POLITECNICO DI TORINO
Repository ISTITUZIONALE

An Electromigration and Thermal Model of Power Wires for a Priori High-Level Reliability Prediction

(Article begins on next page)

10 May 2021

An Electromigration and Thermal Model of Power Wires for *a Priori* High-Level Reliability Prediction

Mario R. Casu, Mariagrazia Graziano, Guido Masera, *Member, IEEE*, Gianluca Piccinini, and Maurizio Zamboni

Abstract—In this paper, a simple power-distribution electrothermal model including the interconnect self-heating is used together with a statistical model of average and rms currents of functional blocks and a high-level model of fanout distribution and interconnect wirelength. Following the 2001 SIA roadmap projections, we are able to predict *a priori* that the minimum width that satisfies the electromigration constraints does not scale like the minimum metal pitch in future technology nodes. As a consequence, the percentage of chip area covered by power lines is expected to increase at the expense of wiring resources unless proper countermeasures are taken. Some possible solutions are proposed in the paper.

Index Terms—*Apriori* power supply interconnect prediction, electromigration (EM) estimation, self-consistent roadmap projections, statistical current model.

I. INTRODUCTION

IN RECENT years, many system simulators such as SUSPENS [1], RIPE [2], BACPAC [3], GENESYS [4], and GTX [5], have been used in order to make *a priori* predictions of system-level parameters like clock frequency, dynamic and static-power dissipation, power supply and crosstalk noise, wiring resources, die size and yield, for very large-scale and giga-scale integrated systems (VLSI/GSI) in ultradeep submicron (UDSM) technologies. A key point that has not been included yet in system-level prediction models, but already claimed as one of the major issues in UDSM technologies, is the reliability of power supply lines subject to the electromigration (EM) risk. This phenomenon, universally known by the integrated circuits designers, is getting more and more important as CMOS technology improves [6]. Not only power wires, but also signal lines suffer from EM, especially when long interconnects and fast signal transitions are involved, as it happens in clock networks [7]. Recently, this issue has been addressed for a chip-level static analysis of signal EM [8].

The reference for the evaluation of the EM is Black's law [9] where the mean time-to-failure of power supply interconnects is inversely proportional to the square of the average current density J_{avg}^2. The reliability also exponentially depends on the actual metal temperature, which is higher than in the silicon substrate; the reason is the self-heating of interconnects due to the finite electrical conductivity of metal alloys and the finite thermal conductivity of intralayer dielectrics. The self-heating phenomenon is primarily determined by the root-mean-square (rms) current. There have been papers aiming at determining the maximum allowed average and rms currents that results in a given reliability measure (e.g., ten years mean time-to-failure) [6], [10], [11]. Other computer-aided design (CAD) works provide the designer with tools able to evaluate the reliability of specific designs where the current activities are known at the gate or transistor level [12], [13]. However, there are not models that link the interconnect current limits, as presented in the SIA International Roadmap for Semiconductors (ITRS) document [14], with a high-level model of self-heating effects in power-supply networks of functional blocks in VLSI/GSI systems and with a high-level model of gate average and rms currents. In this paper, we show how high-level *a priori* models of power-distribution networks and gate currents, in connection with a fanout distribution and a wirelength model, are able to predict a potential "crisis" in future systems for which the minimum width that respects the reliability requirements does not scale with technology nodes.

In Section II, we propose a simple electrothermal model useful for the evaluation of the self-heating in power supply interconnects. In Section III, a statistical system-level model of average and rms currents is developed and the self-heating of interconnects is evaluated in Section IV. Then, Section V shows the above mentioned nonscalability under different conditions. Some possible countermeasures are proposed in Section VI. Finally, we summarize the main results in Section VII.

II. POWER DISTRIBUTION MODEL

In this paper, we consider two possible power distribution models. The first one is the classic interdigitated structure where the gates are connected between power and ground fingers; the second one is a grid-based mesh, which is more common in today's designs and is superposed to the gate area. We assume that the mesh is contacted to two equipotential outer rings of V_{dd} and gnd while the interdigitated structure is connected on one side to V_{dd} and to gnd on the other side. N^2 logic gates are connected to both structures. For the sake of simplicity, we will consider a square layout where each line spans N gates in both models and one gate is connected to each mesh contact point. This simplification allows the use of analytical models. We will limit our analysis to a so called "homogeneous" functional block (also called megacell in some works [15]) of maximum 50–100 kgates. These kind of blocks are used to build complex systems-on-chip (SoC) that are "heterogeneous" because they include several functional blocks. Homogeneous blocks have two
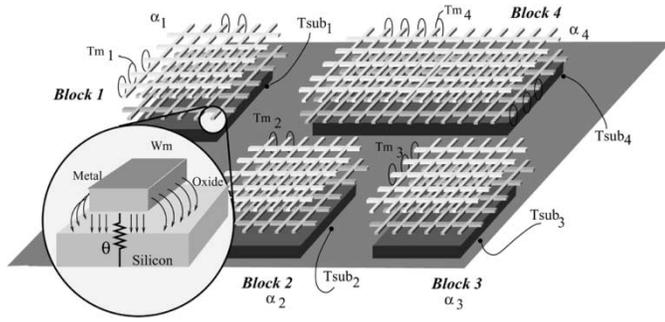
Fig. 1. Functional "homogeneous" blocks with power/gnd grid in a complex "heterogeneous" SoC.

