




Public

 Information Society Technologies	FP6-2004-IST-4-026980-IP	
<b><u>C</u>ontrolling <u>L</u>eakage in <u>N</u>anoCMOS SoCs</b>		

	WP no.	Result no.	Lead participant
	<b>WP5</b>	<b>D5.1.1.3</b>	<b>Polito</b>
<b>Third release of state-of-the art and market survey document</b>			
Project coordinator name:	<b>Roberto Zafalon</b>		
Project coordinator organisation:	<b>STM-Italy</b>		
Document number:	<b>CLEAN/POLITO/D5.1.1.4/V.4.0</b>		
Classification:	<b>CLEAN Confidential</b>		
Preparation date:	<b>May 30, 2008</b>		
Covered period:	<b>May 1, 2007 – April 30, 2008</b>		
<b>Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)</b>			

© Copyright 2005-2006-2007-2008 STMicroelectronics, Infineon Technologies, OFFIS e.V., Politecnico di Torino, Universitat Politecnica de Catalunya, CEA-LETI, Politeknika Warszawska, ChipVision Design Systems, BullDAST s.r.l., edacentrum GmbH, Technical University of Denmark, Consorzio per la Ricerca e l'Educazione Permanente, Budapest University of Technology and Economics.

This document and the information contained herein may not be copied, used or disclosed in whole or in part outside of the consortium except with prior written permission of the partners listed above.

## History of Changes

<b>Rev.</b>	<b>DATE</b>	<b>PAGES</b>	<b>REASON FOR CHANGES</b>
1.0	May 10, 2008	3	Skeleton of document provided to partners
2.0	May 23, 2008	20	State of the art section completed
3.0	May 28, 2008	24	Market survey section completed
4.0	May 30,2008	24	Final revision

*This page was intentionally left blank.*

## Contents

1	Introduction.....	1
2	State-of-the-art in low-leakage design techniques .....	2
2.1	Leakage modeling .....	2
2.1.1	New models for leakage .....	2
2.1.2	Models for CAD tools.....	3
2.1.3	References .....	5
2.2	Transistor, gate and RT level leakage optimization techniques.....	6
2.2.1	References .....	8
2.3	Memory design techniques.....	9
2.3.1	References .....	11
2.4	Behavioral and system-level techniques.....	12
2.4.1	The project approach .....	12
2.4.2	Fluctuation aware synthesis .....	12
2.4.3	Other synthesis for leakage approaches.....	13
2.4.3.1	Behavioral transformation of system description.....	13
2.4.3.2	Improvements in scheduling, binding and allocation process.....	14
2.4.3.3	Temperature Adaptation .....	15
2.4.4	References .....	16
3	Market survey .....	17
3.1	Transistor, gate and RT-level tools.....	17
3.1.1	TMSC's Power Trim service .....	17
3.1.2	Envis' Chill.....	17
3.1.3	Genesys' Design for Leakage Test (DFLT).....	18
3.2	Behavioral and system-level tools.....	18
3.2.1	Cadence .....	18
3.2.2	ChipVision.....	18
3.2.3	Synopsys.....	18
3.2.4	Power Standards .....	18
4	Conclusions.....	20

## **1 Introduction**

This document contains a detailed analysis of the most recent low-leakage design techniques and methodologies, as well as a comprehensive overview of the EDA tools, targeting the leakage power minimization as one of the most relevant optimization parameters, currently available on the market. This document is an update of Deliverable D5.1.1.2.

## 2 State-of-the-art in low-leakage design techniques

### 2.1 Leakage modeling

Recent results and advances on the state of the art on leakage modeling appeared during the last year, with potential impact on future research on low leakage power control have been collected in this subsection.

A summary of the information about new proposals appeared recently in modeling of leakage currents and process variability effects. A summary on the updates for the CAD models Berkeley BSIM4 and Philips PSP is presented.

#### 2.1.1 New models for leakage

The latest advancements on leakage modeling are related to the need of incorporating new leakage mechanisms having higher impact than in the past. In this category we find tunneling for highly doped junctions named Band to Band Tunneling (BTBT) as well as Gate tunneling currents for the new dielectric materials.

New device geometries require specific adjustments of the model parameters to incorporate the impact of the physical phenomena on the novel structures. An example of this trend can be found on [7] where the leakage currents for Double Gate (DG) transistors is modeled taking into consideration different tunneling currents on the special transistor structure.

The traditional charge transport methods are not accurate for nano-transistors and new models for the quantum mechanics transport are currently investigated. In [1] the need for new modeling strategies is presented and the main causes of error of present models are discussed.

The six main causes identified by the authors are:

- 1) **Ballistic transport:** The channel length of the silicon transistors is projected to approach 10 nm in the year 2015 for the 22-nm technology node. This is comparable to the wavelength of an electron with thermal velocity and mean free path (due to electron-phonon scattering), which are both also around 10 nm. The concept of mobility within the context of the drift-diffusion equation is no longer well defined, and transport in the channel is quasi-ballistic.
- 2) **Scattering from impurities:** With decreasing channel lengths, the device to device fluctuations in transistor current-voltage characteristics due to scattering from impurities become important. Scattering from multiple impurities will not be equivalent to the series combination of scattering from independent impurities, as assumed in semi-classical modeling.
- 3) **Tunneling and capacitance corrections:** The thickness of ultra-thin silicon oxides will approach 1 nm unless nehigh- $\kappa$  dielectrics are used. The tunneling current through oxides is expected to increase with a decreasing channel length. In addition, the location of the inversion layer in the silicon channel typically peaks around 1 nm from the oxide-silicon interface. This decreases the capacitance of the MOS structure due to a non negligible increase in effective oxide thickness.
- 4) **Novel channel materials:** In emerging prototype transistors based on nano-materials such as carbon nanotubes and nanowires, a straightforward application of the traditional effective mass approach is of limited validity.

