PATMOS ’04, Santorini Island, Greece, Sept. 15-17, 2004

Leakage reduction at the architectural level and its application to 16 bit multiplier architectures

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

[1]IMT, University of Neuchâtel, Switzerland
{christian.schuster, jean-luc.nagel, pierre-andre.farine}@unine.ch
[2]CSEM, Neuchâtel, Switzerland
christian.piguet@csem.ch

Abstract. In very deep submicron technologies (< 0.13 µm) the leakage power consumption becomes an important contribution to total power consumption. Consequently a new low-power design methodology will be required at circuit, architectural and system levels. This paper focuses on architecture comparison and aims at selecting the one with the minimum total power consumption by simultaneously optimizing static and dynamic power dissipation. This optimal power has been estimated for eleven 16-bit multiplier architectures.

1 Introduction

Predictions of the SIA Roadmap [1] forecast supply voltage (*Vdd*) as low as 0.8 to 0.5 V in year 2018. Since a reduction of *Vdd* requires an additional reduction of the transistor threshold voltage (*Vth*) to maintain speed, this will result in an exponential increase of the static power consumption. As a consequence the static power on its own may already exceed the required maximum total power consumption in stand-by mode for certain applications. When the activity is very low, i.e. when very few gates are switching during a clock cycle, the static power will become a significant contributor to the total power consumption.

Table 1 summarizes predictions of the main parameters required for the next generations of Bulk MOSFET devices in 2018, related to High Performance (HP), Low Operating Power (LOP) and Low Standby Power (LSTP) devices [1][2]. LSTP transistors have high *Vth* and present thus reduced leakage compared to other devices. Conversely, HP transistors present the largest drive current for maximum speed but suffer from a very large leakage.

Table 1. Predicted (year 2018) MOS characteristics for High Performance (HP), Low Operating Power (LOP) and Low Standby Power (LSTP) devices (2003 SIA Roadmap)

MOS Type	Physical length [nm]	Supply Voltage [V]	Threshold Voltage [V]	Ion [µA/µm]	Ioff [nA/µm]	Ion/Ioff
HP	7	0.7	0.11	2190	500	4380
LOP	9	0.5	0.17	950	30	31’667
LSTP	10	0.8	0.4	990	0.1	9’900’000

New design methodologies taking into account the contribution of leakage will be *required* for the conception of innovative System-on-Chips realized in future very deep sub-micron technologies (e.g. 18 nm in 2018), noticing that leakage gains already in importance for state-of-art technologies (e.g. 90 nm in 2004). Similarly to previous design methodologies, solutions can be proposed at different levels, namely at *circuit*, *architecture,* and *system* levels.

At the circuit level many techniques have been proposed:

- Multi *Vth* technology, with fast low *Vth* transistors on critical paths and slow high *Vth* transistors outside critical paths (MTCMOS, Multiple Threshold) [3][4][5][6]
- Electrical regulation of *Vth* (VTCMOS, Variable Threshold, SATS) [7][8][9]
- Gated-Vdd, with *Vdd* switched off in sleeping mode [11][12][13].
- DTMOS (Dynamic *Vth*) with transistor bodies connected to MOS gates [10][11]

Only few publications can be found on the reduction of leakage at architectural level. It is however obvious that some digital architectures are better than others in terms of leakage. The next two sections describe trade-offs between dynamic and static power from a theoretical point of view with the aim of understanding the influence of architectural parameters on this trade-off and on total power consumption. The last part of the paper compares several implementations of 16 bit multipliers and how they compare in terms of total power consumption.

2 Optimal Total Power Consumption

The total power dissipation is defined as the sum of dynamic and static contributions:

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

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; *Is* average transistor current per cell for *Vgs* = *Vth*; *n* slope in weak inversion; *Ut* = kT/q thermal voltage.

Note that power analysis was performed at the library cell level. Short circuit power was included in the switching power as an increased equivalent capacitance. Subthreshold leakage was considered in static power, while gate and junction leakage were neglected. Considering the delay of a particular cell to be constant, the parameters *Vdd* and *Vth* can be related:

$$CellDelay \propto \frac{C_L Vdd}{(Vdd - Vth)^{\alpha}} = cst \quad (2)$$

with C_L cell load capacitance; α velocity saturation factor.

