Performance/Complexity Space Exploration : Bulk vs. SOI

S. J. Abou-Samra and A. Guyot

TIMA laboratory
46, av. Félix Viallet, F-38031 Grenoble - France
Tel : (+33) 476 57 48 12 Fax : (+33) 476 47 38 14
E-Mail: Abou-Samra@imag.fr

Abstract : Performance and complexity are considered here as two orthogonal axes. Performance metrics are recalled. Then different complexity metrics and scales are proposed. Different definitions of complexity are used depending on the considered level of abstraction. Finally, SOI and bulk CMOS technologies are compared in this space.

1. INTRODUCTION

It is now well known [1, 2, 3] that SOI technologies deliver higher “performance” than their bulk counterparts. But today SOI designs are ported from bulk with only slight modifications and without taking advantage of some design subtleties that are SOI-specific [4]. One of these subtleties is the better use of complex cells that one can make in SOI compared to bulk. We will see how the reduced source/drain to body capacitance favours complex cells. Higher performance can then be achieved, but first, detailed explanations of what is meant by performance and complexity are given. Only the gate and transistor levels are addressed here, and not the architecture and algorithm ones [13]. This paper is structured as follows : In the first part, performance metrics are recalled, and the different figures of merit used are detailed. In the second part, an attempt is made to quantify and measure complexity : complexity scales are introduced. In the third part, comparative results for bulk and SOI CMOS are shown and discussed; finally, some concluding remarks are given as well as a brief presentation of the related work.

2. WHAT'S PERFORMANCE ?

The word performance is subjective and the way it is commonly used in microelectronics can sometimes be misleading, as many things can be put behind depending on what is expected : throughput, energy per operation or energy per throughput. The problem of performance metrics has been addressed by C. Piguet [5, 6] amongst others. An overview of the different components of performance are defined or reminded here.

This work is supported by the ESPRIT project HIPERLOGIC (# 20023)

2.1. Throughput

If one is looking for the best throughput, then performance will be the delay D, regardless to any other parameter (area, energy consumption etc.). The delay is a decreasing function with V_{dd}. As a rule of thumb, it can be assumed that

$$D \sim \alpha.V_{dd}/(V_{dd}-V_t)^2 \qquad (1)$$

Where α is a constant depending on the technology for a given design style. This equation does not account for short channel effects, it is only a first order approximation. From equation (1) one can see that the delay requirements can always be fulfilled by increasing the supply voltage - in the limits of the technology tolerance and the model boundaries though (fig. 1). The delay is measured in nanoseconds. The throughput is the inverse of the delay, and is given in MHz.

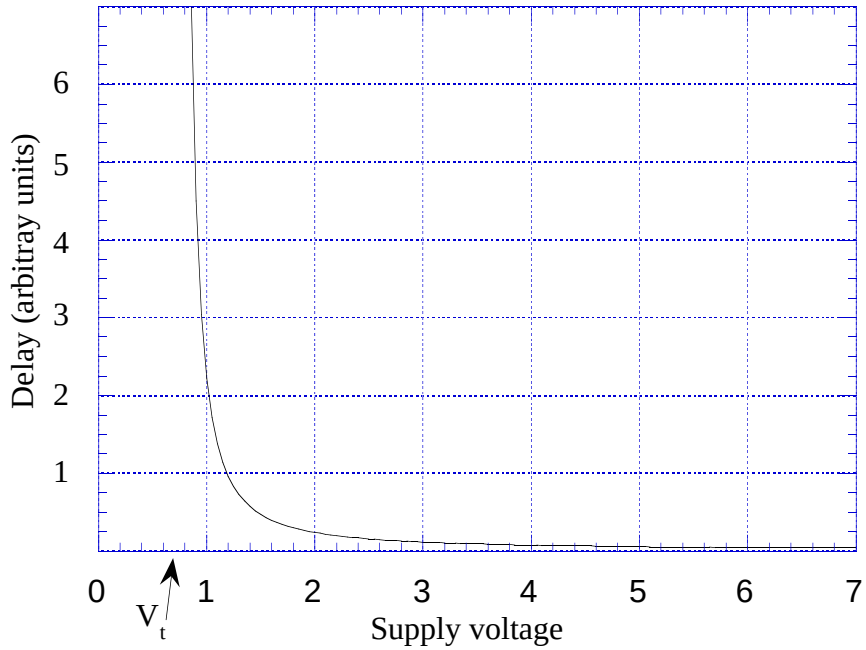

Figure 1: Gate delay vs. supply voltage

Better delay models [12] are of course needed for accurate predictions.