- 5) **Novel device physics:** Nanotransistors may exploit band to band tunneling (between conduction and valence bands) in the channel. These tunnel transistors can have an inverse subthreshold slope that is smaller than 60 mV per decade at room temperature, which is the minimum possible value in conventional transistors.
- 6) **Quantum corrections to semiclassical simulations:** Semi-classical models have incorporated quantum corrections to account for quantum mechanical features that become important with miniaturization. Quantum mechanical simulations are playing an important role in benchmarking these quantum corrections.

The authors discuss an approach to quantum mechanical modeling of transistors using the non-equilibrium Green's function (NEGF). The approach permits to calculate the non-equilibrium distribution function of electrons in the presence of tunneling, scattering mechanisms that cause energy, momentum, and phase relaxation.

Process variability is increasing significantly at each new technology node. Modeling the variability of process, supply voltage and on chip temperature using the classic corner approach provides very pessimistic designs with excessive guard bands. A new design paradigm based on the statistical description of the device modeled and new performance metrics given as functions of random variables has emerged in recent years. An example of statistical design can be found in [2].

In the domain of statistical leakage aware design [4] studies a width-dependent statistical leakage model with an estimation error less than 5% Design examples on SRAMs and domino circuits are presented.

An interesting approach to statistical characterization of standard cell based designs is presented in [VIR 08]. Standard cell libraries for statistical leakage analysis based on models for transistor stacks is the basic idea of this work. Modeling stacks has the advantage of using a single model for many gates there by reducing the number of circuits that need to be characterized.

### 2.1.2 Models for CAD tools

Next a summary of the available models in BSIM4 an Philips public models is summarized

#### BSIM4 model

Two new versions have been released for the BSIM4 models. The BSIM 4.6.0 released on December, 13<sup>th</sup> 2006 and the revision 4.6.1 released on May, 18<sup>th</sup> 2007. In the following, the main updates with respect to the 4.5.0 from July 29<sup>th</sup>, 2005 affecting to the leakage currents are summarized.

Compared with BSIM4.6.0, several new features are added in the BSIM4.6.1 version [3].

1) **New material model** is introduced for the predictive modeling of Non-SiO<sub>2</sub> insulator, Non Poly Silicon gate and Non-silicon channel.

- The following new parameters are added

MTRLMOD : New material model selector

PHIG, EPSRGATE : non-poly silicon gate parameters

EOT, VDDEOT : non-SiO<sub>2</sub> gate dielectric

EASUB, EPSRSUB, NI0SUB, BG0SUB, TBGASUB, TBGBSUB, ADOS, BDOS : Non-silicon channel parameters

2) **Mobility model** (MOBMOD = 0 and MOBMOD = 1) has been improved through predictive modeling of vertical electric field. The improved mobility model is selected through MTRLMOD = 1 for backward compatibility.

3) **GIDL/GISL models** are improved through an improved definition of flatband voltages at S/D ends. The improved GISL/GIDL model is selected through MTRLMOD = 1 for backward compatibility.

4) **Poly-depletion model** is modified to account for new gate and gate-insulator materials.

5) **C-V model** has been improved by adding a new VgsteffCV definition through CVCHARGEMOD = 1

For the 4.6.0 release the improvements consist in the separation of the parameters for different type of components. GIDL and GISL have now independent model parameters. Similarly, the source and drain side junction diodes have totally separated parameters. Finally, the parameters have also been separated for the gate tunneling current in the source and drain overlap diffusion regions.

#### **PSP model [5]**

The PSP model is a compact MOSFET model intended for digital, analogue, and RFdesign, which is jointly developed by NXP Semiconductors (formerly part of Philips) and Arizona State University (formerly at The Pennsylvania State University). The roots of PSP lie in both MOS Model 11 (developed by Philips Research) and SP (developed by Penn State University). PSP is a surface-potential based MOS Model, containing all relevant physical effects (mobility reduction, velocity saturation, DIBL, gate current, lateral doping gradient effects, STI stress, etc.) to model present-day and upcoming deep submicron bulk CMOS technologies. A source/drain junction model, c.q. the JUNCAP2 model, is an integrated part of PSP. In December 2005, the Compact Model Council (CMC) has elected PSP as the new industrial standard model for compact MOSFET modeling.

In version 101.0 the drain induced barrier lowering model has been modified and its behavior at a larger than 1V body potential has been improved. For the rest of versions there are no modifications in relation to the leakage currents. BTBT currents are modeled using 8 parameters and Gate Leakage currents require 7 parameters to be specified by the user. For further details see [6]

### 2.1.3 References

- [1] Anantram, M.P.; Svizhenko, A., Multidimensional Modeling of Nanotransistors, *IEEE Trans on Electron Devices*, Volume: 54 Issue: 9 Sept. 2007 Page(s): 2100-2115
- [2] Bhardwaj, S.; Vrudhula, S., Leakage Minimization of Digital Circuits Using Gate Sizing in the Presence of Process Variations, *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* Volume: 27 Issue: 3 March 2008 Page(s): 445-455
- [3] [http://www-device.eecs.berkeley.edu/~bsim3/bsim4\\_get.html](http://www-device.eecs.berkeley.edu/~bsim3/bsim4_get.html)
- [4] Jie Gu; Sapatnekar, S.S.; Kim, C. Width-dependent Statistical Leakage Modeling for Random Dopant Induced Threshold Voltage Shift, *Design Automation Conference, 2007. DAC '07. 44th ACM/IEEE*Page(s): 87-92
- [5] [http://www.nxp.com/Philips\\_Models/mos\\_models/psp/](http://www.nxp.com/Philips_Models/mos_models/psp/)
- [6] [http://www.nxp.com/acrobat\\_download/other/models/PSP102\\_Summary.pdf](http://www.nxp.com/acrobat_download/other/models/PSP102_Summary.pdf)
- [7] Sarkar, D.; Ganguly, S.; Datta, D.; Sarab, A.A.P.; Dasgupta, S, Modeling of Leakages in Nano-Scale DG MOSFET to Implement Low Power SRAM: A Device/Circuit Co-Design. *VLSI Design, 2007. 6th International Conference on Embedded Systems., 20th International Conference on*6-10 Jan. 2007 Page(s): 183-188
- [8] Viraraghavan, Janakiraman; Das, Bishnu Prasad; Amrutur, Bharadwaj, Voltage and Temperature Scalable Standard Cell Leakage Models Based on Stacks for Statistical Leakage Characterization, *VLSI Design, 2008. VLSID 2008. 21st International Conference on*, 2008 Page(s): 667-672

## 2.2 Transistor, gate and RT level leakage optimization techniques

Target of this Section is to provide an analysis of the most recent “low-level” leakage optimization techniques, appeared in the literature right after the submission of Deliverable D5.1.1.2.