Therefore for every value of *Vdd*, there is one unique corresponding *Vth*, depending on speed requirements, which can be expressed as a function of *Vdd*:

$$Vth(Vdd) = Vdd - \chi Vdd^{1/\alpha} \quad \textbf{(3)}$$

where χ is a parameter depending on C_L, logical depth (LD), frequency and technology parameters.

For any given architecture it is therefore possible to reduce *Vdd* and *Vth* while maintaining a constant throughput. Whereas *Pdyn* decreases in a quadratic manner with the reduction of *Vdd*, *Pstat* increases instead (due to reduced *Vth*). The total power presents thus a minimum i.e. an *optimal* power consumption (marked with a ⊗ symbol in Fig. 1).

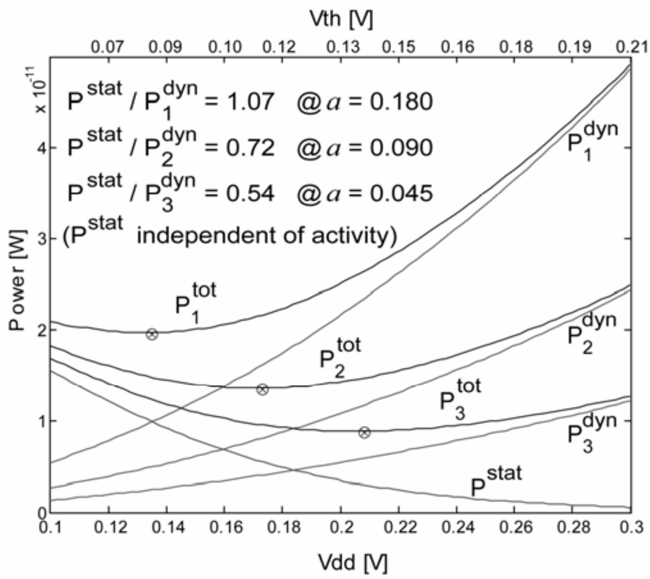

Fig. 1. Total Power (*Ptot*), dynamic Power (*Pdyn*) and static Power (*Pstat*) of a single square transistor, as a function of *Vdd* and *Vth* for three different circuit activities *a*

3 Dynamic over Static Power Ratio at the Optimum *Vdd* and *Vth*

As can be observed in Fig. 1 the ratio of dynamic (*Pdyn*) over static (*Pstat*) power at the minimum of total power consumption is not constant and depends on some some parameters e.g. the activity as illustrated in Fig. 1, but also on the logical depth, working frequency, etc. Without loss of generality we can write:

$$Pdyn^{opt} = k_1 \cdot Pstat^{opt} \quad \textbf{(4)}$$

where k_1 is expected to depend in a complex fashion on architectural parameters.

At the optimal point, the circuit works at its maximal frequency, i.e. the entire clock period is used and the last logical cell on the critical path switches just before the end of the clock cycle. Consequently, the frequency can be rewritten as:

$$f = f_{\max} = \frac{Ion}{LD \cdot C \cdot Vdd} \quad \textbf{(5)}$$

So the ratio between *Pdyn* and *Pstat* (4) becomes:

$$a \cdot N \cdot C \cdot Vdd^2 \frac{Ion^{opt}}{LD \cdot C \cdot Vdd} = k_1 \cdot N \cdot Vdd \cdot Ioff^{opt}$$

$$\frac{Ion^{opt}}{Ioff^{opt}} = k_1 \cdot \frac{LD}{a} \quad \textbf{(6)}$$

A first approximation [16, 17] was obtained by considering that the optimum of the total power consumption is obtained with the same amount of static and dynamic power (k_1=1). From the result shown in Fig. 1 this approximation is clearly too rough.

The parameter k_1 can be obtained by searching the minimum of *Ptot*(*Vdd*) as:

$$\frac{\partial Ptot(Vdd)}{\partial Vdd} = \frac{\partial Pdyn(Vdd)}{\partial Vdd} + \frac{\partial Pstat(Vdd)}{\partial Vdd} = 0 \quad \textbf{(7)}$$

Combining (3) and (7) leads to:

$$k_1 = \frac{Pdyn^{opt}}{Pstat^{opt}} = \frac{(\alpha - 1)(\eta + 1)Vdd^{opt} + Vth^{opt}}{2nUt\alpha} - \frac{1}{2} \quad \textbf{(8)}$$

with α velocity saturation factor of the alpha power law; η drain induced barrier lowering (DIBL) coefficient; *n* slope in weak inversion.