2.2. Battery life

If the only constraint for a design is battery life (like in a watch for example), then the energy per operation is the metrics to use. Energy per operation EPO is the same as power delay product PDP, but the term “Energy Per Operation” is preferred here, as the delay (or the throughput) is not a constraint. Indeed, the energy required for an operation is proportional to V_{dd}^2 (say $E = \beta.V_{dd}^2$) and the power dissipated depends on the frequency :

$$P = \beta.V_{dd}^2.f = \beta.V_{dd}^2/D; \text{ thus, } P \times D = PDP = \beta.V_{dd}^2 = EPO \qquad (2)$$

So, EPO is a monotone function of V_{dd} (fig. 2) meaning that it can be made as small as required simply by lowering the supply voltage. Of course, this strategy is not compatible with increasing the throughput, and furthermore, there are technological limits for the reduction of the supply voltage. The EPO also depends on the architecture, i.e. on the operation to perform. This can be influenced by parallelism or pipelining schemes [6] that do not appear in eq. (2). The EPO is measured in picojoules.

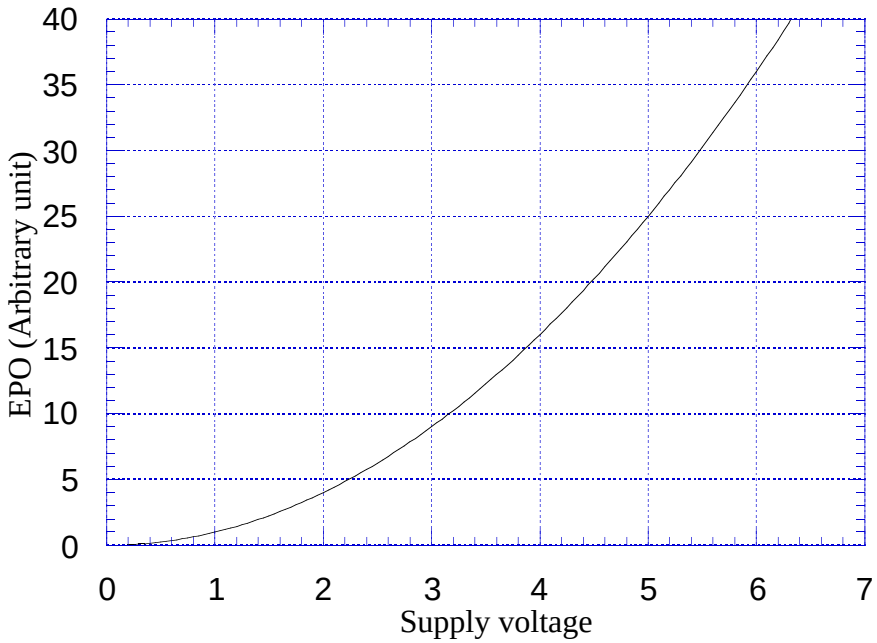

Figure 2: EPO vs. supply voltage

2.3. Efficiency

And what if both constraints, throughput and battery life, need to be satisfied? This means that the EPO has to be minimised in the same time as the delay. A compromise will be necessary, as, on one hand EPO is improved by lowering V_{dd} (fig. 2) and on the other hand, delay improvement requires higher V_{dd} (fig. 1). This compromise is the minimum of the Energy Delay product EDP.

$$EDP = EPOxDelay \tag{3}$$

Using the coarse power and delay models shown in the previous paragraphs, it is easy to solve :

$$\partial EDP / \partial V_{dd} = 0 \tag{4}$$

The result is $V_{dd} = 3V_t$. Actually, this is a first order approximation, further refinements require more technology based considerations. The optimal supply voltage is lower in SOI than in bulk [1]. On figure 3 the EDP is plotted against V_{dd} for a ring oscillator designed in a 0.1µm technology [7, 11, 14] with $(|V_{tn}| + |V_{tp}|)/2 = 0.7V$. the

EDP is measured in pJ/MHz. In this paper, the energy delay product will be called “performance”, and the nearly “optimal” supply voltage of 2V (fig. 3) will be used for further optimisation of the EDP in the performance complexity space for both SOI and bulk technologies.

