Published in IEEE Design, Automation and Test in Europe (Proceedings) 1, 1-6, 2006
which should be used for any reference to this work

Architectural and Technology Influence on the Optimal Total Power Consumption

Schuster Christian[1], Nagel Jean-Luc[1], Piguet Christian[2], Farine Pierre-André[1]

[1]IMT, University of Neuchâtel, Switzerland
[2]CSEM, Neuchâtel, Switzerland
christian.schuster@unine.ch

Abstract

In this paper, an approximated closed-form total power consumption equation for circuits working at their optimal supply and threshold voltage is presented. Comparisons of this formula to the numerical calculation show an error less than 3% on a set of thirteen 16 bit multipliers. Starting from this equation the influence of architecture transformations (including pipelining, parallelization, sequentialization) on the optimal total power is discussed. Finally, by a similar approach, the impact of the technology choice on achievable power saving is considered, showing how a moderated tradeoff between leakage and speed is the key characteristic of a good low power technology.

1. Introduction

Starting from 0.18 μm technologies, static power consumption, due to leaky "off" transistors, is now a non negligible source of power dissipation even in running mode. Thus, the total power consumption (i.e. dynamic plus static power) has to be optimized instead of simply reducing dynamic power, which is due to switched capacitance charge/discharge.

Many research efforts aim at reducing the static power consumption at the device level using for instance MTCMOS, VTCMOS, Gated-Vdd, or DTCMOS [1]. Conversely very few articles considered the joint static-dynamic power optimization at a higher level, namely at system and architectural levels [2][3][4].

For a given architecture, reducing the supply voltage Vdd leads to a reduction of dynamic power consumption, whereas it also results in a decrease of performance or speed. To compensate this effect, the threshold voltage Vth should be reduced too. Unfortunately, lowering the Vth exponentially increases the static power consumption. At a certain point, this increase in static power consumption becomes larger than the gain in dynamic power and the total power consumption becomes larger.

Therefore, between all the combinations of Vdd/Vth guaranteeing the desired speed, only one couple will result in the lowest power consumption (Figure 1). From now on, these working conditions will be called optimal working point or ideal working point. The location of this optimal working point and its associated total power consumption are tightly related to architectural and technology parameters.

Figure 1 illustrates the fact that reducing the activity allows reducing Ptot, whereas it tends to increase the optimal Vdd and Vth. As architectural modifications will change simultaneously several factors (not just the activity), it is necessary to develop a methodology to evaluate the influence of such transformations on Ptot.

One assumption along this contribution is that Vdd and Vth can be freely (and precisely) modified. Whereas the supply voltage is in general easily controllable, it is harder to modify the threshold voltage as body back-biasing becomes less and less efficient in smaller technologies. On the other hand it is possible to select a technology that matches as closely as possible the Vdd and Vth requirements. In any case, the contributions of this paper permit to understand architectural implications on the total power consumption.

The originality of this paper therefore comes firstly from the approximated closed form equation for the total power consumption at its optimal working point, expressed in terms of architectural and technology parameters. This closed form approximation is shown to match precisely the full numerical calculation. Secondly this equation is used to understand the impact of architecture on the minimal achievable power consumption under freely controllable supply voltage (Vdd) and threshold voltage (Vth) assumption. Finally the presented power formula can also be used to select a

technology flavor that is best suited for ultra low power consumption design.

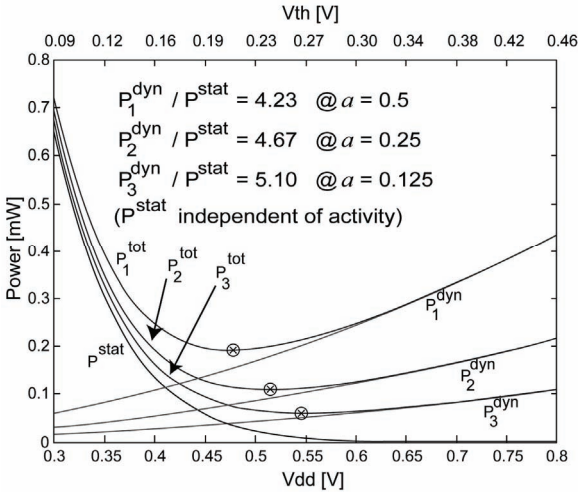

Figure 1 - Total power consumption of a 16 bit RCA multiplier in a STM 0.13 μm technology (HCMOS9GPLL) for different circuit activities (*a*). The optimal working points are marked by the cross mark, and the dynamic over static power ratio at this point is given numerically.

