Power modeling of a NoC based design for high speed telecommunication systems

Philippe Grosse[1], Yves Durand[1] and Paul Feautrier[2]

[1] CEA Léti, Grenoble F-38000, FRANCE
[2] ENS de Lyon/LIP, Lyon F-69364 , FRANCE

Abstract. Considering the complexity of the future 4G telecommunication systems, power consumption management becomes a major challenge for the designers, particularly for base-band modems functionalities. System level low-power policies which optimize dynamically the consumption, achieve major power savings compared to low level optimisations (e.g gated clock or transistor optimisation). We present an innovative power modeling methodology of a 4G modem which allows to accurately qualify such low power solutions. Then, we show the energy savings attended by these power management methods considering silicium technology.

1 Introduction

Managing the complexity of the future 4G telecommunication protocols within the energy constraints of a mobile terminal, is a major challenge for systems designers. In particular, the base-band modem is 10 times more complex in 4G applications compared to current 3G systems. Assuming that the consumption part due to base-band will increase as well (30% in 3G chips), power management of such processing parts must be considered carefully. Whereas low level optimisations (gated clock, low power transistor design...) reduce the consumption by 15 to 45 % [13], higher level methods, by dynamically optimising power consumption, improve further energy savings. But, these optimisations influence the system latency and are constrained by the application. To qualify such low power policies, we require an accurate consumption model which takes into account the real temporal profile of the application. Gate level power analysis is a mature and accurate way to evaluate the power consumption of digital SoC but requires long simulation times and is unpractictable for actual systems. The use of existing power models for circuits like CPU, bus, memories [10][11][12], is another solution. But none exists for such heterogeneous and complex design. Therefore, the first step of our study is to build a consumption model of a base-band modem which would be a trade-off between accuracy and simulation performance.
In this paper, we first expose our modeling methodology and then, evaluate the impact of well known low power policies at logical and system level. The paper is organized as follows: section 2 presents the base-band design under consideration

and the specific consumption issues encountered; section 3 details the modeling method and simulations scenarios; the results of the application low power policies are presented on section 4 and we conclude in section 5 by indicating future works.

2 Power and energy consumption of a 4G modem

Our study and modeling works is focused on the power consumption of a 4G MC-CDMA base band modem. The design under consideration is based on a set of independents hardware fonctionnal units linked by an innovative asynchronous Network On Chip detailed below[1],[2],[3].

2.1 Design overview

Compared to data rates targeted by 3G systems (1 to 2 Mbps), the 4G transmission systems address higher data rates (100 Mbps and more) and so drastically increase the data traffic in the chip itself. Whereas computing power of DSPs or general purpose processors was sufficient for 3G modulations techniques, latency constraints and data rates of MC-CDMA transmission requires the design of dedicated and powerful hardware blocks (FFT, channel estimation) tailored to support such data flows [2]. The design is also constrained by the on-chip connections throughput and traditional bus solutions are no longer appropriate for such applications. [2][3] To satisfy these constraints, the FAUST project (Flexi-

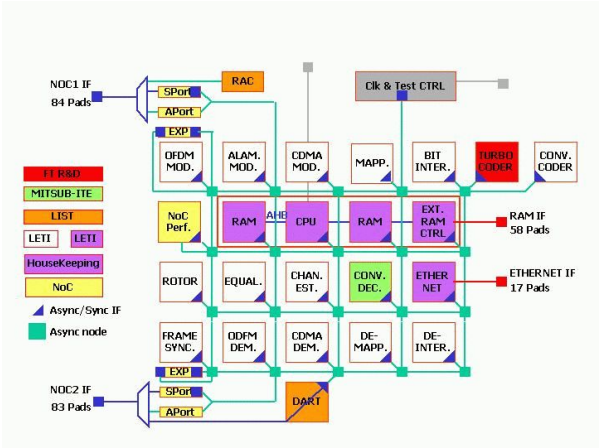


Fig. 1. Architecture of FAUST Chip

ble Architecture of Unified Systems for Telecom), started on 2003, proposes an innovative architecture of an OFDM base-band modem which can be used for

4G transmissions. FAUST consists of a set of independent and re-configurable hardware blocks, which performs the MC-CDMA base-band tasks, controlled by a CPU (ARM 946) and interconnected by an asynchronous network on chip (NoC) as depicted on Fig. 1.