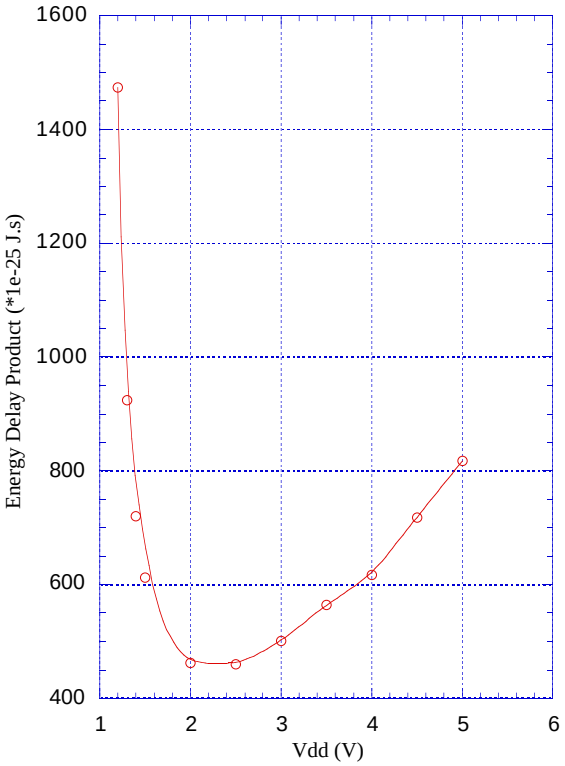

Figure 3: EDP vs. V_{dd} for bulk S_i

3. COMPLEXITY QUANTIFICATION

At first sight, one might be tempted to define complexity as being equal to the number of transistors regardless to the level of abstraction - very simple! But as this might be true at the system level, it becomes false at the gate level. In fact, the complex cell implementation of a given function takes usually less transistors (fig. 4). The reduction of the transistor count has a cost : a logical cost and a performance cost. The logical cost is the loss of the internal nodes (γ on fig. 4). This is a purely formal cost as it does not affect the behaviour of the circuit. The performance cost will be discussed in the following sections.

Figure 4 : Building complex gates

Gate decomposition for higher speed or lower power is a well explored topic [8, 9,

10]. Alternative implementation of a logical function are given, but always without a complexity scale. The attempt is made here to define such a scale.
Complexity metrics actually depend on the level of abstraction one is dealing with. Here, only the gate and transistor levels will be considered. In this section, complexity quantification schemes are proposed, and the experimental protocol is described.

3.1. Gate level

At the gate level, the complexity of a design is related to the length of the critical path, thus at this level of abstraction it seems suitable to define complexity as being the logical depth needed to achieve a given logical function. In this case, a Nand 16 is taken. It can be implemented in one step - with 16 transistors in series (corresponding to Complexity = 1), in three steps like in figure 5a (complexity = 3) or in five steps like in figure 5b which corresponds to a complexity of 5. Other decomposition schemes are possible, but only those in which all gates have the same number of inputs are retained. Thus, the complexity scale at the gate level does not depend on the physical implementation (transistor level) addressed in the next section. The different performance figures can now be plotted on this complexity scale. The experimental protocol is described later. It should be noted that this definition of complexity depends on the logical function performed.

