Abstract

The design of Low Power Systems-on-Chips (SoC) in very deep submicron technologies becomes a very complex task that has to bridge very high level system description with low-level considerations due to technology defaults and variations and increasing system and circuit complexity. This paper describes the major low-level issues, such as dynamic and static power consumption, temperature, technology variations, interconnect, DFM, reliability and yield, and their impact on high-level design, such as the design of multi-Vdd, fault-tolerant, redundant or adaptive chip architectures. Some very low power System-on-Chip (SoC) will be presented in three domains: wireless sensor networks, vision sensors and mobile TV.

I. INTRODUCTION

With the introduction of very deep submicron technologies as low as 45 nanometers and tomorrow down to 32 and even 22 nanometers, integrated circuit (IC) designers have to face two major challenges: first, they have to take into account a dramatic increase in complexity due to the number of components including multi-core processors (“More Moore”) but also due to the significant increase in heterogeneity (“More than Moore”). Secondly, the significant decrease in reliability of the components needs to be taken into account, in particular with the behavior of switches that are very sensitive to technology variations, temperature effects and environmental conditions.

II. INTERDEPENDENCY FROM LOW LEVEL TOWARDS HIGH LEVEL

The interdependency between low level issues mainly due to very deep submicron technologies, and high-level issues related to SoC design, is a major design issue today. One can think that the gap between low level and high level is larger and larger, with the risk that high level designers could totally ignore low level effects and produce non working SoCs. Leakage power, technology variations, temperature effects, interconnect delay, design for manufacturability, yield, and tomorrow “beyond CMOS” unknown devices (ENIAC), are the main low level design aspects that have to be shifted to the high level synthesis. They will impact the high level design methodologies (ARTEMIS), for instance, by rethinking the clocking scheme of processor architectures, by the introduction of redundancy and fault-tolerance, by increasing the number of processor cores, by using multi-voltage domains or by using more dedicated techniques to reduce dynamic and static power. An example of big impact of the low level on high level design is interconnect delays. They are increased due to the smaller and smaller section of wires distributing the clock. So alternate architectures are clockless or asynchronous architectures, moving to multicores organized as GALs (Globally Asynchronous and Locally Synchronous) and using Networks-on-Chips.
A. Dynamic Power

Many techniques have been proposed (and some are widely used today) for reducing dynamic power. One has in a non-exhaustive list gated clock, logic parallelization, activity reduction, asynchronous, adiabatic, bus encoding, standard cell libraries, complex gate decomposition and transistor sizing. The gated clock technique is widely used (to cut the clock when the unit is idle). Parallelism has a strong impact on high level design. Working with many parallel cores or execution units at low supply voltage is always beneficial for the dynamic power. However, it is another story for leakage due to the significant increase in terms of number of transistors.

\[ P = M \cdot C \cdot \frac{f}{M} \cdot V_{dd}^2 = C \cdot f \cdot V_{dd}^2 \]

So the dynamic power is reduced as Vdd can be reduced due to the M times longer clock period.

B. Impact of Leakage on Architectures

The leakage current of switches when they are off is becoming a very dramatic problem regardless of the technology, may it be CMOS, carbon nanotube (CNT) or nanowires. Leakage power increases exponentially with decreasing threshold voltage VT, implying that a significant part of the total power can be leakage for large SoCs. The wasted power is very dependent of the external conditions, such as the chosen technology, the values of VT, the duty cycle defined by the application, etc... There are many techniques [3] at low level and circuit level for reducing leakage, such as using sleep transistors to cut the supply voltage for idle blocks, but other techniques are also available (such as several VT’s, stacked transistors, or body biasing).

In addition to circuit-level techniques, the total power consumption can also be reduced at architectural level. Specific blocks can be operated at optimal supply values (reduced Vdd reduces dynamic power), and optimal VT (larger VT reduce static power) for a given speed, in order to find the lowest total power \( P_{tot} \) depending on the architecture of a given logic block. Therefore, between all the combinations of \( V_{dd}/VT \) guaranteeing the desired speed, only one couple will result in the lowest power consumption [4, 5]. The identification of this optimal working point and its associated total power consumption are tightly related to architectural and technology parameters. This optimal point is depending on activity (a) and logical depth (LD). A not too small activity is preferred in such a way that dynamic power would be not negligible versus static power. A small LD is preferred as too many logic gates in series result in gates that do not switch sufficiently. A gate that does not switch is useless as it is only a leaky gate. The ratio between dynamic and static power is thus an interesting figure of merit, and it is linked to the ratio between \( I_{on}/I_{off} \) of the technology. This ratio is smaller and smaller due to leaky transistors. In [4], this ratio is related to the activity (a) and the logical depth (LD) with the following formula:

\[ \frac{I_{on}}{I_{off}} = k_1 \cdot LD/a \]

