Citation: Beckett, P 2005, 'Why area might reduce power in nanoscale CMOS', in Proceedings of the IEEE International Symposium on Circuits and Systems, 2005 (ISCAS 2005), Kobe, Japan, 23-26 May 2005.

Why Area Might Reduce Power in Nanoscale CMOS

Paul Beckett
School of Electrical and
Computer Engineering
RMIT University
Melbourne, Australia 3000
Email: pbeckett@rmit.edu.au

Seth Copen Goldstein
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891
Email:seth@cs.cmu.edu

***Abstract*— In this paper we explore the relationship between power and area. By exploiting parallelism (and thus using more area) one can reduce the switching frequency allowing a reduction in V_{DD} which results in a reduction in power. Under a scaling regime which allows threshold voltage to increase as V_{DD} decreases we find that dynamic and subthreshold power loss in CMOS exhibit a dependence on area proportional to $\mathbf{A}^{(\sigma-3)/\sigma}$ while gate leakage power $\propto \mathbf{A}^{(\sigma-6)/\sigma}$ and short circuit power $\propto \mathbf{A}^{(\sigma-8)/\sigma}$. Thus, with the large number of devices at our disposal we can exploit techniques such as spatial computing–tailoring the program directly to the hardware–to overcome the negative effects of scaling. The value of σ describes the effectiveness of the technique for a particular circuit and/or algorithm–for circuits that exhibit a value of $\sigma \leq 3$, power will be a constant or reducing function of area. We briefly speculate on how σ might be influenced by a move to nanoscale technology.**

I. Introduction

Tailoring the hardware directly to the program, e.g., spatial computing [1] has the potential to overcome the negative effects of scaling. By eliminating the ISA and allowing tools such as compilers to manipulate the underlying hardware structures directly, one can optimize not only for time but also for other important metrics in the nano-electronics design space, e.g., defect/fault tolerance or power. In this paper we explore the latter: the trade-off between area, performance and power that may be possible in future CMOS devices and how these might interact at the micro-architectural level.

The nexus between area and delay (and therefore power) works at four primary levels:

- device – governing issues of technology choice and transistor sizing (W/L);
- circuit – design style and layout;
- micro–architecture – encompassing implementation issues such as asynchronous vs. synchronous, or serial vs. parallel;
- architectural - including processor decisions - super-scalar VLIW for example, or strategies such as spatial computing.

A fundamental result of early VLSI research is that for many computational functions (for example multiplication, sorting and DFT [2]–[4]) there is tradeoff between the implementation area (A) and the time it takes to compute the function (T) of a form such as:

$$AT^{\sigma} = O(n^{\sigma}), \tag{1}$$

where σ has tended to lie between 1 and 2 for traditional circuit design [3]–[6]. σ can be viewed as an indicator of how inherently sequential a circuit or algorithm is. A higher value of σ means that increasing area will not allow the overall time to be reduced. If we fix the size of the computation (i.e. n^{σ} is constant), then:

$$A^{-1} \propto T^{\sigma} \Rightarrow T \propto A^{-1/\sigma} \qquad \text{for } \sigma > 0. \tag{2}$$

This observation describes the area-time tradeoffs that are possible for a planar circuit: within bounds, one can increase circuit area to reduce circuit delay. The question we examine in this paper is whether we can use more area, not to decrease delay, but rather to reduce power consumption while keeping delay constant. Total computation time is inversely proportional to clock frequency (F) so if we fix the completion time:

$$F \propto A^{-1/\sigma} \tag{3}$$

In other words, as the area dedicated to a circuit increases, it will be possible to reduce the overall frequency of operation in order to control power usage. The fundamental question we address in this paper is: Can one use more area to reduce overall power consumption while maintaining the same delay (i.e., can we reduce energy-delay by increasing area). The power reduction comes from lowering the frequency (which allows V_{DD} to be reduced). The delay remains the same because the circuit harnesses the area to increase parallelism. σ describes the effectiveness of this technique for a particular circuit and/or algorithm the higher the value of σ, the less effective is this approach.