2. Basic equations

The closed-form approximation of total power consumption in optimal conditions that will be developed in Section 3 is based on the following fundamental equations. Total power consumption is expressed as:

$$\begin{aligned} Ptot &= Pdyn + Pstat = \\ &= a \cdot N \cdot C \cdot Vdd^2 \cdot f + Vdd \cdot N \cdot Io \cdot \exp(-Vth/nUt) \\ &= Vdd \cdot N \cdot \left(a \cdot C \cdot f \cdot Vdd + Io \cdot \exp(-Vth/nUt)\right) \end{aligned} \quad (1)$$

with N number of cells; a average cell activity (i.e. the number of switching cells in a clock cycle over the total number of cells); C equivalent cell capacitance; f operating frequency; Io average off-current per cell for $Vgs = Vth$; n slope in weak inversion; Ut = kT/q thermal voltage. Parameters a, C and Io are defined as average values per cell calculated over the full circuits. Hence architectures with different cells distributions could present slightly different parameters even for the same technology.

In Eq 1 the short-circuit power contribution is lumped in the equivalent capacitance C. The static power is here represented by the sub-threshold contribution, which is the main part in present technologies. Neglected leakage sources include: gate tunneling, which exponential depends on the oxide thickness (luckily, it can be kept reasonably low even in future technology by using an high dielectric constant insulator); p-n reverse-bias current, coming from reverse diode conduction between drain/source and body; punchthrough coming from drain and source depletion "touching" deep in the substrate.

The transistor on-current model comes from a modified version of the well known alpha power law [4]:

$$Ion = Io\left(\frac{e}{\alpha n Ut}\right)^{\alpha} (Vdd - Vth)^{\alpha} \quad (2)$$

with Io average off-current per cell for $Vgs = Vth$; n slope in weak inversion; $Ut = kT/q$ thermal voltage; α the alpha power law fitting parameter, e the Euler number; Vdd and Vth the supply and threshold voltages.

The Drain Inducted Barrier Lowering effect (DIBL) is embedded in the threshold voltage definition as:

$$Vth = Vth0 - \eta Vdd \quad (3)$$

with η the DIBL coefficient.

Then the delay can be formulated as:

$$t_{gate} = \frac{\zeta \cdot Vdd}{Ion} = \frac{\zeta \cdot Vdd}{Io\left(\frac{e}{\alpha n Ut}\right)^{\alpha} (Vdd - Vth)^{\alpha}} \quad (4)$$

with ζ (measured in Farad) a fitting parameter, which also includes the switched gate capacitance.

3. Approximated optimal total power consumption

The delay on the critical path, or logical depth (LD), must necessarily match the circuit frequency in order to operate at the optimal power condition. In fact, a positive slack would allow further reducing Vdd, resulting in additional power save. On the other hand a negative slack would correspond to a non working device. This condition can be expressed as:

$$\begin{aligned} t_{LD} &= LD \cdot t_{gate} = f^{-1} \\ Vth &= Vdd - \chi Vdd^{1/\alpha} \end{aligned} \quad (5)$$

With:

$$\chi^{\alpha} = \frac{\zeta \cdot f \cdot LD}{Io\left(\frac{e}{\alpha n Ut}\right)^{\alpha}} \quad (6)$$

Although the optimal Vdd and Vth are now tied together by mean of (5), this equation is not analytically invertible.

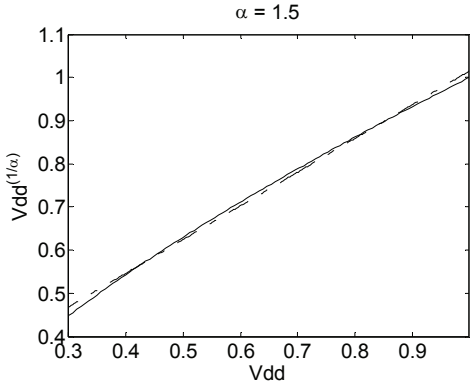

Figure 2 - $V^{1/\alpha}$ [continued line] and its linear approximation [discontinued line]

Figure 2 shows that, in a reasonable range of Vdd, the expression $Vdd^{1/\alpha}$ can be linearized:

$$Vdd^{\frac{1}{\alpha}} \approx A(\alpha)\cdot Vdd + B(\alpha) \tag{7}$$

where A and B are two fitting variables that depend on α and on the fitting range. Eq. (5) can now be rewritten as:

$$Vth \cong Vdd - \chi(A\cdot Vdd + B) = Vdd(1-\chi A) - \chi B$$
$$Vdd \cong \frac{V_{th} + \chi B}{1-\chi A} \tag{8}$$