With a small \( I_{on}/I_{off} \) ratio (100 to 500), it can be observed that LD has to be small and activity quite large. This implies a clear paradigm shift, as activity has been until now a main factor to be reduced, because only the dynamic power was considered. When both static and dynamic power are considered, the activity should not be as small as possible, as very inactive gates or transistors are leaky devices. The parameter \( k_1 \) is the ratio between dynamic and static power; it is roughly between 1 and 5. This optimal power has been estimated for eleven 16-bit multiplier architectures. Figure 3 shows that too sequential multiplier architectures (at the right of the picture) present a very large total power due to the fact that they are not fast enough and consequently have to be operated at high Vdd (large dynamic power) and very low VT (large static power). Conversely, reasonably parallel multiplier architectures such as the Wallace Tree present the best total power. However, if one increases the parallelism too much (2- or 4-Wallace trees in parallel), even at very low Vdd and high VT, leakage power and total power re-increase due to the very large number of logic gates.

C. Interconnect Delays

The wire delays are a main issue: for every technology node with a reduction factor \( S \), the wire delay is increased by a factor \( S^2 \)!! It is a severe problem for busses, but it is an extremely dramatic problem for clock distribution.
Consequently, the influence on architectures is large: everything could be clockless (asynchronous) or GALS (Globally Asynchronous and Locally Synchronous). Any architecture becomes an array of $N \times N$ zones (isochronous), so it leads naturally to multicore architectures and to massive parallelism with very difficult synchronization problems. For such architectures, it is mandatory to consider NoC (Network-on-Chip) for designing efficient complex SoCs.

D. Process Variations

On the same die, there are technology variations from transistor to transistor, which can be systematic or random due to oxide thickness variations, small difference in $W$ and $L$ transistor dimensions, doping variations, temperature and effects of $V_{dd}$ variations. Many of these variations impact the $V_T$, which can impact the delay variations by a factor 1.5 and leakage by a factor 20. Other effects have not to be neglected, such as soft errors. On the overall, these effects have a very dramatic impact on yield and consequently on the fabrication cost of the circuits. In addition to their low-level impacts, the variations described above also affect higher levels. An interesting impact is the fact that multi-core architectures, at the same throughput, are better to mitigate technology variations than single core architecture. With multi-core architecture, one can work at lower frequency for the same computation throughput. Consequently, the processor cores (at lower frequencies) are less sensitive to process variations on delay. At very high frequency, even a very small $V_T$ variation will have a quite large impact on delay variation.

For over-100nm technologies, Adaptive Body Biasing (ABB) is a good technique for compensating the variations [6, 7]. Since ABB changes the $V_T$ value directly, it can control both leakage and delay. Also, the overhead of this technique is small. This technique is very good but has three important weaknesses. First, using ABB for compensating intra-die variations of NMOS transistors need triple-well technology. Second, the increased short-channel effect due to scaling has decreased the body factor of bulk-CMOS drastically. According to the foundry data, at 65-nm technology, ABB can change $V_T$ value effectively less than 60 mV. This amount is much less than PV and temperature effects. And third, body factor is almost zero in emerging Multi-Gate devices which are promising candidate for future electronics [9]. In addition, in multi-gate devices (double-gated FinFET, tri-gated, gate-all-around or GAA), body factor is much smaller than in single-gate devices because of the enhanced coupling between gate and channel. Measurements in [8] show that in GAA devices body factor is exactly zero. So we need to find new compensation techniques as replacements of ABB.

Looking at standard cell libraries and digital block design, some rules could be given regarding technology variations. Resistance to technology variations is better with long critical paths, as the technology variations are better compensated with a large number of cells connected in series. For the same logic function, a way to have more cells in a given critical path is to provide a standard cell library with few simple cells, as shown in [10]: “It can be shown that with a small set of Boolean functions … (and careful selection of lithography friendly patterns)...we mitigate technology variations”. For designing digital block architecture, one can ask the following question: for a full adder, which is the best architecture (ripple carry, carry look-ahead, etc…) and $V_{dd}$ for reducing the effect of technology variations? A ripple carry adder at 500 mV provides same speed and same power than a carry look-ahead adder at 400 mV with 2 times less sensitivity to PV. Using low-power slow circuits in higher $V_{dd}$ voltage is better than using high-power fast circuits in lower $V_{dd}$!