fundamental properties. First, their statistical properties fit well with a Rent's Rule representation [16]; and second, the signal interconnect system within the block is scalable because of the use of scalable local interconnects [3], [17], [18]. In Fig. 1, a floorplan of a SoC is sketched, where many blocks have been placed and the Vdd-gnd grid has been routed over each. 12tvlsi04-casu

The branches of the power structure are characterized by a resistance per square R_{sq} whose value is computed considering the metal resistivity ρ_0, measured at temperature T_0, the metal thickness t_m and the actual metal temperature T_m

$$R_{sq} = \frac{\rho_0}{t_m}[1 + \alpha_0(T_m - T_0)]. \quad (1)$$

The values of the parameters are self-consistent with the 2001 SIA roadmap [14], the extract is reported in Table I, except for the first-order coefficient α_0 which is not mentioned in the SIA document; ρ_0 and the product $\rho_0\alpha_0$ are known being, respectively, 1.68×10^{-8} Ωm and 0.68×10^{-8} $\Omega \cdot$ m/°C for bulk Copper at 20 °C. However, the effective resistivity is higher for the effect of the Cu barrier and is about 2.2×10^{-8} $\Omega \cdot$ m [14]. The temperature dependency changes as well and we derived a higher value of 0.89×10^{-8} $\Omega \cdot$ m/°C from [19]. The actual metal temperature is used for the correct evaluation of the interconnect EM reliability. It can be computed by using a simplified thermal model [10], [11], [20]. The temperature of the metal line of width w_m, length l_m, and thickness t_m over a dielectric of thickness t_{ox} and thermal conductivity k_{ox} is given by

$$T_m = T_{sub} + \Theta_{th}P_m = T_{sub} + \frac{t_{ox}}{k_{ox}l_m(w_m + \phi t_{ox})}P_m \quad (2)$$

where T_{sub} is the substrate temperature at the functional block level (assumed uniform within the block), Θ_{th} is the equivalent thermal resistance due to the finite thermal conductivity of the interlayer dielectric, and P_m is the average power dissipated on the interconnect electrical resistance. ϕ is a coefficient taking into account the heat dissipated from the metal edges (see sketched zoom in Fig. 1). This equation describes the self-heating of power interconnects whose temperature rises above T_{sub}. Equation (2) implicitly assumes a one-dimensional (1-D) heat flow, which is true along the line a few heat diffusion lengths $\lambda_m = \sqrt{t_m t_{ox} k_m / k_{ox}}$ from the vias or contacts that act like heat sinks, and are supposed to be nearly at substrate temperature [10] (k_m is the thermal conductivity of the metal). Considering the roadmap parameters t_m and t_{ox}, the thermal conductivity of low-k materials [11] and of copper, we obtain

values for λ_m on the same order of a gate pitch. This means that 1-2 gate pitches away from the block periphery, the temperature of the interconnects will be given by (2). In any case, the equation represents a worst case for the power interconnects within the block, because the effect of neighbor vias will reduce the temperature. We also neglect the presence of inferior metal layers under the metal considered because they favor the heat dissipation and further reduce the interconnect self-heating.

P_m is given by $R_{sq}l_m/w_m I_{rms}^2$ where I_{rms}^2 is the root mean square current flowing in the line element. By using the explicit R_{sq} definition (1) that contains a linear dependence on T_m, the final form for (2) is

$$T_m = \frac{T_{sub} + \frac{t_{ox}\rho_0(1-\alpha T_0)I_{rms}^2}{k_{ox}w_m t_m(w_m + \phi t_{ox})}}{1 - \frac{t_{ox}\rho_0\alpha I_{rms}^2}{k_{ox}w_m t_m(w_m + \phi t_{ox})}}. \quad (3)$$

T_m is used to compute the limit on maximum current density that is allowed to flow in interconnects. The roadmap document defines a J_{ref} limit for each technology node evaluated at $T_{ref} = 105$ °C as the one leading to a reliability measure of 5FIT (5 failures in 10^9 hours device). However, not only T_{sub} is foreseen to overcome such limit (125 °C junction temperature for low-cost high-volume products, see Table I) but also the self-heating could make T_m higher than T_{sub}. Therefore, it is mandatory to reconsider this limit in a self-consistent manner, as already shown in [10] and [11]. Using Black's law [9], we obtain the following limit on the average current

$$J_{avg} < J_{ref}e^{(Q/2K_b)((1/T_m)-(1/T_{ref}))}. \quad (4)$$

The free parameter in this formula is the width of the line w_m that appears on both members of the inequality because $J_{avg} = I_{avg}/(w_m t_m)$ and T_m depends on w_m as clear from (3). The final task of this work consists in the evaluation of the linewidth necessary to comply with the roadmap expectations and evaluate weather it scales as the minimum metal pitch does or not. In Section III, both average I_{avg} and root mean square I_{rms} current models that are needed to evaluate the EM limits are computed.