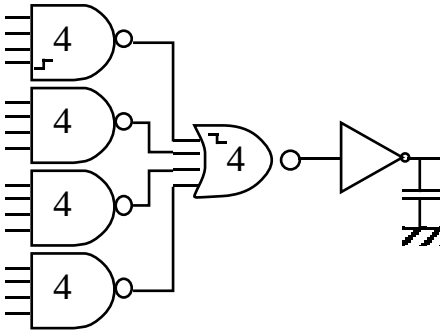

Figure 5a : Complexity = 3

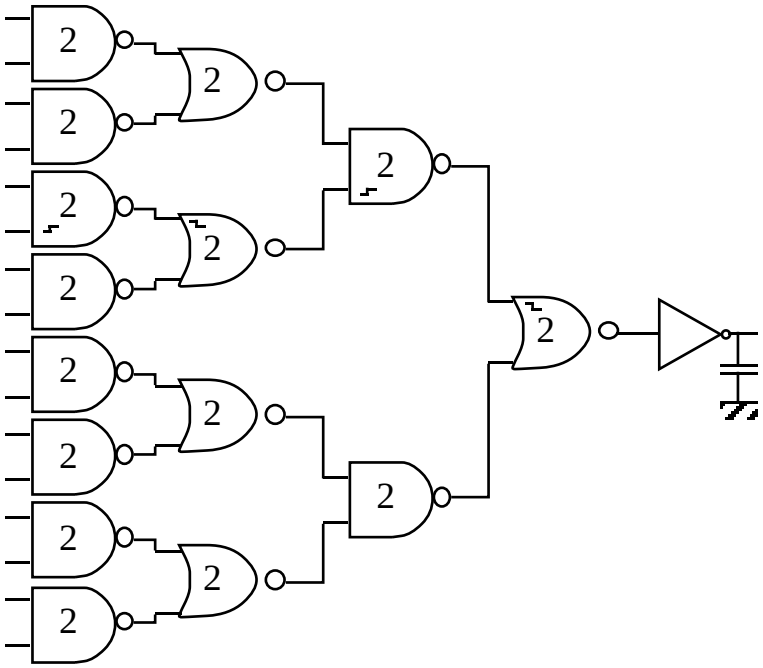

Figure 5b : Complexity = 5

3.2. Transistor level

A finer granularity is needed at the transistor level, and, at this level, the complexity of a gate does not depend on the function it performs. Complexity can though be considered as being proportional to the number of transistors in series. Comparisons can

then be easily performed after normalisation. On this scale, the results are divided by the number of bits, and thus all figures are normalised per bit.

3.3. Experimental Protocol

This section discusses the benchmark circuits used to explore the performance/complexity space. All results are obtained by HSPICE simulations.

3.3.1. Gate level benchmarks

The benchmark used here is a 16 input Nand gate. It is convenient to choose a number of inputs that can be written as 2^{2n}, thus the gate can be decomposed in a balanced tree like manner. For example, the Nand 16 can be decomposed using only 4 input gates (fig. 5a) or 2 input ones (fig. 5b).
The load capacitance is the same for all cases : C_l = 100fF. The measured delay is worst case, i.e. we make sure to always switch the transistor which is the further from the output for all gates. These inputs are marked on figures 5a and 5b.
The EPO is taken as the average of the current flowing through V_{dd} over the duration of the transition, times V_{dd}, times the gate delay.

3.3.2. Transistor level benchmarks

As the scale here is the number of transistors in series, the logical function performed doesn't matter. We assume here a Nand gate with 2 to 8 inputs (fig. 6). All figures are given per bit, thus the loading capacitance has to be proportional to the number of inputs. The load is C_l = n.50fF where n is the number of inputs of the gate. The measured delay is worst case, i.e. the closest input to ground changes.

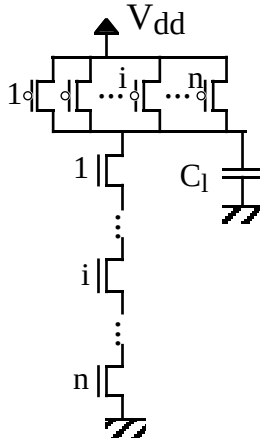

Figure 6 : n input Nand gate

Electrical simulations were carried out for the bulk and the SOI versions of the same technology; the results are discussed in the following section.

4. RESULTS AND DISCUSSION

A bulk and a SOI versions of the same CMOS technology are compared here using the benchmarks discussed in the previous section.
The figure 7 shows the EPO and the EDP as a function of the logical depth for a 16 input Nand gate loading 100fF.