As mentioned in Deliverables D5.1.1.1 and D5.1.1.2, leakage optimization circuit techniques can be categorized into the following classes:

1. Transistor stacking;
2. Multi- $V_{th}$ ;
3. Dynamic  $V_{th}$ ;
4. Supply voltage scaling;
5. Input Vector Control (IVC);
6. Body biasing.

In the time period covered by this document, most of the new contributions deal mainly with Multi- $V_{th}$  and IVC techniques.

MTCMOS technology provides low leakage and high performance operation by utilizing high speed, low  $V_{th}$  transistors for logic cells and low leakage, high  $V_{th}$  devices as sleep transistors. Sleep transistors disconnect logic cells from the power supply and/or ground to reduce the leakage in the sleep mode. There is a performance degradation associated with the sleep transistor insertion. This is due to the IR-drop across the MTCMOS cells in the active mode of operation. For a fixed placement, the amount of the performance degradation depends on the size of the MTCMOS switch cells. The larger the sleep transistors are, the lower the performance degradation is. However, the amount of the power consumption will increase with the size of the sleep transistors. Therefore, there is a trade-off between the amount of the performance degradation and the power consumption of the sleep transistors in an MTCMOS circuit. This makes MTCMOS cell sizing one of the most important issues in the coarse-grain MTCMOS design flows. In some applications performance is too critical and the designer cannot afford any performance degradation due to MTCMOS.

In [1] authors propose to separate timing critical standard cells from the non-critical ones by placing them in different rows and by doing the power gating only for the non-critical standard cell rows. They have shown that a high leakage saving can be achieved while losing a small amount of performance.

In [2], the authors assume that MTCMOS is applied to all standard cell rows. Furthermore, no rail sharing is assumed for the neighbor rows. In this paper the authors present a delay budgeting algorithm to size the sleep transistors in a circuit. The placement of the logic cells and sleep transistor cells are known and given. A delay-budgeting algorithm to optimally use the total available slack and size the sleep transistors optimally is presented.

In [3] the authors propose synchronized dual- $V_{th}$  self-sleep buffer method that eliminates the need for sleep signal distribution and allows easy implementation of MTCMOS wakeup scheduling. Guidelines for designing and sizing the self-sleep buffer circuit are provided. In a 90-nm technology and 2-GHz clock frequency, the self-sleep buffer consumes only 1.46-  $\mu$ W in active mode, while eliminating the sleep distribution network overheads and providing fast, low-energy active- to-standby-to-active transitions.

In [4] a circuit technique is proposed for simultaneously reducing the subthreshold and gate oxide leakage power consumption in domino logic circuits. Only p-channel sleep transistors and a dual-threshold voltage CMOS technology are utilized to place an idle domino logic circuit into a low leakage state. Sleep transistors are added to the dynamic nodes in order to reduce the subthreshold leakage current by strongly turning off all of the high-threshold voltage transistors. Similarly, the sleep switches added to the output nodes suppress the voltages across the gate insulating layers of the transistors in the fan-out gates, thereby minimizing the gate tunneling current. The energy overhead of the circuit technique is low, justifying the activation of the proposed sleep scheme by providing a net savings in total energy consumption during short idle periods.

In [5] an algorithm to determine how to cluster cells to share sleep transistors, while taking both topology and functionality into consideration is proposed. Moreover, one placement refinement algorithm that takes clustering information into account is presented. At the logic level, the results show that the proposed clustering method can achieve an average of 22% reduction in terms of the number of unit-size sleep transistors as compared to a method that does not consider functionality. At the physical level, two placement results are discussed. The first is produced by a traditional placement tool plus topology check (functionality check) for insertion of sleep transistors. It shows that the functionality check algorithm produces 9% less chip area as compared with the topology check algorithm. The second result is produced by a placement refinement algorithm where the initial placement is done in the first placement experiment. It shows that the placement refinement algorithm achieves 5% more reduction in area at the expense of 4% increase in wire length. Totally, around 14% reduction is achieved by utilizing the clustering information.

In [6] two novel approaches to leakage power minimization in static complementary metal oxide–semiconductor circuits that employ input vector control (IVC) are investigated. The authors model leakage effects by means of pseudo-Boolean functions. These functions are linearized and incorporated into an exact (optimal) integer linear programming (ILP) model, called virtual-gate ILP, which analyzes leakage variation with respect to a circuit's input vectors. A heuristic mixed-integer linear programming (MLP) method is also proposed, which has several advantages: It is faster, its accuracy can be quickly estimated, and tradeoffs between runtime and optimality can easily be made. Furthermore, the MLP model also provides a way to estimate a lower bound on circuit leakage current. The proposed methods are used to generate an extensive set of experimental results on leakage reduction. It is shown that average leakage currents are usually 1.25 times the minimum, confirming the effectiveness of IVC. The heuristic MLP approach is shown to be approximately 13.6 times faster than the exact ILP method, whereas finding input vectors whose power consumption is only a few percent above the optimum. In addition, the lower bound estimated by the MLP model is also within a few percent of the optimal value.