The remainder of the paper proceeds as follows. In Section II we examine the tradeoffs between area and power that may allow one to reduce the major sources of power consumption in nanoscale CMOS and propose a threshold voltage scaling function that will allow power, especially subthreshold power, to fall as area increases. In Section III we speculate on the affect that a move to the nanoscale may have on the area-time-power tradeoffs in CMOS and in section IV we conclude and point the way towards future research into this area.

II. Power vs. Area in CMOS

Power consumption in CMOS arises from four main sources:

1) Subthreshold leakage: $P_{\mathrm{SUB}} \propto I_{\mathrm{OFF}} V_{\mathrm{DD}}$;
2) Dynamic power (P_{DYN}), a function of capacitance (C), voltage (V), the activity factor (a), and switching frequency (F) such that P_{DYN} = $\mathrm{aCV^2F}$;
3) Short circuit switching current ($P_{\mathrm{SS}} = I_{\mathrm{ss}} V_{\mathrm{sw}}$) with I_{ss} being a function of rise-time, frequency and transistor size;
4) Gate current: P_{G} a function of logic value and transistor size.

Of these, the dynamic power terms (P_{DYN} and P_{SS}) are primarily a function of the switching frequency and capacitance (fanout and interconnect). One way to reduce dynamic power is to reduce the number of devices that switch per cycle by using asynchronous circuits [7] that eliminate the global clock (and its associated global wire), and are based on local communication and synchronization. An orthogonal approach is to decrease clock frequency by exploiting parallelism. The remaining two terms (P_{SUB} and P_{G}) represent a static power loss that is largely unaffected by either of these techniques. Gate current and leakage current strongly interact [8] and since their total is a function not only of technology (e.g. oxide thickness, dielectric etc.) but also the average gate voltage during operation, static leakage will be strongly data dependant. Static power is expected to be a primary constraint to future device scaling in CMOS [9]. The following sections explore these four different sources of power consumption and relate them to the area used to implement the circuit.

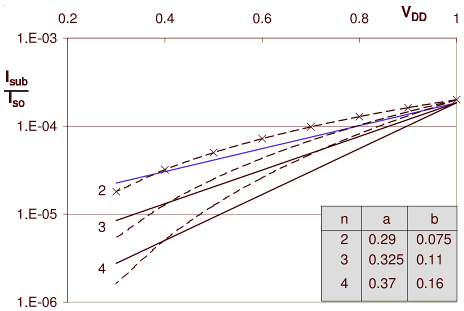

Fig. 1. A comparison of I_{SUB}/I_{SO}=$e^{-40a}e^{40bV_{\rm DD}}$ (solid lines), and $\log V_{\rm DD}^n$ for n=2,3 and 4 (dotted lines). Insert shows values of a and b used to approximate $V_{\rm DD}^n$

A. Subthreshold Leakage Power

Subthreshold leakage arises mainly due to diffusion between the source and drain when the channel is in weak inversion. In bulk CMOS, there is a small contribution from tunneling through the reverse-biased diode junction at the drain/substrate junction, but it will be negligible in future low-voltage SOI technology [9]. Direct source-drain tunneling will also be ignored in this paper as it is likely to be relevant only at gate lengths of less than 10nm.

To analyze this we can start with the BSIM3V3 transistor model for subthreshold drain current [10]:

$$I_{\rm SUB} = I_{\rm SO} \quad 1 - e^{\frac{-V_{\rm DS}}{V_{\rm t}}} \quad e^{\frac{V_{\rm GS}-V_{\rm TH}-V_{\rm OFF}}{nV_{\rm t}}} \tag{4}$$

where $I_{\rm SO}$ is a function of the transistor geometry (W/L) plus a number of process parameters and $V_{\rm OFF}$ represents a small offset from $V_{\rm TH}$ to the subthreshold region. The parameter *n* (≈1 to 2) is related to technology and is adjusted to fit the slope of the curve such that $S = 2.3nV_{\rm t}$ empirically describes $\Delta V_{\rm GS}/\Delta I_{\rm SUB}$ in mV/decade.