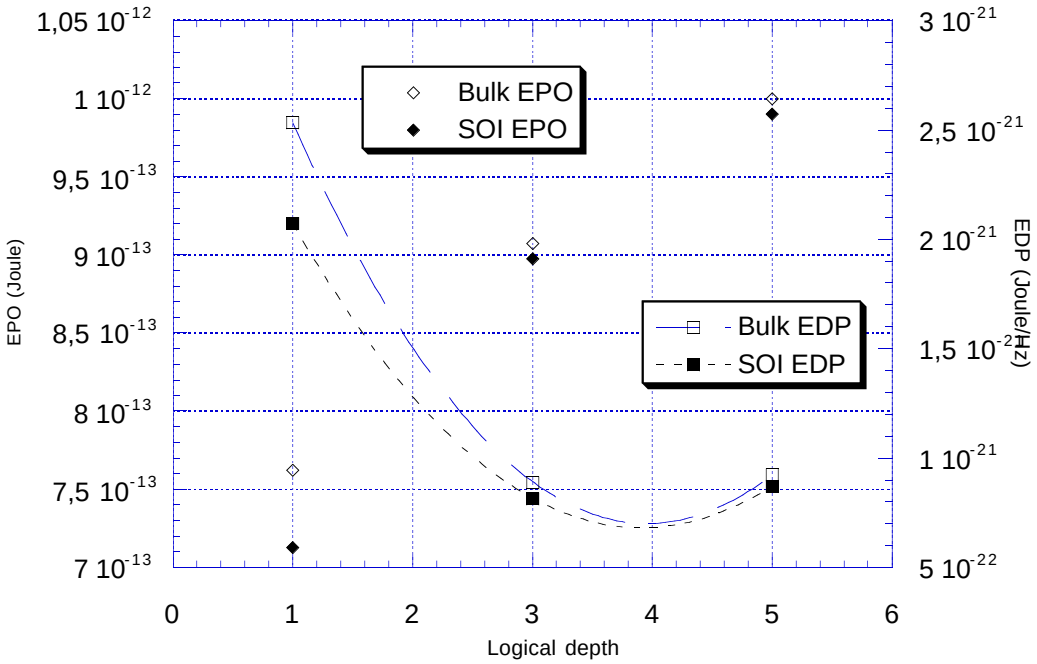

Figure 7 : EPO and EDP vs. logical depth

On the figure 8 the EPO/bit and the EDP/bit are plotted against the number of transistors in series.

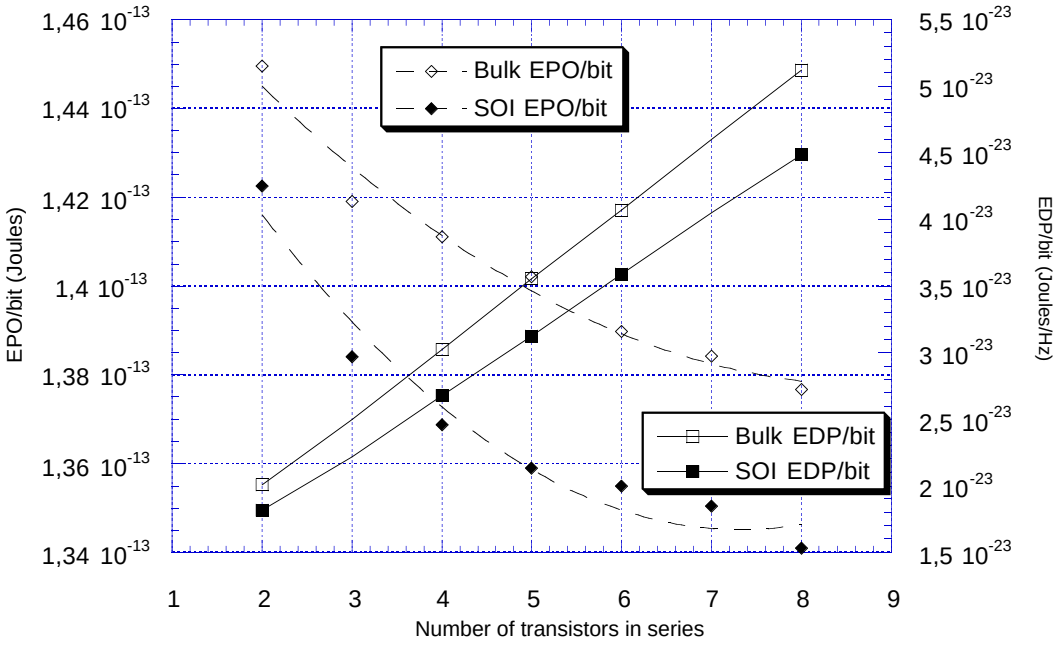

Figure 8 : EPO/bit and EDP/bit vs. number of transistors in series

Some comments and design directives can be drawn from these results :

• There is an optimal decomposition scheme for large functions in terms of EDP but not in terms of EPO (fig. 7). In this case, the implementations with logical depth equals 3 and 5 (fig. 7) give approximately the same EDP. Then of course, one should choose the implementation with the lower EPO and transistor count. The logical depth that minimises the EDP depends on the function being analysed and on the technology.

• From figure 8 it can be seen that the difference in terms of EDP/bit between SOI and bulk increases with the number of transistors in series. This means that SOI favours complex gates design. This increasing difference comes mainly from the very small junction capacitance in SOI as compared to bulk, leading to faster switching of a larger number of transistors in series. The EPO/bit curves for bulk and SOI are parallel with a lower value for SOI.