Introducing the numerical values of a 0.18µm technology (UMC L180 process) in (8), the expression of k_1 (which has a scalar size) becomes (with α=1.5, n=1.3, T=300K, η=0.033):

$$k_1 \approx 5 \cdot Vdd^{opt} + 10 \cdot Vth^{opt} - 0.5 \quad \textbf{(9)}$$

Equation (9) shows that k_1 linearly depends on $Vdd^{\,opt}$ and on $Vth^{\,opt}$ that in turn depend on LD an *a*. Consequently k_1 is as well depending on *a* and LD. This can be observed in Fig. 2 (left) where k_1 is obtained from Eq. 9 and is compared to a k_1 obtained from simulation using *Pdyn* over *Pstat*. Section 4 describes in more details how *Pdyn* and *Pstat* are extrapolated from power analysis at nominal conditions, but it can already be observed that simulation and calculation are in good agreement.

Interestingly, for high activity (*a* > 0.1), k_1 decreases quasi linearly with *a*, while for low activity (*a* < 0.02), k_1 decreases exponentially with *a*. The dependency of k_1 vs. LD can be similarly plotted (right part of Fig.2). It can be observed that k_1 is quasi linearly proportional to the logical depth. When combining both dependencies (i.e. left and right parts of Fig.2), it can be concluded that a little LD/*a* ratio (e.g. small LD and large *a*) results in a small k_1 and hence, referring to (6), an even smaller I_{on}/I_{off}.

The a and LD parameters can be altered by adjusting the degree of pipelining and parallelism of a given architecture. Every different architecture will thus exhibit a certain ratio I_{on}/I_{off} at the optimum of the total power consumption. This ratio I_{on}/I_{off} is generally much smaller than the values reported in Table 1 and more likely around 100 to 1'000 for the best architecture. According to Eq. 6, this would correspond to a high activity a and a small logical depth *LD*. Parallelizing a circuit reduces LD and the activity, so that LD/a remains approximately constant, but the behavior of k_1 depends on the working point on the curves illustrated in Fig. 2. For example, reducing the activity from 0.16 to 0.08 has a much smaller impact on k_1 than reducing the activity from 0.02 to 0.01. This tends to confirm that having architectures with very low activities (i.e. heavily parallel) is probably not a good choice. Pipelining an architecture reduces the LD but does not modify a. The LD/a ratio is consequently reduced, while k_1 is similarly reduce as can be seen in Fig. 2. Indeed, pipelining should be preferred to parallel structures as long as register overhead remains small.

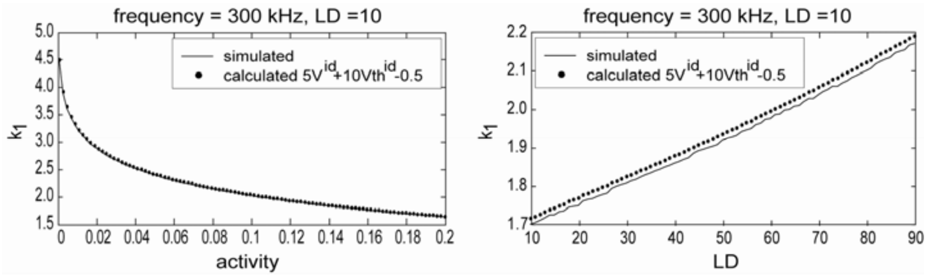

Fig. 2. Example of k_1=Pdyn/Pstat. Left: k1 vs. activity factor; right: k_1 vs. logical depth (*LD*)

4 Comparison of a variety of multiplier architectures

Provided that different architectures present different minima of total power consumption, there is presumably one optimal architecture with the smallest total power consumption. This optimal architecture will be characterized by a reduced leakage so that the dynamic power consumption can be reduced by decreasing *Vdd* and *Vth*.