To find the optimal Vdd, (1) is derived by Vdd and equaled to 0. Using the approximation that Vdd is much larger than nUt combined with previous equations, the two following relationships are obtained:

$$I_0 e^{-\frac{Vth\_opt}{nUt}} \cong \frac{2nUtaCf}{1-\chi A} \tag{9}$$

$$Vdd^{opt} \cong \frac{nUt\ln\left(\frac{I_0}{2nUtaCf}(1-\chi A)\right) + \chi B}{1-\chi A} \tag{10}$$

The optimal total power is defined using (1) and (9):

$$Ptot = aCNfVdd(Vdd + 2nUt/(1-\chi A)) \tag{11}$$

And for Vdd >> nUt/(1-χA), the same equation becomes:

$$Ptot \cong aCNf(Vdd + nUt/(1-\chi A))^2 \tag{12}$$

Finally the optimal Vdd (10), is introduced in (12) resulting in (13) [at the bottom of the page].

Equation 13 is a very important formula because it permits to analytically estimate the optimal total power directly from architectural parameters like activity (a), number of cells (N), frequency (f), logical depth (LD, included in χ) and technology parameters like average off-current (Io), weak inversion slope (n), alpha power law coefficient (α, included in A and B) and delay coefficient (ζ, included in χ). Thus, starting from this formula, it is possible to understand the impact of common architectural transformations, and to compare the performance of different technologies for a given architecture.

Note that (13) does no longer depend on η (DIBL coefficient) although this parameter was introduced during calculation. This can be explained by the fact that the threshold voltage is no more present in Eq. 13, hiding the DIBL effect on the same occasion.

4. Application to architecture selection

Architectural transformations will influence many parameters in (13), e.g. a, N, LD (contained in χ). Knowing the effect of transforming an architecture (e.g. pipelining or parallelization), it is directly possible to see if it will result in a higher or lower total power using (13).

For this discussion, a set of thirteen 16 bit multipliers (described in details in [5][6]) was designed in VHDL and synthesized using Synopsys Design Compiler (V2003.06). The library used for the synthesis is 0.13um CMOS09GPLL from ST Microelectronics.

1. **Ripple Carry Array or RCA** (7 flavors: basic, horizontally pipelined with 2 and 4 stages, diagonally pipelined with 2 and 4 stages, 2 and 4 parallelization): the basic implementation is constructed as an array of 1-bit adders, its speed being limited by the carry propagation. The two horizontal pipelined versions use registers inserted horizontally in the critical path (Figure 3) so that the logical depth is shortened (although not exactly divided by 2 or 4). The two diagonal pipelines (Figure 4) present a diagonal insertion of registers,

$$Ptot^{opt} \cong \frac{aCNf}{(1-\chi A)^2}\left[nUt\left(\ln\left(\frac{I_0}{2nUtaCf}(1-\chi A)\right)+1\right)+\chi B\right]^2 \tag{13}$$

achieving an even shorter LD, at the price of more glitches due to an increased spread of path delays. Finally, both parallelized versions (by 2 or by 4) are obtained by replicating the basic multiplier and multiplexing data across them. This way, each multiplier has additional clock cycles at its disposal relaxing timing constraints.

2. **Wallace Tree** (basic, 2 and 4 parallelization): the Wallace Tree structure adds the partial products using Carry Save Adders in parallel. Path delays are better balanced than in RCA, resulting in an overall faster architecture. Parallelized versions use circuit replication and multiplexing, similarly to the parallel RCA structure.
3. **Sequential** (basic, parallel and "4_16 Wallace"): the basic implementation computes the multiplication with a sequence of "add and shift" operations resulting in a very compact circuit. The intermediate result is shifted, added to the next partial product, and stored in a register.