• For both SOI and bulk, the EPO/bit decreases with complexity (the number of transistors in series) where the EDP/bit increases (fig. 8). Thus, whether using SOI or bulk, if the battery life is the critical parameter, then one should make extensive use of complex cells. On the other hand, if the EDP (or energy per throughput) is the main issue, then complex function should be mapped on smaller cells.

5. CONCLUSION

The originality of this paper is the introduction of quantitative measures for complexity. The scales that are introduced are adapted to the considered level of abstraction. Also, a quantitative comparison between bulk and SOI is carried out, enabling to make a better use of SOI technologies for higher speed and lower energy consumption. The small differences shown here between SOI and bulk are due to the fact that in these benchmarks the loading capacitance C_l are the same for both technologies. This is of course not the case in real designs where the loading capacitance are usually 25 to 30 percent smaller in SOI which leads to much higher improvements in terms of energy delay product (a factor of 3 to 5) [1].
This work is part of wider investigations concerning design methodologies and design tuning for a three dimensional 0.1μm SOI on SOI technology [7, 11, 14].

6. REFERENCES

[1] J.-P. Collinge, *"Silicon-On Insulator Technology : Materials to VLSI"*, 2nd Edition, Kluwer, 1997, ISBN : 0-7923-8007-X

[2] P. Zdebel, *"Low Power/Low Voltage CMOS Technologies, A Comparative Analysis"*, Microelectronics Engineering, Vol. 39, Elsevier, Dec. 1997, pp. 123-137

[3] B. A. Chen et al., In proc. of the 4th Intl. Conference on Solid-State and Integrated Circuits Techniques, 1995, p. 260

[4] S. R. Wilson, M. A. Mendicino, M. L. Alles, *"TFSOI Circuit Applications"*, In proc. of the 8th International Symposium on SOI Technology and Devices (ECS'97), Paris, France, September 1997, pp. 359-372

[5] C. Piguet, *"Circuit and Logic Level Design"*, in Low Power Design in Deep Submicron Electronics, ed. by W. Nebel and J. Mermet, Kluwer, 1997, ISBN 0-7923-4569-X, pp. 105-133

[6] C. Piguet, *"Low-Power and Low-Voltage CMOS Digital Design"*, Microelectronics Engineering, Vol. 39, Elsevier, Dec. 1997, pp. 179-208

[7] S.J. Abou-Samra, V. Dudek, F. Ayache, A. Guyot, B. Courtois and B. Höfflinger, *"Designing With 3D SOI CMOS"*, In proc. of the 8th International Symposium on SOI Technology and Devices (ECS'97), Paris, France, September 1997, pp. 384-388.

[8] J.-M. Masgonty et al., *"Technology and Power-Supply -Independent Cell-Library"*, IEEE CICC'91, San Diego, CA., USA, May 12-15 1991

[9] C. Piguet et al., *"Low Power Design of a Standard Cell Library"*, Low-Voltage Low-Power Workshop during ESSCIRC'95, Lille, France, Sept. 22 1995

[10] C. Mead, M. Renn, *"Cost and Performance of VLSI Computing Structures"*, IEEE JSSC-14, April 1979, pp. 455-462

[11] S.J. Abou-Samra, P. A. Aisa*, A. Guyot and B. Courtois, *"3D CMOS SOI for High Performance Computing"*, In proc. of the 1998 International Symposium on Low-Power Electronics and Design (ISLPED), Monterey, CA, August 10-12, 1998 (* University of Bologna, Bologna, Italy)

[12] D. Auvergne, J. M. Daga and S. Turgis, *"Power and Delay Macro-Modelling for Submicronic CMOS Process : Application to Low Power Design"*, Microelectronics Engineering, Vol. 39, Elsevier, Dec. 1997, pp. 209-233

[13] J. Smit, *"Energy Complexity & Archtecture"*, In proc. of the 7th International Workshop Power and Timing Modeling Optimization and Simulation (PATMOS'97), Louvain la Neuve, Belgium, September 1997, pp. 357-368

[14] S.J. Abou-Samra, J. Arweiler* and A. Guyot, *"Low Power SOI CMOS Multipliers : 2D vs. 3D"*, In proc. of the 24th European Solid State CIRcuits Conference (ESSCIRC'98), The Hague, The Netherlands, September 22-24 1998 (* Technische Hochschule Darmstadt, Germany)

⌘