The same logical function (arithmetic, filter banks, microprocessors, data paths, etc.) can be implemented using different architectures varying in their degree of parallelization and pipelining, length of critical path, number of transistors, etc. The following design parameters are proposed to compare digital architectures regarding total power and leakage: activity factor (a), logical depth (*LD*), the number of cells (*N*), the equivalent cell capacitance (*C*) and the ratio *Ion/Ioff*.

Eleven 16 bit multipliers architectures have been described and synthesized. Their power consumption was analyzed using the same stimuli for all circuits and the optimal *Vdd* and *Vth* were estimated.

The test circuits properly populate the space spanned by the above parameters, and can be classified as:

- *Ripple Carry Array* or RCA (basic, pipelined 2, pipelined 4, parallel 2, parallel 4): the basic implementation is constructed as an array of 1-bit adders (see for instance Fig. 3(a) for a 4x4 multiplier). The speed of this structure is limited by the carry propagation. The two pipelined versions (2 or 4 stages) use registers in the critical path so that the logical depth is directly shorten by the number of stages, as shown in Fig. 3(b). Finally, both parallelized versions (factor of 2 or 4) are obtained by replicating the basic multiplier and multiplexing data at each clock across different multipliers (see Fig. 3(c)). This way, the multiplier has additional clock cycles at its disposal relaxing timing constraints.
- *Wallace Tree* (basic, parallel 2, parallel 4): the Wallace Tree structure adds the partial products (P0-P7 in Fig. 3(d)) using Carry Select Adders in parallel. Path delays are better balanced than in RCA, resulting in an overall faster architecture. Parallelized versions use a circuit replica as for the RCA structure.
- *Sequential* (basic, 4x16 Wallace, parallel): the basic implementation computes the multiplication with a sequence of “add and shift” resulting in a very compact circuit. The intermediate result is shifted and added to the next partial product, and stored in a register (see Fig. 3(e)). This type of structure needs as many clock cycles to complete as the operand width, but only one 16-bit adder is necessary. The architecture called 4x16 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 (see Fig. 3(f)). The parallelized version is also a simple replica of the basic version.

For all the architectures, the throughput has been fixed to 31.25 MHz. All architectures have been synthesized with Synopsys using Virtual Silicon Technology library for the UMC 0.18μm 1.8V process. Static and dynamic power consumption were then estimated by back-annotating the cell activities with a test bench generating the same random stimuli for all circuits. The power consumption was then extrapolated at different *Vdd* and *Vth* using Eq. 1 and Eq. 3. The *Vdd* and *Vth* providing the minimum total power consumption were finally found numerically.

Results are summarized on Table 2. The architecture providing the lowest total power consumption is the structure based on the Wallace tree, operating at very low *Vdd* and *Vth*. As could be expected at these low *Vdd* and *Vth*, the *Ion/Ioff* ratio for optimal total power consumption is rather small, approximately from 100 to 1’000 (compared to that of LOP devices reported in Table 1). Moreover, the percentage of static power over the total power is between 15% and 30%, i.e. k_1 between 3 and 6, far from k_1=1. *Ion/Ioff* being very close to k_1*LD/a, Eq. 6 seems to be a good approximation.

However, it is difficult to link optimal total power consumption (last column) with Ion/Ioff (or k_1*LD/*a*) because similar k_1*LD/*a* result in quite different total power consumptions. It is the case for sequential architectures that present a very high total power consumption but average k_1**LD/a*. Indeed LD of the basic sequential architecture is very important (192, i.e. 12 times that of the basic Wallace structure) but since the activity is also huge (220%) the ratio *LD/a* stays roughly the same. The constant k_1 is similarly constant since it is increased due to the large *LD*, but simultaneously reduced due to the large *a*. Besides, *Pdyn* is large due to the high *Vdd* (relatively to other architectures) and *Pstat* is also very large due to the very low *Vth*.