2.2 Problematic of power modeling

In this paragraph, we recall briefly the main consumption parameters that we consider meaningful in our context. Next, we describe the main instruments that we use to reduce consumption, called here "low power policies".

State of the art Power consumption estimation is the purpose of numerous tools and studies. Physical models deduced from architecture and technology exist for memories or on-chip connections [11]. For CPU, models deduced from instructions profiling are common for embedded processors [10][14]. These models are accurate but specific. Thus, adapting and combining them to build a consumption model of our chip does not guarantee accurate results.
CAD tools propose others solutions to perform accurate power analyses at each step of the design. SPICE [15] (transistor level) and SYNOPSYS "Prime-Power" (gate level) [7] relate the power dissipation with the activity by using activity files coming from behavioral simulations. CADENCE "SoC Encounter"[17] integrates statistical analysis functionnalities to quickly evaluate the power dissipation of a whole SoC. Qualify low power managements methods implies numerous iterations and simulation times becomes unpracticable for real design.
Hereafter, we identify the main characteristics of our base-band consumption in order to build a model which can be a trade-off between accuracy and simulation performance.

Power consumption of a 4G modem Power consumption of such architecture corresponds to the summation of the functionnal units contributions (which performs MC-CDMA base-band functions), NoC consumption, IO pads and CPU. Like any digital architecture, the consumption of theses design units corresponds to the sum of their static (leakage) and dynamic (switching) power dissipation which is given by [6]:

$$P_{tot} = P_{leak} + P_{switch} = V_{dd} * I_{sub} + \frac{1}{2} * \alpha * C_{load} * Vdd^2 * f_{clock} \qquad (1)$$

Where α represents the switching activity; C_{load} is the output capacitive load; linked with the technology and the number of gates, V_{dd} is the supply voltage and f_{clock} the frequency of the unit. I_{sub} represents the major source of static power dissipation, the sub-threshold leakage, and is given by [5]:

$$I_{sub} = k * e^{\frac{-Q*Vt}{a*kT}} \qquad (2)$$

Thus, Vt represent the threshold voltage which is a technology linked parameter. Power dissipation of our base-band modem is so function of :

– **Technology** Vt,C_{load}
– **Runtime parameters** f_{clock}, V_{dd},α

We consider two low power policies, usually applied on system level, for our base-band modem.

System level low power policies Two common ways are used to reduce power consumption dynamically: *resource shut down* (Dynamic Power Management) and *resource slow down* (Dynamic Voltage scaling)[5][6].
DPM techniques avoid static and dynamic dissipation by shutting down all supply sources during idle states. These methods may be tailored to our design and save the static and dynamic energy wasted during idle states. But they have two major consequences on the timing of the application :

– Reconfiguration after shut down is needed
– Respond time of the supply source must been considered (almost 5μs for our blocks)

DVS, by reducing Vdd is a well known method to reduce P_{switch} due to its quadratic contribution. However, reducing this supply voltage increases the gate traversal delay (and thus the global delay) D as given by [4]:

$$D \propto \frac{Vdd}{(Vdd - Vt)^2} \tag{3}$$

Considering this delay D, frequency must be decreased linearly with the supply voltage in first order approximation ($f \propto V_{dd}$). This, in turn, increases the computation time, inversely proportional to the voltage supply. Moreover, Vt is also reduced in new technologies and causes an exponential growing of the standby current (Equ. 2) [6]. Energy savings of a DVS policy will be also both function of application (computation times and timing constraints) and technology (Vt scaling).
These power management methods, usually applied on the CPU [15] [16], may as well apply to our design considering our set of hardware units as a set of "task" schedulable by our central processing unit.
To summarize, we identify keys parameters of the power dissipation of our architecture:

– **Technology** (Vt,C_{load})
– **Architecture** (reconfiguration times)
– **Temporal consumption profiles and run time parameters** (idle time, active time,f_{clock}, V_{dd},α)
– **Real time constraints**

At this stage, it should be noted that the two first parameter are technological, whereas the two others require the knowledge of the actuel run-time profile of the application.

3 Modeling the consumption of the base-band modem

We base our model on two complementary analysis: a gate power analysis, which links our model with the technology, and a functional analysis which gives an accurate image of the activity profile of our application.