We are interested in the worst case power when the gate is off, i.e., the off-current ($I_{\rm OFF}$). This is the point at which the gate voltage is zero and the voltage drop from the drain to the source is highest, i.e., $V_{\rm GS} = 0$ and $V_{\rm DS} = V_{\rm DD}$. Under these conditions the first exponential is $e^{\frac{-V_{\rm DD}}{V_{\rm t}}}$. Since V$_t$(=$\frac{kT}{q}$) is small compared to $V_{\rm DD}$—and is likely to remain so into the nanoscale region—this term is approximately 0. If we assume that $V_{\rm OFF}$ is small and set $n = 1$ and $V_{\rm t} = 0.025$V, then $I_{\rm OFF}$ becomes:

$$I_{\rm OFF} \propto (W/L)\, e^{-40V_{\rm TH}} \tag{5}$$

For a given *fixed* threshold, the exponential term is a constant—independent of area—and therefore total subthreshold current will be a linear function of the number of devices (N) and therefore of increasing area i.e., $P_{\rm SUB} = NI_{\rm OFF}V_{\rm DD} \propto A$. However, if we relate $V_{\rm TH}$ and $V_{\rm DD}$ via a function of the form:

$$V_{\rm TH} = a - bV_{\rm DD}, \text{ where } a \text{ and } b \text{ are constants} \tag{6}$$

then the behavior of the second exponential term changes and it becomes possible to reduce the effect of subthreshold current—obviously with some performance cost.

Given a pair of scaling factors (a and b), the exponential term e$^{-40V_{\rm TH}}$ is transformed to $e^{-40(a-bV_{\rm DD})}$ and therefore to a product of two terms: e^{-40a} (i.e. a constant) and $e^{bV_{\rm DD}}$. As illustrated in Fig. 1, the subthreshold power can be approximated by $V_{\rm DD}^n$ for $n =$ 2 to 4 down to $V_{\rm DD} = 0.4$—the 2018 ITRS target for low-power SOC [11]. For example, if we choose to set *a* and *b* such that the curve approximates $V_{\rm DD}^2$, then subthreshold current becomes $\propto AV_{\rm DD}^3$ and finally:

$$P_{\rm SUB} \propto A^{(\sigma-3)/\sigma} \tag{7}$$

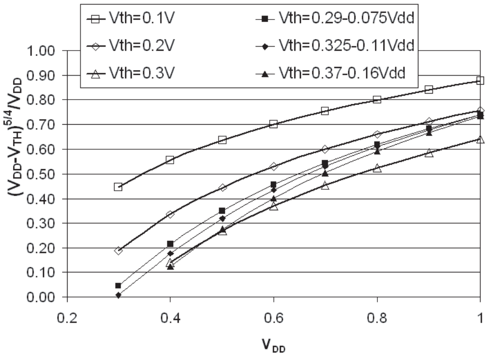

Fig. 2. $(V_{\rm DD}$-$V_{\rm TH})^{\frac{5}{4}}/V_{\rm DD}$ vs. $V_{\rm DD}$ for both the fixed $V_{\rm TH}$ case and $V_{\rm TH}$ = a-b$V_{\rm DD}$ for various *a,b* as in fig. 1.

Thus with careful management of the relationship between threshold and supply voltage, subthreshold power can be made to be a reducing function of area for micro–architectures for which $\sigma \leq 3$.