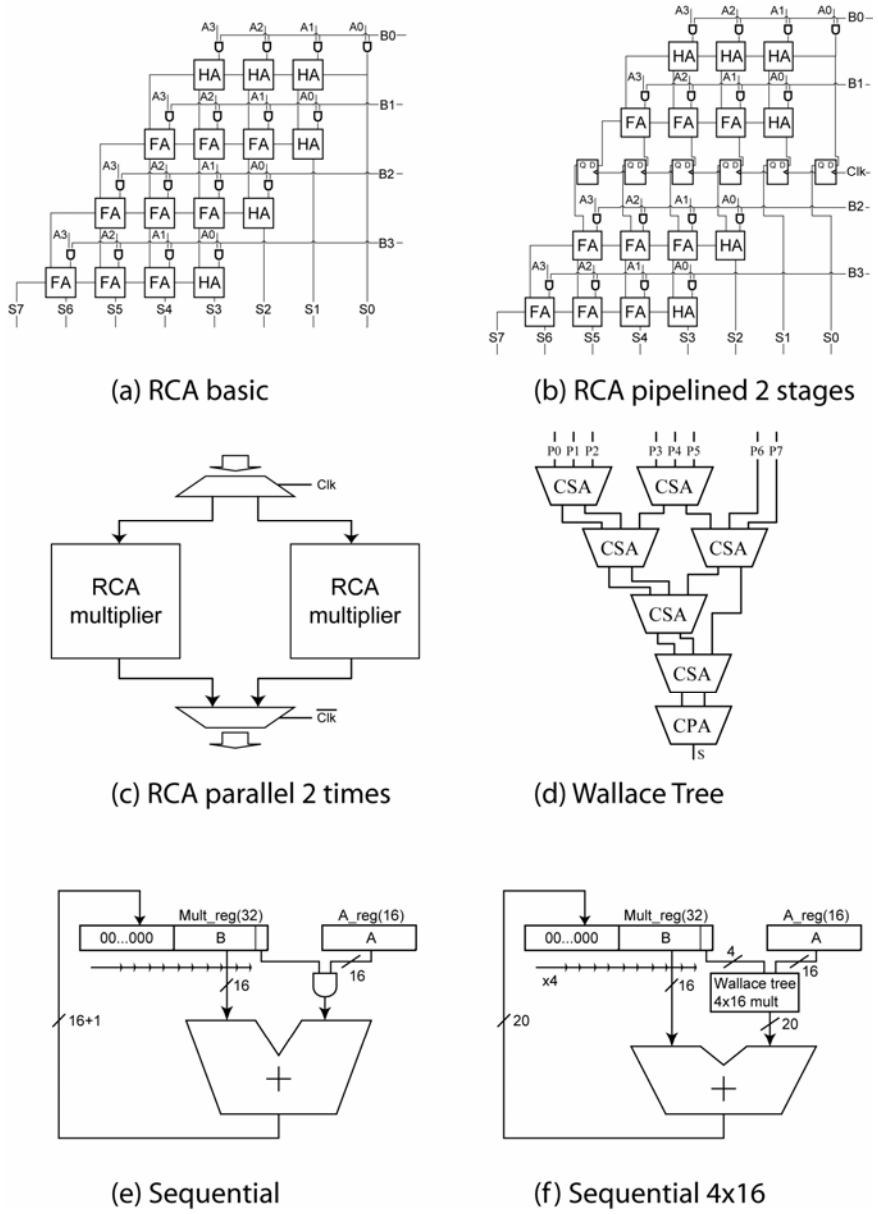

Fig. 3. Various architectures implementing 16-bit multipliers

Table 2. Main parameters characterizing the 16 bit multiplier architectures

	Area [*10^3 µm^2]	**# of cells**	**# of cell transitions**	**a**	**LD**	**k1**	**k_1LD/a**	**I_{on}/I_{off}**	**Vdd [V]**	**Vth [V]**	**Pdyn [µW]**	**Pstat [µW]**	**Ptot [µW]**
RCA	23	577	171	0.20	46	3.8	856	919	0.43	0.21	92	24	116
RCA parallel 2	47	1156	181	0.11	23	3.6	761	740	0.35	0.24	76	19	94
RCA parallel 4	92	2316	208	0.06	12	3.5	660	452	0.31	0.25	70	22	92
RCA pipeline 2	27	641	187	0.21	31	3.3	497	448	0.35	0.21	73	25	98
RCA pipeline 4	33	769	220	0.21	24	3.1	350	278	0.31	0.21	75	29	104
Wallace	25	755	199	0.20	16	3.1	254	226	0.29	0.21	43	15	58
Wallace parallel 2	50	1506	208	0.11	8	3.1	237	147	0.27	0.23	45	18	63
Wallace parallel 4	97	3017	238	0.06	4	3.3	218	112	0.27	0.25	54	21	75
Sequential	10	257	654	2.23	192	6.2	530	439	1.06	0.14	3269	522	3792
Sequential 4x16 wallace	14	355	255	0.59	100	3.8	646	432	0.53	0.17	302	84	386
Sequential parallel 2	16	325	502	1.28	160	5.6	701	634	0.91	0.16	2053	356	2409