### 2.2.1 References

- [1] A.Sathanur, A.Pullini, L.Benini, A.Macii, E.Macii, M.Poncino, “Timing-driven row-based power gating,” Proc. Int’l Symp. on Low Power Electronics and Design, pp. 104-109, 2007.
- [2] E. Pakbaznia and M. Pedram, “Coarse-Grain MTCMOS Sleep Transistor Sizing Using Delay Budgeting”, DATE '08: Design, Automation and Test in Europe, pp. 385 – 390, 2008.
- [3] C. J. Akl, M. A. Bayoumi, “Self-Sleep Buffer for Distributed MTCMOS Design”, Proc. of the 21st International Conference on VLSI Design, pp. 673-678, 2008.
- [4] Z. Liu, and V. Kursun,” PMOS-Only Sleep Switch Dual-Threshold Voltage Domino Logic in Sub-65-nm CMOS Technologies”, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 15, n. 12, 2007.
- [5] A-C. Hsieh, T-T. Lin, T-W. Chang, T. Hwang, “A functionality-directed clustering technique for low-power MTCMOS design—computation of simultaneously discharging current”, ACM Transactions on Design Automation of Electronic Systems, Vol. 12 , n. 3, 2008.
- [6] F. Gao and J. P. Hayes, “Exact and Heuristic Approaches to Input Vector Control for Leakage Power Reduction”, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 25, n. 11, 2006.

## 2.3 Memory design techniques

As already experienced in the previous surveys of the state-of-the-art in low-leakage memories, many new contributions deal mostly with circuit-level techniques or even involving the design of the very memory cell. References [1] and [2] are the two most relevant examples. The former paper uses non bulk-CMOS technologies (namely SOI) for designing memory cells with low leakage, whereas in the latter paper leakage is achieved by isolating the data from the bit lines during a read operation using a 9-transistor memory cell.

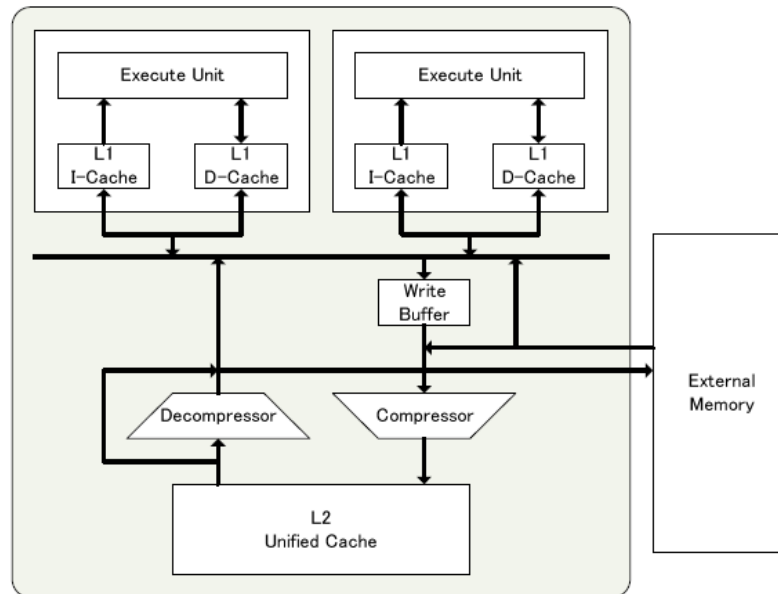
A few solutions at higher levels of abstractions have however been proposed in the last year , mostly related to caches ([3],[6],[7],[7]).

The solution proposed by [3] is based on the observation that most data values (mainly integer values) stored in the cache are composed of a majority of bits set to zero, and proposes propose the *zeros switch-off* (ZSO) technique, that switches off the power supply to some chunks of SRAM cells, using the gated-VDD technique (i.e., similar to a sleep transistor like the ones used power gating). The implementation requires to add extra bits to every word stored in the cache (1 bit per byte, indicating zero or non-zero value, controls the power supply of that byte). This idea is not new by itself and was already exploited in already published papers (e.g, the FV-cache or the STV Cache [4],[5]) and in addition it requires like similar approaches modification to the internal cache structure.

Tanaka proposes in [6] the combination of software techniques with hardware compression in the context of chip multiprocessors, They use gated-Vdd as a primary leakage control mechanism, which is controlled by the L1 cache memory by means of the execution of some special load and store instructions (called by the author “software self-invalidation”).

On the other hand, a data compression technique is applied to the L2 cache and its vacant portions are turned off by the gated-Vdd, which does not cause additional cache misses. The reference architecture (Figure 1) assumes coherent L1 caches and compression/decompression units between L1 and L2 caches. As in all gated-Vdd based schemes, this approach requires modification to the internal cache structure to accommodate for gating transistors.

The approach described in [7] targets cache-coherent chip multiprocessors and plays with principle of *inclusion* on which caches rely, that is, that property in a cache hierarchy which requires that if a cache line is present in a lower level cache (e.g. L1), it should also be present in all the higher levels (e.g. L2 and beyond). Inclusion is normally the default in multi-processors because it facilitates efficient implementation of cache coherence. The authors show that multi-level inclusion makes approaches such as cache decay unfeasible, and propose *virtual exclusion* as a way to enable existing leakage-aware memory management solutions. Virtual exclusion saves leakage energy by keeping the data portion of repetitive cache lines off in the large higher level caches while still manages to maintain multi-level inclusion. The method has minimal overhead because it can exploit the existing state information in conventional snoop-based cache coherence protocols.



**Figure 1. Architecture proposed by [6].**

The last approach we consider here is the one by Goudarzi and Ishihara [7], who, by exploiting again the dominance of 0's in caches (instruction caches in this specific case), propose to enforce the use of asymmetric SRAM cells that are properly designed for dissipating less leakage when storing a "0". This idea is not new by itself, but it is managed at the software level by carefully choosing register operands of instructions in such a way that the number of 0 bits is maximized. For example, if we need to store 4 variables simultaneously, a conventional approach would use registers in numerical order (e.g., R0, R1, R2, and R3); in the proposed approach, it is more beneficial to use R4 instead of R3 since binary representation of 4 has more 0's than that of 3. They apply this static renaming of register directly on the binary code (and not on source code), taking into account control and data-dependencies among instructions so that the producer-consumer relation among instructions are not affected.

