the European Commission, the project TEC is also financed by FEDER founds.

REFERENCES

[1] Gordon E. Moore “Lithography and the future of Moore’s Law”, Proc. SPIE, Vol. 2439; DOI:10.1117/12.209195
[2] International Technology Roadmap for Semiconductors Report, www.sematech.org
[3] P. Mazumder, et al., “Digital circuit applications of resonant tunneling devices”, Proc. IEEE, April 1998, Vol. 86, issue 4. DOI: 10.1109/5.663544
[4] R.H. Chen et al., “Single electron transistor logic”, Appl. Phys. Lett. 68, 1954 (1996); DOI:10.1063/1.115637
[5] Avi Aviram, “Molecular electronics, Science and Technology”, Angewandte Chemie, Vol. 28, issue 4 (2003), DOI: 10.1002/anie.198905201
[6] D.A. Allwood et al., “Submicrometer ferromagnetic NOT gate and shift register” Science 14, June 2002, Vol. 296, no. 5575, DOI: 10.1126/science.1070595
[7] G. Zorpette, “The quest for the SPIN transistor”, Spectrum IEEE, Vol. 38, issue 12 (2001), DOI: 10.1109/6.969357
[8] D.B. Strukov et al., “The missing memristor found” Nature 453, (2008), DOI: 10.1038/nature.06932
[9] AJ KleinOsowski and David J. Lilja. “The NanoBox project: Exploring fabrics of selfcorrecting logic blocks for high defect rate molecular device technologies. In IEEE Symposium on VLSI, pages 1–10, 2004.
[10] Source: Selantec 2007.
[11] M. Mishra and S.C. Goldstein, “Defect tolerant at the end of the roadmap”, Nano, Quantum and Molecular Computing, Springer, US, DOI: 10.1007/b116438
[12] J.E. Green et al., “A 160-kilobit molecular electronic memory patterned at 10E(11) bits per square centimeter” Nature, 445 (2007), DOI: 10.1038/nature05462
[13] G. Moore, “VLSI: Some fundamental challenges”, IEEE Spectrum 1979.
[14] E. Pop et al, “Heat Generation and transport nanmmeter-scale transistors” Proceedings IEEE, Vol. 94, issue 8, (2006), DOI:10.1109/JPROC.2006.879794
[15] D. Andrade et al., “Power supply noise and logic error probability”, Proc. European Conference on Circuit Theory and Design 2007, DOI; 10.1109/ECCTD.2007.4529559
[16] K.U. Stein, “Noise-induced error rate as limiting factory for energy per operation in digital ICs”, IEEE Journal Solid-State Circuits, Vol. 12, issue 5, 1977, pp. 527-530.
[17] J. von Neumann et al., “Automata Studies”, Annals of Mathematical Studies, Princeton University Press.
[18] R. Amerson et al., “Teramac configurable custom computing”, IEEE Symp. On FPGAs for Custom Computing Machines, 1995, DOI: 10.1109/FPGA.1995.477406
[19] K. Nicolik et al., “Fault-tolerant techniques for nanocomputers”, Nanotechnology 13, 357-362, 2002, DOI: 10.1088/0957-4484/13/3/323
[20] K. Nicolik et al., “Architectures for reliable computing with unreliable nanodevices” IEEE Conf. on Nanotechnology, 2001, pp. 254-259, DOI: 19.1109/NANO.2001.966429
[21] F. Martorell et al., “An analysis of internal parameter variations effects on nanoscaled gates” IEEE TR. On Nanotechnology, Vol. 7, no. 1, 2008, pp. 24-33, DOI:10.1109/TNANO,2007.913429
[22] S. Cotofana et al., “A design exploration framework for reliable nano electronics”, Proc. 2005 IEEE Conf. on Applied Specific Systems, Architecture Processors, pp. 260-267.
[23] Shekhar Borkar, "Designing Reliable Systems from Unreliable Components: The Challenges of Transistor Variability and Degradation," IEEE Micro, vol. 25, no. 6, pp. 10-16, Nov./Dec. 2005, doi:10.1109/MM.2005.110