B. Dynamic Switching Power

Dynamic (switching) power, given by $P_{\rm DYN} = aFC_{\rm L}V^2$, is intrinsic to CMOS. Combining this with (3) leads to:

$$P_{\rm DYN} \propto C_{\rm L}V^2A^{-1/\sigma} \tag{8}$$

In scaled CMOS, power may be further reduced because the maximum switching speed is a function of supply voltage. For example, Chen et al [12] determined that, if load capacitance is held constant, frequency scales with voltage as $F(=\frac{1}{T}) \propto (V - V_{\rm TH})^{\frac{5}{4}}/V$. If we fix $V_{\rm TH}$, then as Flynn et al. [13] point out, $F \propto V$. The assumption of constant load capacitance is valid for organizations such as asynchronous spatial architectures that exhibit small Rent exponents for which the fanout and interconnect length do not depend greatly on the size of the circuit [14]. We can thus conclude that $P_{\rm DYN} \propto C_{\rm L}F^3$. Combining this with (3) leads to $P_{\rm DYN} \propto C_{\rm L}A^{-3/\sigma}$. Since total capacitance is roughly proportional to area, the dynamic power becomes:

$$P_{\rm DYN} \propto A^{(\sigma-3)/\sigma} \tag{9}$$

Eqn. 9 states that one can hold delay constant and still decrease power by utilizing more area. There are two main requirements for this to hold: (1) capacitance per node must be constant and $V_{\rm TH}$ must remain fixed as $V_{\rm DD}$ is lowered. As we argued above, the first holds for design styles which emphasize local wires. The second arises from the frequency being proportional to $V_{\rm DD}$ as long as $V_{\rm TH}$ remains fixed. However, in order to obtain Eqn. 7 for subthreshold power we related $V_{\rm TH}$ and $V_{\rm DD}$ as in Eqn. 6. As Fig. 2 shows, allowing $V_{\rm TH}$ to vary as $V_{\rm DD}$ is scaled still maintains the necessary near linear relationship between F and $V_{\rm DD}$. However, it also indicates that there will be a penalty in delay of up to a factor of two. This penalty is necessary in order to control the subthreshold power. In other words, while power can be reduced, there may be some increase in energy–delay, depending on the value of σ.

Thus, if one correlates the scaling of supply and threshold voltages as described in the previous section, power can be made to exhibit the form A$^{(\sigma-3)/\sigma}$ for both dynamic and subthreshold power and one can exploit area to reduce power.

C. Short Circuit Power

Short circuit power represents only a small percentage—typically 10-20%—of the overall dynamic power figure as long as the gate is loaded such that the input and output signals exhibit approximately equal rise and fall times. If this is not the case—for example with small fanout and local interconnect—then the short-circuit dissipation may exhibit the same order of magnitude as the load switching power. The unloaded case therefore represents an upper bound on the short circuit power.

We can start with the equation for average short circuit current (I_{AVE}) derived by Veendrick [15] for the unloaded case—$C_{\mathrm{L}} = 0$ and with $\mathrm{W_P}$ and $\mathrm{W_N}$ adjusted to compensate for mobility differences:

$$I_{\mathrm{AVE}} = \frac{1}{12}\frac{\beta}{V_{\mathrm{DD}}}[V_{\mathrm{DD}} - 2V_{\mathrm{TH}}]^3\frac{\tau}{T} \tag{10}$$

where β = device gain, τ = input rise/fall time and T is the clock period (1/F). As mentioned previously, we are assuming that we are dealing with circuits that will exhibit low Rent exponents, such that the average fanout and interconnect length asymptotes to a small fixed value as the size of the circuit increases. As a result, both β ($\propto W/L$) and τ ($\propto C_{\mathrm{L}}\ V_{\mathrm{DD}}\ /I_{\mathrm{d}}$) can be taken to be independent of area and since device capacitance and drive current are both directly proportional to gate area, τ will also be independent of device size and proportional only to the supply voltage, V_{DD}.