These solutions compare in different ways with respect to the solutions developed in the CLEAN project. The first two solutions resort to an underlying low-level, internal, low-leakage scheme (e.g., drowsy cache), which requires modification of the internal memory implementation, which we cannot support in a standard design flows. The third solution is actually the most innovative but targets a very specific class of architectures, high-end chip multi-processors and does not apply to a system with a single core and a single level of memory hierarchy. The last solution relies on a non-standard asymmetric memory cell structure, which however might be available by some silicon foundries (it does not imply particular overhead); the technique is orthogonal to any other architectural-level solution since it is simply based on modifying register encoding.

### 2.3.1 References

- [1] D. Levacq, V. Dessard, D. Flandre, “Low Leakage SOI CMOS Static Memory Cell With Ultra-Low Power Diode”, *IEEE Journal of Solid-State Circuits*, Vol. 42, No. 3, March 2007, pp. 689 – 702.
- [2] L. Zhiyu V. Kursun, “High Read Stability and Low Leakage Cache Memory Cell” *ISCAS 2007. IEEE International Symposium on Circuits and Systems*, May 2007, pp. 2774 – 2777.
- [3] R. Ubal, J. Sahuquillo, S. Petit, H. Hassan, P. Lopez, “Leakage Current Reduction in Data Caches on Embedded Systems”, *IPC’07, 2007 International Conference on Intelligent Pervasive Computing*, Oct. 2007 pp. 45 – 50.
- [4] C. Zhang, J. Yang, F. Vahid, “Low Static-Power Frequent-Value Data Caches,” *DATE’04: Design, Automation, and Test in Europe*, February 2004, pp. 210-215.
- [5] K. Patel, L. Benini, E. Macii, M. Poncino, “STV-Cache: a leakage energy-efficient architecture for data caches”, *GLSVLSI ’08: 16th ACM Great Lakes Symposium on VLSI*, April 2006, pp. 404-409.
- [6] K. Tanaka, “Cache Memory Architecture for Leakage Energy Reduction”, *IWIA 2007. International Workshop on Innovative Architecture for Future Generation Processors and Systems*, Jan. 2007 pp. 73 – 80.
- [7] M. Ghosh, H.H. Lee, “Virtual Exclusion: An architectural approach to reducing leakage energy in caches for multiprocessor systems”, *ICPDS’07, International Conference on Parallel and Distributed Systems*, Volume 1, Dec. 2007 pp. 1 – 8.
- [8] M. Goudarzi, T. Ishihara, “Instruction cache leakage reduction by changing register operands and using asymmetric SRAM cells”, *GLSVLSI ’08: 18th ACM Great Lakes Symposium on VLSI*, May 2008, pp. 383-386.

## 2.4 Behavioral and system-level techniques

Since the last report D5.1.1.2, two major estimation-optimization frameworks and some additional ideas have been proposed optimizing the synthesis steps binding, allocation and scheduling to reduce the leakage itself and to improve the system's variation robustness. As this is a public report, also the framework, developed in the CLEAN project and the leakage aware synthesis ideas reported in D2.4.1.1 are outlined.

### 2.4.1 The project approach

One of the major approaches is the one, basing on the CLEAN project results: In a first step, the target technology is analyzed. In order to describe the leakage of the system, the BSIM specification of the technology is abstracted into a transistor model describing the technologies behavior on a change of one of the major parameters as temperature, supply voltage, body voltage, channel length, oxide thickness, or channel doping. By synthesizing several RT components, a bottom-up model, accurately describing the leakage of entire RT components is characterized.

The delay modeling of this framework starts with an analysis of the SPICE description of the standard cell gates. Using an accurate inverter-delay model, which is sensitive on temperature, supply- and body-voltage, the process parameters mentioned above, the fanout load and the input slope, the rise and fall delay of each input pin of each gate can be modeled. These delay models can then be integrated for each path through the component finding the (potentially temperature dependent) critical path, still under consideration of the PTV (process, temperature, and voltage) variation [1].

Both models for delay and leakage are combined by a variation engine, which is characterized by an analysis of the description of the variation (inter- and intra die, with consideration of complex distributions and correlated input parameters) [2].

If included into the synthesis tool PowerOpt (ORINOCO), the models report the effect of each design modification on the distribution of leakage and delay [3]. By a modification of the scheduling and binding heuristic, PowerOpt is able to generate a system optimally exploiting power gating without performance degradation. Additionally, the floorplanning and binding steps were combined to generate DVS and ABB islands under consideration of the overhead [4].

### 2.4.2 Fluctuation aware synthesis

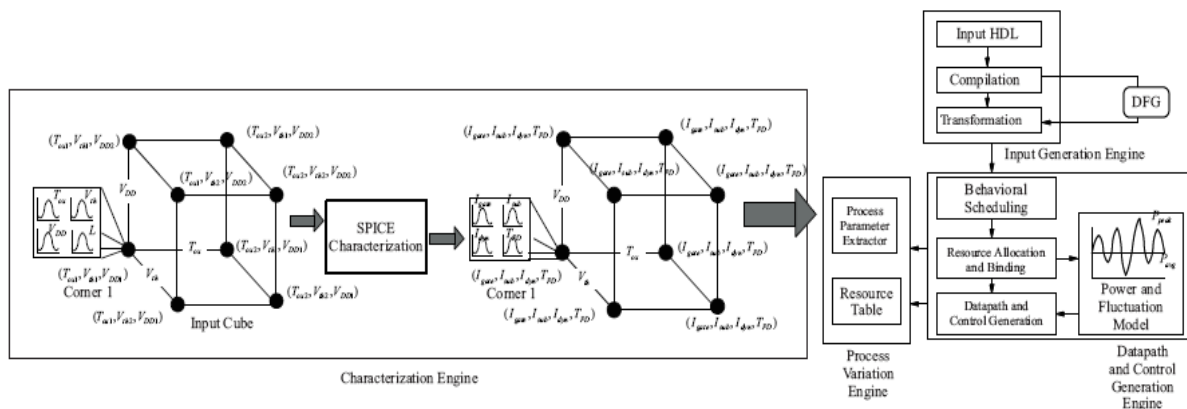
A comparable, yet simpler framework basing on modeling of the variation, followed by a synthesis optimization is presented by [5]. This approach is divided into 4 different engines: the characterization engine analyzing SPICE files, the variation engine, describing the process variation and its correlation, the input generation engine reading in the system description, and the datapath and control generation engine, optimizing the system's resulting netlist.