These *Vdd* and *Vth* values are necessary to maintain the throughput of the architecture. Nevertheless the ratio *Pdyn* to *Pstat* is roughly similar to the ratio of the more parallel architectures.

Parallelizing the RCA (i.e. dividing its *LD* and *a*) shows improvements in total power. Similarly, using a two-stage pipeline RCA (i.e. dividing its *LD*, but keeping the same *a*) reduces the total power as compared to the basic RCA, as could be expected [16]. However, the register overhead becoming too important using a four-stage pipeline leads to a poorer solution. It should be noted that *LD* is not perfectly divided by the number of stages, and that the area is increased due to this overhead. As a result, the optimization of the number of pipeline stages should clearly take into account the register overhead.

Although increasing the parallelism of the RCA was beneficial, increasing the parallelism of the Wallace structure does not lead to a better solution. Even though k_1*LD/a* is reduced with the increase of parallelism, the static power consumption becomes too large due to the increased number of cells, which are inactive and leaky. The parallel versions of the Wallace architecture can no longer be operated at significantly lower *Vdd* because the static power is already important, resulting in larger dynamic and total power compared to the basic Wallace architecture.

Comparing the four-time parallel RCA and twice parallel Wallace in Table 2, it can be observed that the first is composed of 2316 cells, while the second has only 1506 cells, the ratio of areas being over 1.8. Still, both architectures present 208 cell transitions for executing a 16x16 multiply, showing that various architectures with a quite different number of cells can achieve their function with the same number of transitions. The optimum *Vdd* and *Vth* conditions used in the table are not the same for both architectures. Fig. 4 on the other hand report the same architecture in the same conditions (*Vdd* = 0.31 V and *Vth* = 0.25 V). The architecture containing fewer cells still presents a larger activity, a smaller leakage but also a smaller dynamic power (second bar in Fig. 4, at same *Vdd* and *Vth*). However, this architecture can be supplied with smaller *Vdd* and *Vth* thanks to its smaller LD (third bar in Fig. 4) to meet the same speed performances than the RCA parallel 4, with better results in total power consumption. Accordingly, an attractive goal to reduce the total power consumption would be to have fewer transistors to perform the same logic function, i.e. with the same number of transitions but more activity.

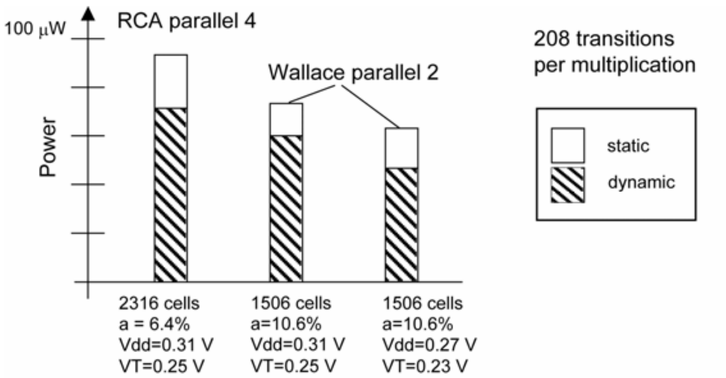

Fig. 4. Comparison of two multiplier architectures (RCA parallel 4 and Wallace parallel 2) showing the same number of transitions per multiplication but different total power

5 Conclusions

Total power consumption resulting from the sum of the static and dynamic contribution shows an optimum corresponding to very low *Vdd* and *Vth*. As illustrated in this paper, this minimum is characterized by a dynamic power somewhat larger than the static power (between 3.1 and 6.2 times in the considered multiplier architectures). The optimal ratio *Ion/Ioff* is also discussed and reported to be related to the architectural parameters *LD* and *a*. This relationship has been demonstrated for eleven 16 bit multipliers showing optimal values of *Ion/Ioff* between 100 and 1000.