Even if this is not the case, typical values of τ/T tend to be small (< 0.1)—which is why P_{SS} is ignored in most power analyses. Further, Equ. 10 holds only where V_{DD} is greater than the sum of the device thresholds ($\mathrm{V_{TP}}$+$\mathrm{V_{TN}}$). When V_{DD} falls below this point, I_{AVE} becomes zero, as it is not possible for both transistors to be on simultaneously over the full range of gate voltages. Substituting $\tau \propto V_{\mathrm{DD}}$ into Equ. (10) and eliminating both β and the constant $\frac{1}{12}$, simplifies it to:

$$I_{\mathrm{AVE}} \propto [V_{\mathrm{DD}} - 2V_{\mathrm{TH}}]^3 F \tag{11}$$

Figure 3 plots $(V_{\mathrm{DD}} - 2V_{\mathrm{TH}})^3$ against V_{DD} over the range 0.4V $\leq$ V_{DD} $\leq$ 1.0V, and for various fixed values of V_{TH} along with V_{TH} =0.29-0.075V_{DD}. Also shown are plots of V_{DD}^n for various values of n. It can be seen that, just as for the subthreshold discussion above, it is always possible to select a value of n (>3) such that V_{DD}^n becomes an upper bound on $(V_{\mathrm{DD}} - 2V_{\mathrm{TH}})^3$. For example, with V_{TH} =0.29-0.075V_{DD} (as in Section II-A), the term is bound by approximately $0.18V_{\mathrm{DD}}^6$. Thus in this case we have $I_{\mathrm{AVE}} \propto FV_{\mathrm{DD}}^6$ for V_{DD} >0.6V, and thus $P_{\mathrm{SS}} \propto FV_{\mathrm{DD}}^7$. Substituting $V \propto F \propto \mathrm{A}^{-1/\sigma}$ and multiplying by A to get total average current:

$$P_{\mathrm{SS}} \propto A^{(\sigma-8)/\sigma} \tag{12}$$

The $(\sigma - 8)$ term implies that short circuit power will continue to contribute only a very small fraction of the overall dynamic term as area increases. As V_{DD} approaches $2V_{\mathrm{TH}}$, the short circuit current rapidly tends to zero. It is extremely sensitive to V_{DD} and, as for switching power, can be easily traded off against area.

D. Gate Leakage Power

Gate leakage is projected to exceed sub-threshold leakage at the 65nm technology node [11] although there is recent evidence that problems have already arisen at 90nm [16]. As it is due to direct tunneling through the gate oxide it varies exponentially with oxide thickness and is extremely sensitive to gate voltage [17]. The current density (J_{FN}) at the transistor gate will have the general form of the Fowler-Nordheim tunneling equation:

$$J_{\mathrm{FN}} = C_0 E^2 e^{-Y/E} \tag{13}$$

where E $\approx \frac{\mathrm{V_G}}{\mathrm{T_{OX}}}$ is the surface electric field; $\mathrm{T_{OX}}$ = oxide thickness, and Y is a function of the effective barrier height between the oxide and the silicon surface. Assuming that C_0 is fixed for a given technology, the gate current will have the form:

$$I_{\mathrm{G}} \propto A \quad \frac{V_G}{T_{OX}} \quad ^2 \quad e^{\frac{Y T_{\mathrm{OX}}}{V_{\mathrm{G}}}} \tag{14}$$

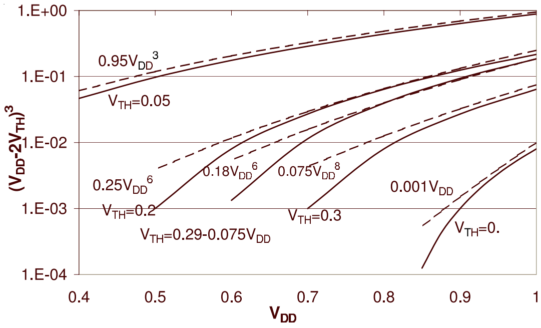

Fig. 3. $(V_{\mathrm{DD}} - 2V_{\mathrm{TH}})^3$ (for $0.4V \leq V_{\mathrm{DD}} \leq 1.0V$) for various fixed V_{TH} (solid lines) and $V_{\mathrm{TH}} = 0.29 - 0.075V_{\mathrm{DD}}$ (dashed line at center). In all cases, a value of n can be selected such that kV_{DD}^n (dotted lines in the diagram, k constant) represents an upper bound on $(V_{\mathrm{DD}} - 2V_{\mathrm{TH}})^3$.