PCMOS or Probabilistic CMOS, is a new very promising technique [11]. It is based on the fact that each logic gate has a probability of failure. So it characterizes an explicit relationship between the probability ($p$) at which the CMOS switch computes correctly, and its associated energy consumed by each switching step across technology generations. Each basic logic gate (NOT, NAND, NOR) has a given probability to provide a correct result for a given input. For instance, a truth table indicates that for input 100 (correct output is “0”), probability for the output to be “1” is ¼ whereas probability for the output to be “0” is ¾. Using such basic gates to synthesize more complex functions (adder, flip-flops, etc…), over many different schematics that perform the same function, the optimized schematic is chosen in such a way of minimizing the probability of failure.

Logic circuits based on transistors operated in weak inversion (also called subthreshold) offer minimum possible operating voltage [12], and thereby minimum $P_{dy}$ for a given $P_{sta}$. This technique has been revived recently and applied to complete subsystems operated below 200 mV. It has been demonstrated that minimal energy circuits are those operated in subthreshold regime with $V_{dd}$ below VT, resulting in lower frequencies and larger clock period. Therefore, dynamic power is reduced, static power is decreased, although the static energy is increased as more time is required to execute the logic function, meaning that there is an optimum in energy. This optimal energy is also depending on logic depth and activity factor [13]. The minimal $V_{dd}$ (and minimal energy) is smaller for small logical depth and for large activity factors. Reference [14] shows this optimum for $V_{dd}$=0.4 Volt with VT at 0.4 Volt.

Another approach is to introduce spatial or timing redundancy for implementing Fault-Tolerant architectures. It is a paradigm shift, as any system would not be composed of reliable units, but one has to consider that every unit could fail, without inducing the entire system to fail. A possible architecture is to use massive parallelism while presenting redundant units that could take over the work of faulty units. One can have spatial redundancy (very expensive) or timing redundancy (quite expensive in terms of throughput). However, all redundant architectures face the same problem: the overhead in hardware or in throughput is huge, which is a contradictory effect for energy efficient architecture. An example for limiting the hardware overhead is to compare the result of a given operation at 2 different time frames. But as the same operation is executed two times, it reduces the throughput by a factor of 2.
E. Yield and DFM

For very deep submicron technologies, the smallest dimensions of transistor geometries on the mask set are well below the lithographic light wavelength. This yields a variety of unwanted effects, such as bad end line extension, missing small geometries, etc... They can be corrected by OPC (Optical Proximity Correction) which is a technique available for DFM (Design For Manufacturability). However, to facilitate the process of mask correction by OPC, it is recommended to have regular circuit layout. Regular arrays implementing combinational circuits like PLA or ROM memories are therefore more and more attractive. Figure 4 shows three examples of regular layout. A first example back to 1988 [15] is shown at right of Fig. 4 in micronic technology, called gate-matrix style. It was used to facilitate the automatic layout generation. The two other pictures describe a SRAM cell as well as nanowires [16] for which it is mandatory to have very regular structures. This has a huge impact on architectures and systems: SoC architectures should be based on regular arrays and structures, such as PLAs and ROMs for combinational circuits and volatile memories such as SRAM for data storage. Consequently, SoC design should be fully dominated by memories and array structures.

F. Alternative Energy Sources

SoCs used in portable devices may be powered by a variety of energy sources and sometimes energy will be scavenged from the environment. Primary or rechargeable batteries may be used and ultimately miniature fuel cells. To implement energy scavenging, one could use vibrations, thermoelectricity, solar cells, human energy sources, etc... Considering the SoC itself, one has to generate inside the chip multiple supply voltages with very diverse peak currents (some µA, some mA, up to 10 or 100 mA). This requires « Power Management » circuits that may be very complicated circuits (DC-DC, regulators) in particular for high-efficiency implementations required by low-power applications. On top of this, one requires to add DVS and DVFS (Dynamic Voltage Frequency Scaling). It turns out that the power management circuit has to manage many aspects, i.e. energy sources, the multiple supply voltages that have to be generated, DVFS as well as idle modes, resulting in a complex control that is most of the time performed in software by the Operating System. In addition, this part of the embedded software has to interact with the application embedded software, which increases the overall complexity.