3.1 Modeling methodology

We run the gate level analysis with a real MC-CDMA transmission/reception scenario on the gate-level netlists of each functionnal unit, synthesized on the current technology of the chip (130nm). We use for that a commercially available tool, SYNOPSYS Prime-Power [7]. To evaluate the impact of the technology scaling on the consumption (dynamic and static), we run a similar analysis on the same units synthesized in other technologies (65 nm). These simulations, based on real scenario, give us an accurate timing profile of the architecture power dissipation and its consumption results will be the basis of our model. In a second time, we characterize the main consumption phases and correlate them with the algorithms to build a functional model of the architecture with PTOLEMY [8]. This software, developed by Berkeley University, is designed to model embedded systems. With this software we build components simulating the consumption of each blocks of FAUST architecture.
The figure below summarizes the modeling flow:

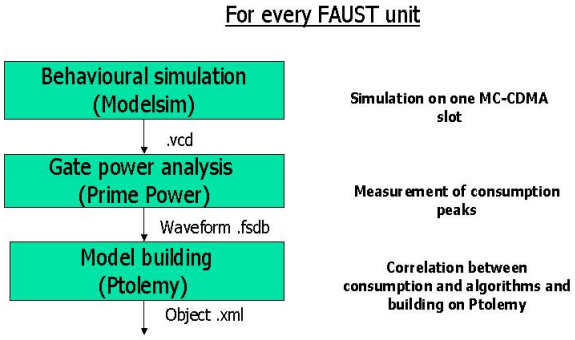


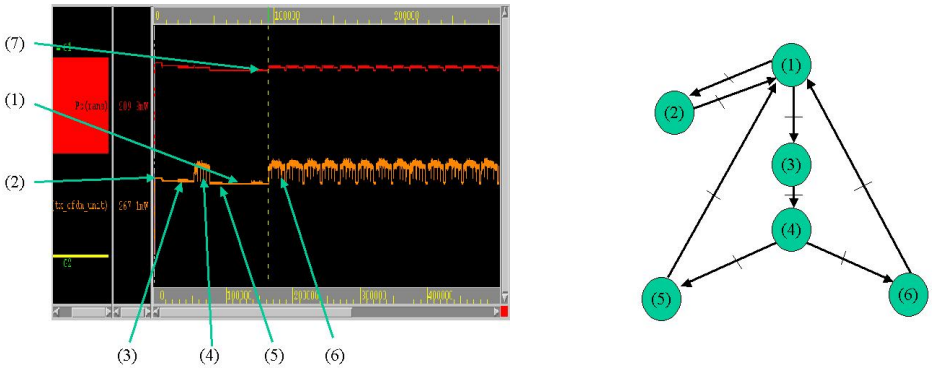
Fig. 2. Modeling flow

In the next section we detail the application of this methodology on FAUST architecture.

3.2 The FAUST consumption model

As we mentioned in section 2, FAUST architecture can be considered like a set of hardware blocks connected by a NoC and controlled by a CPU.

– **Modeling hardware blocks:** In order to build a functional model of the digital blocks consumption, we simulate every processing units by a Finite State Machine (FSM) built with PTOLEMY. The exhaustive modeling of each unit considering the consumption of each arithmetical operation (sum,division ...) is time-consuming and implies long simulation time. To avoid this drawback we build an abstracted model of each functionnal unit consumption. From the gate analysis, we identify different phases of consumption which correspond to a set of macroscopic events of our block activity, such CPU configuration burst, computation of the core or incoming data (Fig. 3).



(1) (2)

Fig. 3. Consumption profile of the OFDM modulator (1) and its FSM model (2)

The FSM models the sequence of these phases. Consumption "events" (Fig. 3) (idle state (1), configuration (2), data in (3),computation of the first symbol(4), data out (5) and computation of the other symbol (6)) derived from the power analysis constitute states and algorithmic execution conditions, transitions.

DVS instrumentation

From gate power analysis, we extract $\beta_{phase,tech}$ such as, for each phase:

$$P_{switch} = \beta_{phase,tech} * V_{dd}^2 * f_{clock} \quad (4)$$

$\beta_{phase,tech}$ depends on the technology of our base-band design (one value for the 130 nm design and one for the 65 nm model). For each consumption

phase, we calibrate this parameter with the initial conditions of the gate power analysis (supply: 1.2V, fclock : 200 Mhz). In order to instrument our model for DVS, each unit have a modifiable V_{dd} and f_{clock} value. Then, modify these parameters shows the impact of a voltage scaling on the power dissipation.
Each FSMs is synchronized by its own model of variable clock initialized at the normalized frequency of each block ($\frac{f_{clock}}{200Mhz}$) and linked with the supply voltage of the unit ($f_{clock} \propto V_{dd}$). A watch-dog monitors the global duration of transmission, which shall not exceed 666μs. Then, we guaranteed the fulfillment of our application's timing constraints.
DPM instrumentation
In our base-band design, each block independently adheres to a DPM policy. When units are idle, i.e. neither computing nor awaiting data, we shut their core down. Our abstract model reproduces this policy by setting the FSM on a non-consumptive state as soon as all pending data transfers are cleared. Therefore, data traffic regulation is explicitely simulated in each unit: it is based on a credit mechanism which used the exchange of signaling message between destination and source.