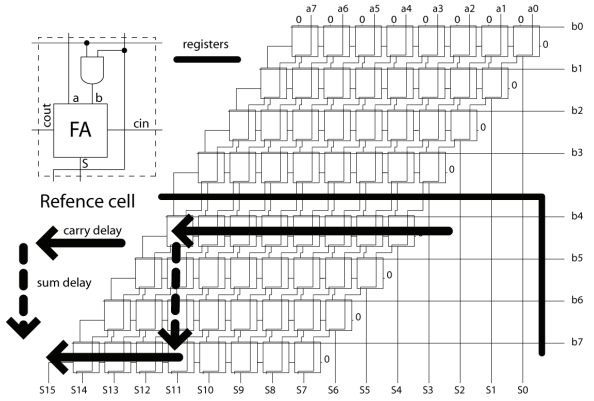

Figure 3 - 8bit RCA horizontal pipeline

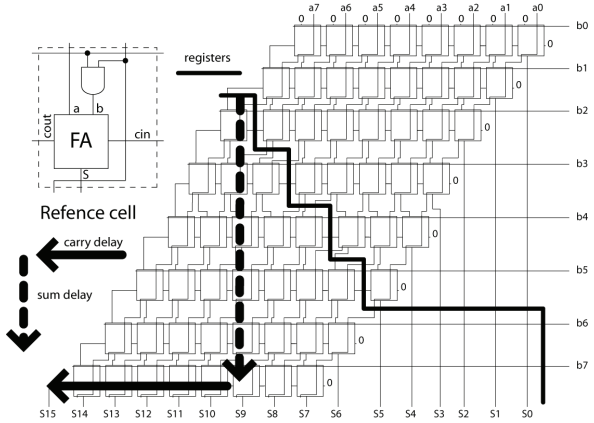

Figure 4 - 8bit RCA diagonal pipeline

This type of structure needs as many clock cycles as the operand width to complete, but only one 16-bit adder is necessary. Note, this corresponds to an internal clock running 16 times faster than the 31.25 MHz data clock that defines the throughput. The architecture called 4_16 Wallace reduces the number of clock cycles per multiplication from 16 to 4 by using a 4x16 Wallace tree multiplier i.e. by adding 4 partial products in parallel. The parallelized version is a simple replication and multiplexing of the basic version.

Starting from the values of static and dynamic power at the nominal supply voltage (Vdd = 1.2V) with activity annotated through timing annotated simulations of the netlist in ModelSIM (Mentor Graphics), the optimal total power was calculated twice. Firstly numerically from Eqs. (1)-(6) by calculating the total power for all reasonable Vdd/Vth couples, then using Eq. (13). Results are shown in Table 1.

The values of A and B used in Eq.13 were obtained by minimizing the approximation error (7) for Vdd in the range of 0.3-1.0V. All technology parameters have been estimated with Spice simulations (ELDO from Mentor Graphics) for inverters cells.

Numerical values are: A = 0.671; B =0.347; α = 1.86; n =1.33; Vt0_nom = 0.354 V; Vdd_nom = 1.2 V

The first remark that can be made on this table is that the approximation of the optimal total power based on (13) presents an error lower than ±3% compared to a numerical solution based on not approximated equations.

Moreover, by looking at the influence of architecture on optimal power consumption, several things can be observed on Table 1 and explained thanks to (13).

It is clear that sequential multipliers are not suited for low power design, unless the circuits have to work at a very low data frequency. This happens due to two additive factors. Firstly, the activity (defined with respect to the throughput frequency and not the internal clock frequency) can be very high and even bigger than 1 in some cases. This will present a high dynamic consumption at nominal conditions, but also at the optimal working point as shown by the first fraction of (13). Secondly, such architecture is very slow, resulting in a large χ, hence penalizing the total power consumption by increasing χB and reducing 1-χA (present in a square form on the denominator of the pre-factor in Eq. 13). The effect of a slow architecture can also be observed on the optimal Vdd and Vth. In fact, to respect the desired working frequency, sequential designs present high Vdd (i.e. high dynamic power) and low threshold voltage (i.e. high static power).

The RCA architecture is based on a very regular structure that permits many variations to be implemented. Both parallelization and pipelining transformations shorten the effective logical depth (which is reflected in a reduction of χ, although not in a linear manner). In this case the benefit of the relaxed timing constraints permits

Table 1 - 16 bit multipliers. All values refer to the optimal working point (f=31.25MHz). Tech. ST LL