In the characterization engine, the leakage current and delay distribution is performed by an 8 design corners based model (see

Figure 2). Assuming Gaussian distribution of 4 parameters ( $V_{DD}$ ,  $L_{ch}$ ,  $T_{ox}$ ,  $V_{th}$ ) and (implicitly) the linearity of the leakage, dynamic power, and delay functions, the Gaussian distribution of the gate and subthreshold leakage, the dynamic power and delay are

determined<sup>1</sup>. The process variation engine now supplies the tool with statistical data for each parameter using a resource table which is filled by the characterization engine.

The system description is translated into a graph (data flow graph) by the input generation engine. In the data path and control generation engine, a target function of the leakage and dynamic power, the delay and the variation of these parameters can be specified. By a simulated annealing based optimization, the system synthesis (without the scheduling, which is fixed at optimization time), this target function can be minimized.



**Figure 2: Left: Characterization engine of the fluctuation aware synthesis approach. Assuming a Gaussian variation of the supply voltage, the oxide thickness, the channel length and the resulting threshold voltage, for 8 design corners the gate and subthreshold leakage as well as the dynamic (charge) currents and the transition delay are stored. Right: Using the variation aware model, each component's dynamic and static power as well as its delay can be predicted. The heuristic optimization basing on these predictions can now reduce power, delay, and/or variation.**

## 2.4.3 Other synthesis for leakage approaches

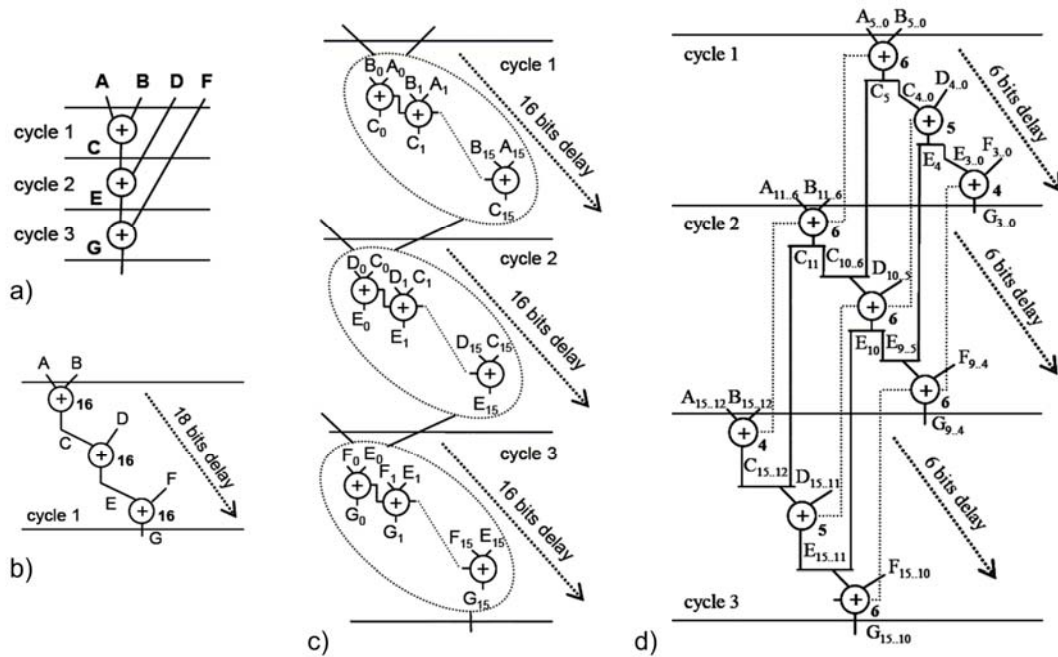
### 2.4.3.1 Behavioral transformation of system description

[6] presents an approach that performs pre synthesis optimization by transforming the behavioral level system description with the objective of increasing the optimization potential within the synthesis process.

The authors address the slack that is produced by conventional scheduling algorithms adjusting the minimum clock cycle duration to the execution time of the slowest operation in the system. Therefore, appropriate operations in the behavioral level system description are split into smaller ones resulting in the execution of the whole operation distributed over several not necessary consecutive clock cycles. These kinds of transformations result in more space for following scheduling, allocation and binding optimization algorithms.

An example of such a transformation is presented in Figure 3.

<sup>1</sup> It is neither fair to assume Gaussian distribution of the input parameters, nor is the number of variables sufficient (at least the temperature should be added). The leakage currents are also not Gaussian distributed. It is not analyzed by the authors, how high the error of these simplifications is.