– **NoC consumption model:** All units are linked on the NoC consumption model which simulate the power dissipated by a flit (32 bits), the elementary network data unit, in an asynchronous node (we assume that energy dissipated by the wires as negligible). In FAUST, the NoC is based on a deterministic routing protocol, the number of node and the latency between two units is known. Therefore, we model our network with a "routing" table

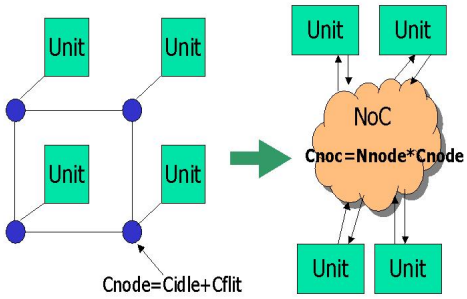


Fig. 4. Faust Noc (1) and its Ptolemy representation (2)

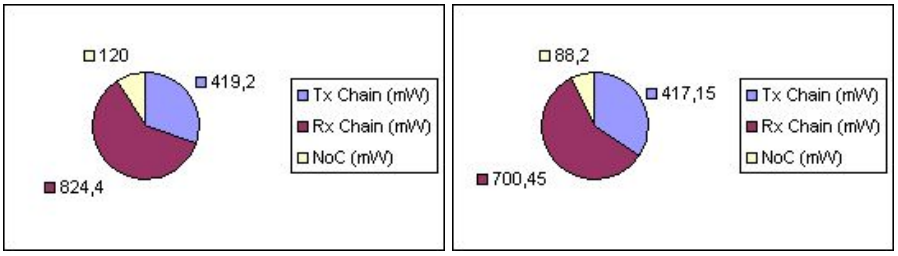
(Fig. 4), which indicates the number of asynchronous nodes that must be traversed between two units. Flit latency through the node is simulated by a clock initialized at the crossing time of a node, deduced from the gate simulation.

– **Memories consumption:** We based our RAMs models on data sheets avaiable for the 130nm technology. We relate power dissipation with the size and the state of the memories units (enable, write/read, selected...) and simulate their activity with a FSM.
– **Power model of the CPU:** FAUST integrates a CPU (ARM 946) which manages the configuration of each blocks (telecom protocol used, frequency, timing constraint). To compensate the absence of gate model we based our model on a bibliographic study which give us average consumption values of an ARM integrated in a HIPERLAN/2 and 802.11a modem [9]. In our application, the CPU is only utilized to configure the base-band modem and its activity has a limited impact on the consumption (less than 10 %). However, we can assume that an average value is sufficient to our model.
– **Final mapping:** These design units are finally assembled and form the complete consumption model of the FAUST architecture. Silicon technology, timing constraints and telecom standards are used as global parameters and are applied on all units (by modifying β for each consumption phase of each unit). Finally, all consumption values of each block are linked with a sequence plotter which give a temporal representation of the power dissipation.

3.3 Validation

In order to check the validity of our model, we compare the results of our simulation to a statistical analysis performed by a commercial tool (CADENCE SoC Encounter). On a 130 nm technology, with a statistical activity of 20% for each blocks(usually considered for such application) and a frequency of 166 Mhz we obtained the results displayed on Fig. 5.

The results given by these two analysis are almost equivalent (30 % of rela-



(1) (2)

Fig. 5. Comparison between SoC encounter (1) and Ptolemy simulations (2)

tive variation for Tx, 15% for Rx and 26% for the NoC), thus, unlike a simple statistical power analysis, our model gives an accurate timing profile of the consumption and can be useful to test low power policies.