Parallelization and pipelining can improve the power consumption when this leads to a reduction of *Vdd* and an increase of *Vth* (e.g. with RCA multiplier) but are not interesting for structures where the increase in static power does not allow this reduction (e.g. Wallace structure).

References

1. http://public.itrs.net/
2. Belleville, M., Faynot, O.: Evolution of Deep Submicron Bulk and SOI Technologies, Chapter 2 in Low Power Electronics Design. C. Piguet, CRC Press, (will be published in 2004)
3. Kang, S.-M., Leblebici, Y.: CMOS Digital Integrated Circuits. 2nd edn. McGraw-Hill. (1999) 466-469
4. Anis, M., Elmasry, M.: Multi-Threshold CMOS Digital Circuits. Kluwer Academic Publishers. (2003)
5. Usami, K., Kawabe, N., Koizumi, M., Seta, K., Furusawa, T.: Automated Selective Multi-Threshold Design For Ultra-Low Standby Applications. Proc. Int'l Symp. on Low Power Electronics and Design. (2002)
6. Kao, J., Chandrakasan, A.: MTCMOS Sequential Circuits. Proc. European Solid-State Circuits Conf. Villach, Austria. (2001, Sep.)
7. Von Kaenel, V., et al. : Automatic Adjustment of Threshold & Supply Voltage Minimum Power Consumption in CMOS Digital Circuits. Proc. Int'l Symp. on Low Power Electronics and Design. San Diego. (1994, Oct.) 78-79
8. Kim, C.H., Roy, K.: Dynamic Vt SRAM: A leakage Tolerant Cache Memory for Low Voltage Microprocessor. Proc. Int'l Symp. on Low Power Electronics and Design. (2002)
9. Mizuno, H., et al.: An 18µA Stand-by Current 1.8V,200 MHz Microprocessor with Self-Substrate-Biased Data-Retention Mode. IEEE J. Solid-State Circuits, Vol. 34, No. 11. (1999, Nov.) 1492-1500
10. Assaderaghi, F., et al.: A Dynamic Threshold Voltage (DTMOS) for Ultra-Low Voltage Operation. IEEE IEDM Tech. Dig., Conf. 33.1.1. (1994) 809-812
11. Soeleman, H., Roy, K., Paul, B.: Robust Ultra-Low Power Sub-threshold DTMOS Logic. Proc. Int'l Symp. on Low Power Electronics and Design. (2000)
12. Enomoto, T., Oka, Y., Shikano, H., Harada, T.: A Self-Controllable-Voltage-Level (SVL) Circuit for Low-Power High-Speed CMOS Circuits. Proc European Solid-State Circuits Conf. (2002) 411-414
13. Cserveny, S., Masgonty, J.-M., Piguet, C.: Stand-by Power Reduction for Storage Circuits. Proc. Int'l Workshop on Power and Timing Modeling, Optimization and Simulation. Torino, Italy. (2003, Sep.)
14. Clark, L.T., Deutscher, N., Ricci, F., Demmons, S.: Standby Power Management for 0.18mm Microprocessor. Proc. Int'l Symp. on Low Power Electronics and Design. (2002)
15. Piguet, C., Narayanan, V.: Guest Editorial. IEEE Trans. on VLSI Systems, Vol. 12, No 2, (2004, Feb.)
16. Brodersen, R.W., et al.: Methods for True Power Minimization. Proc. Int'l Conf. on Computer Aided Design. San Jose, California. (2002, Nov.) 35-42
17. Heer, C., et al.: Designing Low-Power Circuits: An Industrial Point of View", Proc. Int'l Workshop on Power and Timing Modeling, Optimization and Simulation. Yverdon, Switzerland. (2001, Sep.)
18. Piguet, C., et al.: Techniques de Circuits et Méthodes de Conception pour Réduire la Consommation Statique dans les Technologies Profondément Submicroniques", Proc. FTFC. Paris, France. (2003, May) 21-29