The value of Y depends somewhat on the model of gate leakage: it varies from 1.43x10^8V/cm [18] to 1.9x10^8V/cm [19]. In either case, one can fit curves of the form aV_G^n for particular values of T_{OX} to the $\frac{1}{T_{OX}^2}e^{-\frac{YT_{OX}}{V_G}}$ term in (14). This first-order analytical result is confirmed by various simulations and experiments conducted by other researchers [20]–[22]. Using their data we obtain a minimum value of $n \geq 3$ for viable values of T_{OX}. As a result, one can conclude that as T_{OX} shrinks, gate leakage will become more of a constraint to lowering power by increasing area. However, in no case will it be the dominant factor. Substituting in V $\propto$ $\mathrm{A}^{-1/\sigma}$, we obtain: $\mathrm{I}_G \propto \mathrm{A}^{(\sigma-5)/\sigma}$. We thus conclude that the gate leakage component of power can be easily reduced with increased area for effective oxide thicknesses $\geq$ 0.2nm, so that:

$$P_G \propto A^{(\sigma-n)/\sigma}, n > 6 \tag{15}$$

E. Total Power

The total power is given by the sum of the contributions from dynamic switching, short-circuit, sub-threshold and gate leakage. Gate current and leakage current strongly interact and therefore the total leakage current (I_G + I_{OFF}) will be a complex function of transistor area, layout topology, interconnect stacking and the state of the system [8]. As a result, while the tradeoff against leakage power will be more complex than just a simple dependency on supply voltage, it will exhibit an $\mathrm{A}^{(\sigma-\chi)/\sigma}$ relationship with area, where χ will be no smaller than 3 and will probably lay between 3 and 4.

We observed in Section II-C the short circuit power term ($\propto$ $\mathrm{A}^{(\sigma-8)/\sigma}$) will tend to vanish so that the aggregated power will tend to be dominated by the $\mathrm{A}^{(\sigma-3)/\sigma}$ terms describing the limits of the dynamic and total leakage power. For values of $\sigma \leq 3$, the total power will be a constant or reducing function of area.

In summary, it can be seen that it is possible to arrange for overall power to be a reducing function of area with an impact on the maximum operating frequency that can be compensated for at the architectural level. The adjustment of V_{TH} with reducing V_{DD}, so that it becomes a greater fraction of $\mathrm{V_{DD}}$, might be achieved by various means such as gate work function engineering [23], body bias adjustment in bulk CMOS or back gate bias modulation in double-gate fully-depleted SOI CMOS transistors [24]. The caveat here is that V_{TH} will certainly become harder to control at smaller channel lengths, due to a combination of short channel effects and increased sensitivity to nanometer fluctuations in channel length [25].

In the next section we briefly speculate on the meaning of σ, and how a move to nanoscale technology might affect it.

III. Some Speculations on σ

As introduced in Section II, traditional measures of complexity for VLSI circuits across a range of computational functions such as multiplication, sorting and DFT (e.g. [4]) have the general form of Equation 1. These were extensively studied in the early days of VLSI (e.g. [2]) in order to establish bounds on the performance of computations that were to be distributed over the surface of a (planar) chip and to determine limits to the growth of the area-time metric with computation size. In this work, we are interested in the tradeoffs that can be made between area and time rather than just their growth with computation size. Thus we consider the case where $AT^{\sigma} = O(constant)$ which leads to: $\mathrm{A}^{-1} \propto T^{\sigma}$. Obviously, σ will not be a single value, but will exist within a range determined by a combination of design style and application. At the circuit level, different optimization techniques can be used to hit a particular area-performance point, while at the architectural level, it will be the ability to exploit parallelism (e.g., IPC) inherent in the computation that will determine the final area-time relationship.