4 Impact of low power policies combined with technological improvements

We have used our model to compare the results of different power optimization methods at each step of the design. On a first time we compare the influence of two low power policies (DPM and DVS) on a non optimized design. Then, we show the influence of one low level optimization (gated clock) on the savings performed by these high level methods. As energy is the main characteristic of batteries supply in mobile applications, our simulations are characterized in term of dissipated μjoules (mW*ms).

4.1 Energy savings of system level policies on a non low level optimized design

Gate power analysis of our initial design (130 nm) have shown that the power consumption during idle states are equivalent, in average, to 80% of the dissipated power of the computation phases. Therefore, save energy during idle phases by shutting down the units is a tempting solution compared to a voltage scaling.
This table represent the comparison between DPM and DVS policies during the transmission of one MC-CDMA slot (the macro unit of an 4G transmission duration : 666 μs) detailed by each transmission chain units. For each block the scenario is functionnally equivalent but the temporal profile, and so results of power optimizations, may vary for one to another.
E_{ref} is the reference energy dissipated by the base-band modem (130nm implementation) without any optimization and the energy savings corresponding to $\frac{E_{ref}-E_{optim}}{E_{ref}}$ with E_{optim} the energy dissipated after the use of one low power management method.

Unit	**Eref(μj)**	**Energy savings DVS**	**Energy savings DPM**
Encoder	10.34	49%	99%
Bit interleaving	24.40	54.8%	99.5%
Mapping	24.40	54.8%	99.5%
Fast Hadamard Transform	33.5	6.1%	38.8%
Ofdm modulation (IFFT)	202	14.45%	0%
Total	294.64	15.48%	24.38%

Table 1. Comparison between DVS and DPM on a 130 nm non optimized base-band modem

For a non low level optimized base-band design, a DPM policy appears more efficient than a DVS policy given to the timing constraint of the transmission. In particular, shutting down the first blocks of the transmission chain (encoding,

interleaving and mapping units) , is an easy way to save energy. Apply DPM on these units saves almost 99% of their dissipated energy. Moreover, these units represent just 20% of our transmission chain consumption, so savings performed by a DPM represent just 24% of the initial energy dissipation.
The most consuming unit (OFDM modulator), receives data request permanently during one slot transmission and can not be shut down. The only way to save energy on this functionnal unit is to scale its supply voltage (DVS).
Thus, we adopt this methodology:

- On a first time, we slow our OFDM modulator frequency (and supply) to reach the minimum admissible througput in established mode
- Then we adjust the speed of the other transmission units in accordance to our block througput

In this case, timing constraints and technology allows just a small scaling (90% of the initial value). Therefore energy savings of a DVS policy represent just 15% of our reference dissipated energy. For a non low level optimized design, DPM is so more efficient than DVS. This conclusion depend on the application and is only valid for the current power dissipation profile of the non-saturated blocks (i.e the first units).
However, technology scalings and low level optimization have an huge influence on the power dissipation profile. In the next section, we show how these low level modifications impact the choice of high level policies.

4.2 Low level optimization and consequence on the choice of low power policies

Conclusions presented above are only valid for the non power-optimized librairies used for the first release of the chip. We have to actualize this comparison between DVS and DPM for the 65nm libraries used for the second release. Netlist utilized to perform the gate power analysis were optimized with a gated clock policy generated by the synthesis tool with a low leakage 65nm library. Gated clock, by avoiding useless gate commutation, reduce the dynamic idle power dissipation by a factor of 10. Use of low power libraries allow us to neglect the leakage power dissipated (less than 5% in our application).
These low level optimizations implies less energy waste during idle state and so decrease the influence of a DPM policy. Another aspect of the technology scaling is the reduction of Vt and transistor delay which increases the maximum speed of the architecture. Considering these evolutions a DVS applied on such architecture become more interesting. Table 2 presents the results of the comparison between DPM and DVS on our base-band optimized design.

We optimize the power consumption regarding the more consuming unit, and adhere to the DVS methodology described in the precedent section. The 65nm technology allows better scaling capabilities and so better results. Table. 2 shows that, if DVS achieve less energy savings than DPM on the first blocks of the MC-CDMA chain (50% of their initial dissipated energy compared to 97%), the global

Unit	(1)	(2)	(3)	(4)	(5)
Encoder	10.34	1.02	90.2%	97.46%	49%
Bit interleaving	24.40	2.37	90.3%	97.46%	54.8%
Mapping	24.40	2.37	90.3%	97.47%	54.8%
Fast Hadamard Transform	33.5	11.03	67.4%	13.56%	54.8%
Ofdm modulation (IFFT)	202	116.4	42.37%	0%	50.5%
Total	294.64	133.19	54.79%	21.26%	51%