	Cells (N)	Area [um^2]	Activity (a)	LD_{eff}	Vdd [V]	Vth [V]	Pdyn [uW]	Pstat [uW]	Ptot [uW]	Eq.13 Ptot [uW]	Eq.13 Err [%]
RCA	608	11038	0.5056	61	0.478	0.213	154.86	36.57	191.44	191.09	0.182
RCA parallel	1256	22223	0.2624	30.5	0.395	0.233	117.20	30.37	147.57	150.29	-1.844
RCA parallel 4	2455	43735	0.1344	15.75	0.359	0.256	100.51	26.39	126.90	129.93	-2.384
RCA hor.pipe2	672	12458	0.3904	40	0.423	0.225	100.51	25.27	125.78	127.25	-1.166
RCA hor.pipe4	800	15298	0.2944	28	0.394	0.238	81.54	20.94	102.48	104.34	-1.819
RCA diagpipe2	670	12684	0.4064	26	0.407	0.224	98.65	25.50	124.15	126.11	-1.581
RCA diagpipe4	812	15762	0.3456	14	0.366	0.233	82.83	22.52	105.35	108.04	-2.559
Wallace	729	11928	0.2976	17	0.372	0.236	56.69	15.17	71.86	73.56	-2.376
Wallace parallel	1465	23993	0.1568	8	0.341	0.256	55.64	15.06	70.69	72.58	-2.676
Wallace par4	2939	47271	0.0832	4.75	0.333	0.277	58.04	15.26	73.30	75.01	-2.335
Sequential	290	4954	2.9152	224	0.824	0.173	1134.00	184.48	1318.48	1318.94	-0.035
Seq4_16	351	6132	0.2464	120	0.711	0.228	184.69	31.59	216.29	212.62	1.696
Seq parallel	322	7276	1.3280	168	0.817	0.192	888.19	142.07	1030.26	1028.97	0.124

to further reduce Vdd and increase Vth, reducing this way the optimal total power consumption.

The diagonal pipeline versions present shorter logical depth but higher activity (due to more glitches) compared to horizontal pipeline, thus preferring the latter in low power pipelining techniques. In fact, when diagonally pipelining the basic RCA the critical paths will be effectively reduced more than using a horizontal pipeline, but the shortest paths will be reduced even more. This greater spread of paths delays results in a glitch increase and hence in a higher activity. This example illustrates very well how simple architectural transformations can modify the parameters like *a* and LD in a complex, and difficult to predict, manner.

Finally the Wallace family presents the fastest circuits of our set. By applying a parallelization to the basic version, we observe that, as for the RCA family, the logical depth is reduced and hence χ is also reduced. This, one more time, results in a lower Vdd and higher Vth meaning a slight power save. However, optimal total power of the further parallelized structure (Wallace par4) becomes higher than for the previous structures even if, as expected, Vdd is further reduced and Vth increased. The explanation comes from the overhead introduced by parallelization. In fact, the Wallace parallel architecture being already a fast circuit (compared to the desired working frequency), the reduction of χ is only marginal and its benefit is cancelled by the overhead in power consumption introduced by data multiplexing.

5. Application to technology selection

While Eq. (13) was discussed considering variations on the architectural parameters in Section 4, the optimal total power is also highly dependent on the technology parameters. Because current technologies often propose a choice of a few flavors or because it is sometimes possible to select one technology among several different available, we discuss here the influence of those parameters on total power consumption.

The CMOS09 0.13µm ST Microelectronics technology exists in three different flavors, namely High Speed (HS), Low Leakage (LL) and Ultra Low Leakage (ULL). The technology parameters for these cases were obtained with ELDO simulations by fitting delays on inverter chains ring oscillators:

Table 2 -STM CMOS09 technology

	Vdd nom [V]	Vth0 nom [V]	Io [E-6 A]	ζ [E-12 F]	α
ULL	1.2	0.466	2.11	7.5	1.95
LL	1.2	0.354	3.34	5.5	1.86
HS	1.2	0.328	7.08	6.1	1.58

The optimal total power was calculated for the 16 bit multipliers introduced in Section 4. Due to space limitations, only the results for the Wallace family are presented here. The results for the LL type were already reported in Table 1, while the values for the remaining two types are reported in Table 3 and Table 4.

Table 3 – Wallace family optimal power for ULL technology. (f=31.25MHz).

	Vdd [V]	Vth [V]	Ptot [uW]	Eq.13 Ptot [uW]	Eq.13 Err [%]
Wallace	0.409	0.231	84.79	86.03	-1.47
Wallace par	0.363	0.253	76.24	78.02	-2.33
Wallace par4	0.360	0.281	80.61	82.21	-1.98

Table 4 – Wallace family optimal power for HS technology. (f=31.25MHz).