The models used in the early studies of computational complexity assumed that signal propagation across on-chip interconnections could be achieved in linear time, with the area term adjusted up slightly to compensate. While this is certainly the case for circuits today, it is arguable that this will become more difficult in nanoscale CMOS for which interconnect costs will be high. For example, if the *width* of a silicon nanowire device is related to its diameter (a discrete value that will be difficult to control), it is not clear how the W/L ratio of an individual device could be made larger than unity - in order to directly drive a long interconnection line, for example. Such drivers might have to be formed from parallel groups of devices or else the nanoscale devices might be interfaced via conventional (micro-scale) CMOS [26]. A simpler alternative may be to constrain the layout style to mesh-style topologies with localized (or adjacent) connectivity only. In any case, area and delay overheads in nanoscale CMOS are likely to be significantly higher than at present.

IV. Conclusions

We have shown that area can be used to reduce power consumption in CMOS for a certain class of algorithms – those that are parallelizable. While this result is not overly surprising with respect to dynamic power, our analysis shows that it can also be made to hold true for static power. Further, we can characterize the algorithms for which this is true, those for which $\sigma \leq 3$ in $AT^{\sigma} = O(constant)$.

Our results suggest that attempts to increase drive current in nanoscale CMOS may be counter-productive. Where large numbers of devices are available, a better approach will be to exploit the extra devices to implement highly parallel versions of the algorithm (i.e. spatial computing) that will allow a reduction in operating frequency. The end result will be a reduction in power with little or no loss of performance. In future work we will be studying the types of spatial micro-architectures that would be best suited to nanoscale CMOS and how these might exploit parallelism at multiple levels (e.g. instruction level, multiple-threads etc.) in order to achieve the sort of tradeoffs between power and area and performance that we have determined are possible. Also important to study is how the expected increase in process variability will affect the area–power tradeoff.

References

[1] S. C. Goldstein and M. Budiu, "Nanofabrics: Spatial computing using molecular electronics," in *Proc. 28th Int'l Symposium on Computer Architecture*, Goteborg, Sweden, 2001, pp. 178–189.

[2] C. D. Thompson, "Area-time complexity for vlsi," in *Proc. Eleventh Annual ACM Symposium on Theory of Computing*, 1979, pp. 81–88.

[3] H. Abelson and P. Andreae, "Information transfer and area-time tradeoffs for vlsi multiplication," *Communications of the ACM*, vol. 23, no. 1, pp. 20–23, 1980.

[4] R. J. Lipton and R. Sedgewick, "Lower bounds for vlsi," in *Proc. Thirteenth Annual ACM Symposium on Theory of Computing*, 1981, pp. 300–307.

[5] R. P. Brent and H. T. Kung, "The chip complexity of binary arithmetic," in *Proc. Twelfth Annual ACM Symposium on Theory of Computing*, Los Angeles, CA, USA, 1980, pp. 190–200.

[6] ——, "The area-time complexity of binary multiplication," *Journal of the ACM*, vol. 28, no. 3, pp. 521–534, 1981.

[7] I. E. Sutherland, "Micropipelines," *Communications of the ACM*, vol. 32, no. 6, pp. 720–738, 1989.

[8] D. Lee, W. Kwong, D. Blaauw, and D. Sylvester, "Analysis and minimization techniques for total leakage considering gate oxide leakage," in *Proc. 40th Conference on Design Automation*, 2003, pp. 175–180.

[9] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device scaling limits of si mosfets and their application dependencies," *Procs. of the IEEE*, vol. 89, no. 3, pp. 259–288, 2001.

[10] P. K. Ko, J. Huang, Z. Liu, and C. Hu, "Bsim3 for analog and digital circuit simulation," in *Proc. IEEE Symposium on VLSI Technology CAD*, 1993, pp. 400–429.

[11] SIA, *Int'l Technology Roadmap for Semiconductors-2003 update*, 2003.