**Figure 3: a) Conventional schedule b) Schedule using operation chaining c) Zoom into conventional schedule d) Zoom into the schedule of the transformed specification**

Having three chained 16 bit additions a conventional schedule would look as presented in Figure 3 a). From Figure 3 c), one can see that in case of a ripple adder, the clock cycle delay is determined by the sum of the delays of 16 single bit adders. The whole chain requires 3 clock cycles to perform all operations which results at least in 48 single bit delays. In case of operation chaining, as presented in Figure 3 b) only one clock cycle is needed. This time, the whole execution time is reduced to 18 single bit adders, due to the rippling effect that allows the execution of some bits of the three operations in parallel. The drawback of this approach is that the duration of the clock cycle is increased compared to the first approach. The optimized method presented in Figure 3 d) combines the advances of both approaches. All three operations are performed within three clock cycles but this time in a way that the duration is limited to 18 single bit adder delays. The clock cycle duration in this case is determined by only 6 single bit adder delays.

#### 2.4.3.2 Improvements in scheduling, binding and allocation process

Several new scheduling, binding and allocation algorithms were developed having the objective of statically reducing the leakage power or providing better support for power management technologies such as power gating (MTCMOS).

In [7] a framework that supports power gating within the behavioral level synthesis process in the binding and allocation step. Therefore, they developed a library that contains power gated RTL components. Using this library a clique partitioning-based approach is presented that selectively replaces components with much idle time with functional equivalent candidates, taken from the power gating library. Additionally, the size of the gating transistor is selected in a way that the timing slack of the component is minimized, resulting in minimal additional leakage caused by the transistor itself. Applying the knapsack algorithm, a leakage optimal binding can be found that meets a given area constraint.

The European CRAFT project MAP<sup>2</sup> focuses on the integration of power management support into the high level synthesis tool ORINOCO<sup>®</sup>, complementing the activities in CLEAN. Some work related to dynamic power management was also done. In this work, mainly the binding algorithm within ORINOCO<sup>®</sup> was improved to support power

management. Therefore two algorithms were developed. The first one advances the one, actually integrated, by extending the cost function also considering leakage power and power managed components. The drawback of this approach is a high amount of computing power, due to the complex power management and especially leakage models. To overcome this drawback, a second approach abstracts from any power value only taking idle times of components into account. This far less computing power intensive algorithm is well suited in large designs to create bindings with largest possible idle times between two consecutive operations at one resource. Applying power management techniques then result in reasonable savings of static power, which currently dominates the overall power consumption especially in low utilized designs.

An approach simultaneously performing optimization for scheduling, allocation and binding is proposed by [8]. This work focuses the gate leakage, which is one major part of static power consumption. Here, multi oxide thickness technique is selected to minimize area and power overhead. A simulated annealing based algorithm is selected to find the optimum within the design space spawned by power, area and performance dimensions. Savings of gate leakage with up to 77% were reported.

### **2.4.3.3 Temperature Adaptation**

The static power consumption caused by subthreshold leakage exponentially depends on the surrounding temperature. Thus, an increase of temperature results in a large increase of leakage. [9] presents an approach limit this effect and to prevent a possible thermal runaway. The basic idea of this work is the reduction of the supply voltage in case of high temperature but differing to other approaches without reducing the clock frequency as well resulting in only low performance loss. To prevent wrong computation caused by timing violations, the authors propose a novel design methodology that combines critical path isolation with clock stretching technology. At the beginning the paths are identified that due to temperature variations can possibly get critical ones. Then, it must be unsecured that these paths are rarely used and can tolerate possible delay failures when in low supply voltage mode. When the temperature reaches a threshold value the design is switches to low voltage taking advance from reducing leakage and dynamic power consumption as well. Depending on the input data, some of the critical paths will fail their timing constrains in this mode. Using delayed shadow latches, these deterministic failures can be detected so that an additional clock cycle introduced in such cases can prevent the design from resulting in computational failures caused by timing violations.

As this approach is planned being applied at gate level, there are no further information about the RTL schedule at this level. Having such information enables this approach being extended in a way that possible slack of components can be used instead of increasing the run time using clock stretching technique.

#### 2.4.4 References

- [1] Marko Hoyer, Domenik Helms, Wolfgang Nebel: Modelling the impact of high level leakage optimization techniques on the delay of RT-components. *PATMOS 2007*
- [2] Domenik Helms, Marko Hoyer, Sven Rosinger, Wolfgang Nebel: RT Level Makro Modelling of Leakage and Delay under Realistic PTV Variation. *IEEE LPonTR*, 2008.
- [3] Domenik Helms, Wolfgang Nebel: Logic design techniques for 65 to 45nm and below for reducing total energy and solving technology variations problems. 14th IEEE International Conference on Electronics, Circuits, and Systems, 2007.
- [4] Domenik Helms, Olaf Meyer, Marko Hoyer, Wolfgang Nebel: Voltage- and ABB\_Island Optimization in High Level Synthesis. Intl Symposium on Low Power Electronic Design, 2007.
- [5] Saraju Mohanty, Elias Kougianos: Simultaneous Power Fluctuation and Average Power Minimization during Nano-CMOS Behavioural Synthesis. 20th IEEE Conf. on VLSI Design 2007.
- [6] R. Ruiz-Sautua, M.C. Molina, J.M. Mendías, R. Hermida: Behavioural Transformation to Improve Circuit Performance in High-Level Synthesis. *DATE 2005*
- [7] C. Gopalakrishnan, S. Kathoori: KnapBind: An Area-Efficient Binding Algorithm Low-leakage Datapaths. *ICCD 2003*
- [8] S.P. Mohanty, R. Velagapudi, E. Kougianos: Physical-Aware Simulated Annealing Optimization of Gate Leakage in Nanoscale Datapath Circuits. *DATE 2006*
- [9] S. Ghosh, S. Bhunia, K. Roy: Low-Overhead Circuit Synthesis for Temperature Adaptation Using Dynamic Voltage Scheduling. *DATE 2007*

### 3 Market survey

This deliverable complements the market surveys reported in D5.1.1.1 and D5.1.1.2, and, has therefore incremental nature. The tools offered by Synopsys, Magma, Cadence, Sequence, Prolific and GoldenGate Technology (surveyed in Deliverable D5.1.1.1) and those offered by Blaze, Apache, and Incentia (surveyed in Deliverable D5.1.1.2) are still supported by these companies, although with different levels of commercial success.