G. Complexity

With technology scaling, increasingly more low-level effects have to be taken into account. Consequently, the impacts of these low level effects on to the high level SoC synthesis process are more and more difficult to understand and to be taken into account. Only the low level effects have been presented here, but there are also effects at high level that have to be taken at low level, such as architectures for executing efficiently a given language, asynchronous architectures requiring special Standard Cell Libraries or parallelizing compiler onto N processors, and their constraints on to the processor architectures.

III. HETEROGENEOUS SOC EXAMPLES

This Section shows some SoC examples designed at CSEM for research projects or for industrial customers. These circuits are extremely low power SoCs for radio communication, image recognition or mobile TV applications. The first SoC is called WiseNET [17] and is a circuit designed for supporting radio communication and has been leveraged and industrialized into a home security application for industrial customer. The second SoC is a vision sensor integrated with the processor and memory on the same chip. The third SoC has been designed by a Swiss company named Abilis, using a CSEM DSP core. The fourth SoC is a radio communication circuit using a powerful CSEM processor core.

A. Wisenet SoC

Figure 5. Wisenet SoC
The Wisenet SoC contains an ultra-low-power dual-band radio transceiver (for the 434 MHz and 868 MHz ISM bands), a sensor interface with a signal conditioner and two analog-to-digital converters, a digital control unit based on a CoolRISC microcontroller with SRAM low-leakage memories and a power management block. In terms of power consumption, the most critical block is the RF transceiver. In a 0.18-micrometer standard digital CMOS process, in receive mode, the radio consumes 2.3 mA at 1.0 Volt and 27 mA in transmit mode for 10dBm emitted power. However, as the duty cycle of any WSN application is very low, using the WiseNET transceiver with the WiseMAC protocol [18], a relay sensor node consumes about 25 microwatts when forwarding 56-byte packets every 100 seconds, enabling several years of autonomy from a single 1.5V AA alkaline cell. Figure 5 shows the integrated WiseNET SoC.

B. Vision Sensor SoC

Icycam is a circuit combining on the same chip a CSEM 32-bit icyflex 1 processor [19] operated at 50 MHz, and a high dynamic range versatile pixel array integrated on a 0.18 µm optical process.

Figure 6. icycam SoC

Icycam has been developed to address vision tasks in fields such as surveillance, automotive, optical character recognition and industrial control. It can be programmed in assembler or C code to implement vision algorithms and controlling tasks. The icyflex 1 processor communicates with the pixel array, the on-chip SRAM and peripherals via a 64-bit internal data bus. The pixel array has a resolution of 320 by 240 pixels (QVGA), with a pixel pitch of 14 µm. Its digital-domain pixel-level logarithmic compression makes it a low noise logarithmic sensor with close to 7 decades of intra-scene dynamic range encoded on a 10-bit data word. One can extract on the fly the local contrast magnitude (relative change of illumination between neighbour pixels) and direction when data are transferred from the pixel array to the memory. Thus it offers a data representation facilitating image analysis, without overhead in term of processing time. Data transfer between the pixel array and memory or peripherals is performed by group of 4 (10 bits per pixel) or 8 (8 bits per pixel) pixels in parallel at system clock rate. These image data can be processed with the icyflex’s Data Processing Unit (DPU) which has been complemented with a Graphical Processing Unit (GPU) tailored for vision algorithms, able to perform simple arithmetical operations on 8- or 16-bit data grouped in a 64-bit word. Internal SRAM being size consuming, the internal data and program memory space is limited to 128 KBytes. This memory range can be extended with an external SDRAM up to 32 MBytes. The chip has been integrated and is pictured in Figure 6.

C. Mobile TV SoC

CSEM has licensed a DSP core (called MACGIC [20]) to Abilis [21], a Swiss company of the Kudelski group. This DSP core has been used in a SoC for broadband communication in a wireless multipath environment using Orthogonal Frequency Division Multiplexing (OFDM).