[12] K. Chen, C. Hu, P. Fang, M. R. Lin, and D. L. Wollesen, "Predicting cmos speed with gate oxide and voltage scaling and interconnect loading effects," *IEEE Trans. on Electron Devices*, vol. 44, no. 11, pp. 1951–1957, 1997.

[13] M. J. Flynn, P. Hung, and K. W. Rudd, "Deep-submicron microprocessor design issues," *IEEE Micro*, vol. 19, pp. 11–22, 1999.

[14] P. Zarkesh-Ha, J. A. Davis, W. Loh, and J. D. Meindl, "Prediction of interconnect fan-out distribution using rent's rule," in *Proc. Int'l Workshop on System-level Interconnect Prediction (SLIP00)*, 2000, pp. 107–112.

[15] H. J. M. Veendrick, "Short-circuit dissipation of static cmos circuitry and its impact on the design of buffer circuits," *IEEE Journal of Solid-State Circuits*, vol. 19, no. 4, pp. 468–473, 1984.

[16] R. Ball, "Smaller processes lead to fpga leakage crisis," *Electronics Weekly*, vol. September 30, 2003. [Online]. Available: http://www.reed-electronics.com/electronicnews/article/CA332188

[17] C. K. Huang and N. Goldsman, "Modeling the limits of gate oxide scaling with a schrodinger-based method of direct tunneling gate currents of nanoscale mosfets," in *Proc. 1st IEEE Conference on Nanotechnology.* IEEE-NANO 2001, 2001, pp. 335–339.

[18] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deepsubmicrometer cmos circuits," *Proc. of the IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[19] S. Keeney, R. Bez, D. Cantarelli, F. Piccinini, A. Mathewson, L. Ravazzi, and C. Lombardi, "Complete transient simulation of flash eeprom devices," *IEEE Trans. on Electron Devices*, vol. 39, no. 12, pp. 2750–2757, 1992.

[20] Y. C. Yeo, Q. Lu, W. C. Lee, T. J. King, C. Hu, X. Wang, X. Guo, and T. P. Ma, "Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric," *IEEE Electron Device Letters*, vol. 21, no. 11, pp. 540–542, 2000.

[21] H.-S. P. Wong, D. J. Frank, P. M. Solomon, C. H. J. Wann, and J. J. Welser, "Nanoscale cmos," *Proc. of the IEEE*, vol. 87, no. 4, pp. 537–70, 1999.

[22] Y. C. Yeo, T. J. King, and C. Hu, "Direct tunneling leakage current and scalability of alternative gate dielectrics," *Applied Physics Letters*, vol. 81, no. 11, pp. 2091–2093, 2002.

[23] J. Kedzierski, E. Nowak, T. Kanarsky, Y. Zhang, D. Boyd, R. Carruthers, C. Cabral, R. Amos, C. Lavoie, R. Roy, J. Newbury, E. Sullivan, J. Benedict, P. Saunders, K. Wong, D. Canaperi, M. Krishnan, K.-L. Lee, B. A. Rainey, D. Fried, P. Cottrell, H.-S. P. Wong, M. Ieong, and W. Haensch, "Metal-gate finfet and fully-depleted soi devices using total gate silicidation," in *Proc. Int'l Electron Devices Meeting, IEDM '02*, 2002, pp. 247–250.

[24] P. Beckett, "Exploiting multiple functionality for nano-scale reconfigurable systems," in *Proc. 10th Reconfigurable Architectures Workshop, RAW2003*, 2003, pp. 50–55.

[25] Y. Naveh and K. K. Likharev, "Modeling of 10-nm scale ballistic mosfet's," *IEEE Electron Device Letters*, vol. 21, no. 5, pp. 242–244, 2000.

[26] M. M. Ziegler and M. R. Stan, "A case for cmos/nano co-design," in *Proc. Int'l Symposium on Circuits and Systems, ISCAS'02*, 2002, pp. 348–352.