	Vdd [V]	Vth [V]	Ptot [uW]	Eq.13 Ptot [uW]	Eq.13 Err [%]
Wallace	0.398	0.328	99.56	100.33	-0.78
Wallace par	0.383	0.349	110.27	111.39	-1.01
Wallace par4	0.390	0.376	118.89	119.99	-0.93

The optimal total power of the parallel version of the Wallace multiplier is higher than that of the basic one when using the HS process (Table 4), whereas it is the opposite for ULL and LL processes (Table 3,Table 1). This can be explained by the fact that parallelization (where the number of cells is more than doubled) is more penalized with technologies presenting a very high leakage. Moreover speed gain resulting from a logical depth reduction of an already rapid structure in "fast" technologies is often extremely limited.

Similarly, the optimal total power for ULL is always larger than for LL in corresponding architectures. This can be explained by the low Io and high ζ of ULL, which both lead to slower architectures as can be observed in (4). This corresponds to a higher optimal Vdd (higher dynamic power) and lower Vth (higher static power).

On the other hand the HS technology is characterized by a low α (reflected in a high A) and increased capacitance C. Both effects tend to the increase the optimal total power as predicted by Eq. (13) and confirmed by Table 4.

From these examples, it appears that under such conditions (a Wallace architecture working at 31.25MHz) the technology presenting the lowest optimal power consumption is the LL, showing that extreme technology flavors (ULL and HS) are penalized.

Starting from these observations, we can understand that a smaller technology node with ultra-high speed and large leakage might consume more than a larger techno with better balanced α, Io, ζ, etc. at its optimal working point when considering the same performances.

6. Conclusions

In this paper, three important subjects have been discussed around the theme of total power optimization for adjustable values of supply voltage (Vdd) and threshold voltage (Vth). In the first part, an analytical approximated formula for total power consumption (static plus dynamic consumption) at the optimal working point (where the minimum power is obtained while still maintaining speed requirements) is derived. Practical results show an error lower than 3% as compared to full numerical computations.

Starting from this equation, a discussion of the architecture influence on the total power was presented. The first observation was that sequential circuits are highly penalized due to the high activity and large effective logical depth.

Then, parallelization was beneficial as long as the architecture did not already present a short LD. Otherwise the multiplexing overhead completely cancelled the benefit brought by relaxed timing constraints. This was for instance the case for Wallace structures.

For pipeline transformations, it was interesting to observe that a diagonal pipeline, presenting a shorter logical depth than the horizontal one, was penalized due to the increased number of glitches (reflected by the increase in activity).

In the last part of the article, Eq. 13 was used to discuss the impact of the technology on the optimal total power. Through simple examples, it was shown how extreme technology flavors (here in the case of a STM 0.13µm technology) like Ultra Low Leakage and High Speed were less suited for low power than Low Leakage. In fact, slow or highly leaky technologies perform worst than a moderated trade-off of these characteristics when working at the optimal point condition.

Acknowledgment

This work was supported by the Swiss National Science Foundation (SNSF), under grant 105619. The authors would also like to thank TIMA for providing the STM libraries.

References

[1] K. Roy, A. Agarwal, C. H. Kim. Circuit Techniques for Leakage Reduction, chapter 13 of *Low-Power Electronics Design*, CRC Press, edited by C. Piguet, 2005.

[2] A. P. Chandrakasan, R.B. Low Power CMOS Digital Design.*IEEEJSSC, 473-484, vol. 27, no.4,* April 1992.

[3] S. Martin, K.F. Combined Dynamic Voltage Scaling and Adaptive Body Biasing for Lower Power Microprocessors under Dynamic Workloads. *ICCAD,* 2002

[4] K. Nose, T.Sakurai. Optimization of Vdd and Vth for Low-Power and High-Speed Applications.*469-474, ASP-DAC,* January 2000.

[5] C. Schuster, J.-L. Nagel, C. Piguet, P.-A. Farine. Leakage reduction at the architectural level and its application to 16 bit multiplier architectures. *Proc. Int'l Workshop on Power and Timing Modeling, Optimization and Simulation*, Santorini Island, Greece, September 15-17, 2004.

[6] C. Piguet, C. Schuster, J-L. Nagel. Optimizing Architecture Activity and Logic Depth for Static and Dynamic Power Reduction. *Proc. of the 2nd Northeast Workshop on Circuits and Systems, NewCAS'04*, June 20-23, 2004, Montréal, Canada.

[7] C. Schuster, J.-L. Nagel, C. Piguet, P.-A. Farine. An Architecture Design Methodology for Minimal Total Power Consumption at Fixed Vdd and Vth. *Journal of Low Power Electronics*, Vol.1 No.1, April, 2005.