Figure 7. Abilis SoC

The SoC developed by Abilis (Fig. 7) is an OFDM digital TV receiver for the European DVB-T/H standards containing a multi-band analog RF tuner, immediately followed by an analog-to-digital-converter (ADC) and a digital front-end implementing time-domain filtering and I/Q channels mismatch correction. Several algorithms are executed on chip, such as mismatch correction, Fast Fourier Transform (FFT), equalizer, symbol de-mapping and de-interleaving, forward error correction (FEC) through Viterbi decoder, de-interleaver
and Reed-Solomon decoder. The main algorithms implemented by the software programmable OFDM demodulator are the frequency compensation, the FFT and an adaptive channel estimation/equalization. Abilis has designed a 90nm single-die digital mobile TV receiver platform (Fig. 7), from which two different chips, the AS-101 and AS-102 have been developed (for DVB-T/H applications). The programmable OFDM demodulator is implemented as a set of 3 CSEM’s MACGIC DSPs customized for OFDM applications (Fig. 8). The SoC contains also an ARC 32-bit RISC core as well as four hardware accelerators (RS decoder, Viterbi decoder, de-interleaver, PID filter).

IV. DISRUPTIVE ARCHITECTURES AND SYSTEMS?

By looking at various Roadmaps, the end of CMOS «scaling» is predicted around 11 nanometers, around 2013 to 2017. So we could conclude of this that after 2017, we should move to «Beyond CMOS ». However, today, there is no clear alternating route to replace CMOS. If one is looking at CNT, nanowires, molecular switches etc…, one can conclude that it is not so clear how to use these devices for architectures and systems requiring billions of switches and how to interconnect them with billions of wires. Nevertheless, there is an interesting approach in hybrids CMOS and nano-devices, it will be heterogeneous… With these nano-elements, one has sometimes the same problems at low level (leakage, process variations), but we could also imagine or hope that some of these effects would disappear!

It is sometimes interesting to revise completely the classical ways of thinking and to try to elaborate disruptive heterogeneous SoC architectures. A first idea could be to design a single universal SoC platform: the motivation is that all applications have to rely on the same hardware, and consequently, the design and differentiator between various applications is fully concentrated on embedded software. Such a SoC platform would be very expensive to develop, about 100 M€, and one could ask whether it remains reasonable for applications sensitive to power consumption or to other specific performances.

A second idea is a SoC dominated by memories. Memories are automatically generated, implying that the hardware part to design is very small and yields low development. It means that one has to maximize the on-chip memory part, with very small processors and peripherals. In this case, the design of a new chip mainly consists in the development of embedded software. It is therefore similar to the first idea, the difference being that a new chip is designed with the required amount of memory, but not more. A third idea is a SoC with 1’000 parallel processors. It is very different from multicore chips with 2 to 32 cores. With 1’000 cores, each core is a very small logic block of 50K gates combined with a lot of memory. A fourth idea is the design of SoC architectures with nano-elements. The design methodology will be completely different, consisting in a bottom-up design methodology and not in a top-down one. It is due to the fact that the fabrication process will produce many nano-devices with few of them being functional. So the design methodology will consist of checking if the fabricated chip can be used for something useful. However, the applications which will be completely different than existing microprocessors; one can think more about neural nets, biological circuits or learning circuits.
V. CONCLUSION

The diagnostic is clear: complexity increases, interdisciplinary too. There are increasingly more interactions between all design levels from application software down to RF-based MPSoC and even MEMS and SiP. Consequently, engineers have to design towards higher and higher design levels but also down to lower and lower design levels. This widening gap will call for design teams that are more and more heterogeneous, with increasingly challenging objectives: to perform focused research for providing outstanding and innovative blocks in a SoC, but also interdisciplinary research which becomes the “key” to successful SoC designs.

VI. ACKNOWLEDGEMENTS

The author wishes to acknowledge the CSEM design teams that contributed to the SoC cases described above: C. Arm, M. Morgan, D. Séverac, S. Gyger, J.-L. Nagel, F. Rampogna, S. Todeschini, R. Caseiro of the “SoC and Digital Group”, E. Franz, P.-F. Ruedi, F. Kaess, E. Gernet, P. Hein, P-A. Beuchat, of the “Vision Sensor Group”; V. Peiris, D. Ruffieux, F. Pengg, M. Kucera, A. Vouilloz, J. Chabloz, M. Contaldo, F. Giroud, N. Raemy of the “RF and Analog IC Group”; E. Le Roux, P. Volet of the “Digital Radio Group”. The author also wishes to acknowledge the EU project UltraSponder partners (FP7 ICT 20007-2, project 224009), for using the icycom SoC.

The author also acknowledges the industrial contributions from Hager and Semtech for the WiseNET SoC, and Abilis for the MACGIC-based SoC for mobile TV.

VII. REFERENCES