Table 2. Low level optimization and influence on system level policies;(1):Initial dissipated energy 130 nm without Gated-Clock during slot transmission;(2):Initial energy for a 65nm Low level optimized base band design during slot transmission;(3):% of energy savings achieved by technology scaling and low power optimization;(4):% of energy savings achieved by DPM methods on low level optimized design;(5):% of energy savings achieved by DVS on low level optimized design

savings are more significant due to the OFDM block consumption which is the critical unit of our transmission chain. From these studies and comparisons we conclude that:

- A DPM policy is only useful for blocks which are in idle state a long period of time, i.e first units performing binary treatments in our base-band.
- Low power optimizations by reducing idle power consumption, decrease the influence of a DPM policy
- Consumption of our transmission chain is conditioned by the OFDM block. An adequate DVS policy must be optimized for this critical block to achieve the higher energy savings.

Clearly,the optimal power management method consist to associate DPM and DVS, on a first time we adjust frequency and supply voltage in order to saturate the timing constraints of all blocks. In a second time, we shut down the first block of the transmission chain at the end of their treatment.
Such power management applied on a low level optimized design can reduce the consumption by 60%.

5 Conclusion and future works

Our methodology is an easy way to obtain an accurate and technology linked model of a complex SoC. With our model based on a gate power analysis, we have been able to qualify low power system level solutions on our 4G modem. Results of simulations shows how energy savings performed on computation units with system policies are dependent to the algorithm and the technology of the initial architecture and allows to target an optimal power management method. We have progressed in the exploration of low power policies for high speed telecommunication designs, but our approach remains exploratory and not very practical for industrial applications. Futures work will adress analytical tools to improve the accuracy ouf our model and methodology.

References

1. **Stefan Kaiser and al.** "4G MC-CDMA Multi Antenna SoC for Radio Enhancements", IST summit Lyon June 2004
2. **Yves Durand and al.** "FAUST: On chip distributed Architecture for a 4G Baseband Modem SoC", IP-SOC Grenoble 2005
3. **Fabien Clermidy and al.** "A NoC based communication framework for seamless IP integration in Complex Systems", IP-SOC Grenoble 2005
4. **Yann-Hang Lee, C.M Krishna** "Voltage-Clock Scaling for Low Energy Consumption in Real Time Embedded Systems"Proceedings of the Real-Time Computing Systems and Applications Conference IEEE,december 1999
5. **Ravindra Jejurika, Rajesh Gupta** "Dynamic Voltage Scaling for Systemwide Energy Minimization in real time embedded Systems", ISLPED IEEE, Newport beach 2004
6. **Kamal S. Khouri** "Leakage Power Analysis and Reduction During Behavioral Synthesis", IEEE Transactions on VLSI systems, December 2002
7. **Synopsys** "Prime Power" website : http://www.synopsys.com/products/power/
8. **Berkeley University** "The Ptolemy Project" website :http://ptolemy.eecs.berkeley.edu/
9. **Easy project team** "The EASY project design story 4", Easy Project IST 2004
10. **Amit Sinha ,Anantha Chandrakasan** "JouleTrack A web based Tool for Software energy Profiling", DAC IEEE, June 2001
11. **Dake Liu and Christer Svensson** "Power Consumption Estimation in CMOS VLSI Chips" , IEEE Journal of Solid State Circuits, June 1994
12. **Tajana Simunic and al.** "Cycle Accurate Simulation of Energy Consumption in Embedded Systems", DAC IEEE, New Orleans 1999
13. **David Chillet** "Basse consommation dans les systèmes embarqués", Roscoff-Ecole thématique, avril 2003
14. **Vivek Tiwari and al** "Power analysis of Embedded Software : A first Step Towards Software Power Minimization", Transactions on VLSI systems IEEE, december 1994
15. **Gupta R** "Formals Methods for Dynamic Power Management",ICCAD IEEE, November 2003
16. **Woonseok Kim and al** "Performance Comparison of Dynamic Voltage Scaling Algorithms for Hard Real Time Sytems", Proceedings of the Real-Time Computing Systems and Applications Conference IEEE,2002
17. **CADENCE CAD Tools** "SoC Encounter design tools" website: http://www.cadence.com/products/digital_ic/soc_encounter/index.asp