In the timeframe between D5.1.1.2 and this deliverable, the picture is basically unchanged; in Section 3.1 the few new transistor/gate/RT-level solutions are surveyed.

Regarding the high levels of abstraction, described in Section **Fehler! Verweisquelle konnte nicht gefunden werden.**, none of commercially available EDA tools but those provided by ChipVision are currently available for estimating/optimizing leakage power above RTL.

#### 3.1 Transistor, gate and RT-level tools

An accurate analysis of the market offer in the last year has not shown any relevant new tool at the transistor/gate/RT-level. Existing tools already described in previous deliverables have been consolidated and are still supported by their providers. The only leakage-related efforts at the transistor/gate/RT-level are relatively marginal, in the sense that none of them truly targets a direct reduction of leakage power.

##### 3.1.1 TSMC's Power Trim service

Thanks to an exclusive agreement with startup Blaze DFM Inc (which provides the Blaze MO tool surveyed in D.5.1.1.2), Taiwan-based semiconductor manufacturing foundry Taiwan Semiconductor Manufacturing (TSMC) is now offering a Power Trim service that combines Blaze's power optimization technology with special variations of TSMC's advanced manufacturing process.

The Power Trim service is meant to allow reductions in leakage power in addition to what is possible with existing tools and techniques, while also reducing leakage power variability, which is a critical power issue to overcome in next generation designs.

This is believed to be the first offering of its kind because it blends a layer of design technology software with advanced semiconductor processing to tune the manufacturing process to the specific chip design.

Nevertheless, it cannot be regarded as a "tool", in the sense that it basically provides leakage-efficient implementation of library cells. In that sense, it cannot be considered by any means a competitor with the gate-level/RTL tools developed in CLEAN.

##### 3.1.2 Envis' Chill

Chill, a tool provided by the US-based company Envis ([www.envis.com](http://www.envis.com)) actually features a new clock-gating approach that combines an analysis and insertion software tool along with proprietary silicon elements that together shuts down many of the clocked elements.

Chill combines two approaches to clock gating. The first one consists of a search and decision process about which elements to gate and what signals can be used or combined to effect the gating, whereas the second one is sequential, looking for conditions one or two clock cycles before that can be used to gate elements on subsequent clock cycles.

Although mainly a tool for dynamic power reduction, the vendor advertises it as a tool providing dynamic *and* static leakage reduction. In fact, static power reduction is actually achieved as a by-product by means of a reduction of the number of cells and no explicit technique for static reduction is implemented.

### 3.1.3 Genesys' Design for Leakage Test (DFLT)

This tool, provided by Genesys Testware, Inc., a supplier of yield and cost optimization tools, is not for leakage optimization, but it is mentioned here because it can be considered as a possible complement for any leakage optimization tools, and in particular those that use power gating..

Design For Leakage Test (DFLT) provides the automatic insertion of so-called DFLT circuitry around power controllers, power switches and isolation gates in designs that use power gating to reduce leakage, with the objective of improving the testability.

In fact, faults in power controllers, power switches and isolation gates cannot be detected by traditional DFT schemes and ATPG software. In some sense, it can be seen as a circuit-level leakage-related tool, although it does not provide any optimization; thus, it is not a competitor of RTL/Gate CLEAN tools.

## 3.2 Behavioral and system-level tools

### 3.2.1 Cadence

[http://cadence.com/lowpower/index.aspx?lid=low\\_power](http://cadence.com/lowpower/index.aspx?lid=low_power)

Cadence announced a system level tool with power capabilities. However, the tool is not yet available, nor is clear, what exactly it will support. The Company proposes a flow from specification to GDSII, using CPF to import and refine the power intent in tools on every level of abstraction.

### 3.2.2 ChipVision

<http://chipvision.com/press/2008-04-22.php>

ChipVision released the new high-level power synthesis tool PowerOpt. A prototype of that software was already shown at the last CLEAN review. At the same time, ChipVision coled the high level power estimation tool ORINOCO DALE. Most of the functionality of DALE, however, will also be supported by the new tool. This means especially the prototypes, developed in the CLEAN project and integrated in ORINOCO DALE.

Additionally ChipVision has released a system level power analysis tool. SystemC-based system descriptions can be instrumented with API-calls. The power models of the different parts of the design will be modeled by a power state machine. While simulating the design, the API communicates with the model and a power over time graph can be displayed. The power can be broken down to a dynamic and a leakage part.

### 3.2.3 Synopsys

<http://www.synopsys.com/eclipse/eclipse.html>

Synopsys offers a complete toolbox, called "Eclipse Low Power Solution" implementing an end-to-end low-power multi-voltage methodology. The solution supports static, as well as dynamic voltage scaling and power gating.

### 3.2.4 Power Standards

#### UPF

The standardization body "IEEE p1801" was kicked of after the release of UPF v.1.0 The UPF 1.0 specification was donated to that committee and has to developed to UPF v.1.1. The current version is in the process of ballot and will become almost certainly the IEEE 1801 standard later this year.

**CPF**

The standardization body “Low Power Coalition” has voted for a new minor release of that standard, too. There is a working-group instantiated, defining the extended format. The main features will be:

- Hierarchical flow support
- Memory modeling styles and support
- Gate level verification flow support
- Power estimation support
- Clocking and related updates are required to drive power optimization

## **4 Conclusions**

As anticipated in the previous release of this state-of-the art and market survey (i.e., D5.1.1.2), the large amount of research contributions in the domain of leakage power modeling and optimization at different levels of abstraction is not paired by a corresponding effort from the EDA stand-point. The market of leakage optimization tools is still quite poor, and this clearly points to the opportunities that the activities in CLEAN will create for filling the gap.

Obviously, market analysis and survey will continue throughout the entire duration of the project, in order to guarantee an up-to-date picture of possible competitors, and thus enable prompt reaction by the CLEAN Consortium.