III. AVERAGE AND rms CURRENT MODEL

Each gate current is modeled with a triangular waveform as in many other works. We will only consider the V_{dd} part of the power distribution, therefore this current is drawn from V_{dd} during low–high transitions. The area of the triangle is simply the total charge drawn during the transition of the output gate and is $Q = CV_{dd}$. C is the total output load, including output capacitance of the gate itself, input capacitance of the fanout gates and interconnect capacitance. We will neglect in the following the I_{sc} short-circuit current and the I_{off} current (subthreshold plus reverse-biased junctions) contributions. We have verified using the BACPAC tool that I_{sc} scales with technology nodes and is always about 5% of the total current. I_{off} does not scale and increases progressively. Its impact is getting more and more important especially for what concerns the power budget [14]. However, we have verified that its effect on EM is still negligible compared to the dynamic current.

The wiring capacitance is estimated assuming a wirelength distribution like in the original Donath's work [21] that has been

| Year of Production	2001	2002	2003	2004	2005	2006	2007
DRAM 1/2 pitch (nm)	130	115	100	90	80	70	65
MPU Printed Gate Length (nm)	90	75	65	53	45	40	35
MPU Physical Gate Length (nm)	65	53	45	37	32	28	25
T_{ox} Electrical Equivalent (nm)	2.3	2.1	2.0	2.0	1.9	1.9	1.4
Nominal Power Supply Voltage (*Vdd*) (V)	1.2	1.1	1.0	1.0	0.9	0.9	0.7
Number of Metal Levels	8	8	8	9	10	10	10
J_{ref} (A/cm^2)-wire at 105°C	9.6E5	1.1E6	1.3E6	1.5E6	1.7E6	1.9E6	2.1E6
Local Wiring Pitch (nm)	350	295	245	210	185	170	150
Local Wiring Aspect Ratio (for Cu)	1.6	1.6	1.6	1.7	1.7	1.7	1.7
Intermediate Wiring Pitch (nm)	450	380	320	265	240	215	195
Intermediate wiring Aspect Ratio (Cu wire/via)	1.6/1.4	1.6/1.4	1.7/1.5	1.7/1.5	1.7/1.5	1.7/1.6	1.8/1.6
Global Wiring Pitch (nm)	670	565	475	460	360	320	290
Global Wiring Aspect Ratio (Cu wire/via)	2.0/1.8	2.0/1.8	2.1/1.9	2.1/1.9	2.2/2.0	2.2/2.0	2.2/2.0
ILD effective dielectric constant (k)	3.0-3.6	3.0-3.6	3.0-3.6	2.6-3.1	2.6-3.1	2.6-3.1	2.3-2.7
Dielectric Thermal Conductivity (W/°C m)	0.6	0.6	0.6	0.25	0.25	0.25	0.25
Frequency (GHz) for High-Performance Devices	1.7	2.32	3.09	3.99	5.17	5.63	6.74
Low-Cost Devices Max Junction Temperature (° C)	125	125	125	125	125	125	125
High-Perf. Devices Max Junction Temperature (° C)	90	85	85	85	85	85	85
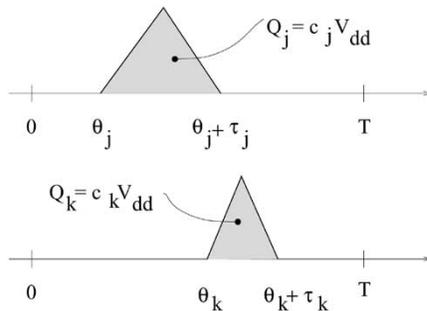
Fig. 2.   Triangular models of current waveforms.

integrated in the BACPAC tool [3], [18]. There are many other valid and more accurate works on wirelength prediction such as the model by Davis [22]. Our choice was justified by reasons of simplicity and because another more accurate model will not have an impact on the generality of our results. The focus of this paper is not on the gate load model, therefore, one could switch to a better one while retaining the philosophy of this work.

The height of the triangle is $I_{\max}$ and the base is then $\tau = 2Q/I_{\max}$. The waveform is delayed in the clock period $[0, T]$ of Θ which is a stochastic variable representing the starting time point of the gate output transition. Fig. 2 represents two possible waveforms $i_j(t)$ and $i_k(t)$ delayed of Θ_j and Θ_k and width τ_j and τ_k, respectively.

All the parameters are functions of the gate fanout and of the technology node considered. We assume that the gate sizing has been done according to the optimization rule used in BACPAC. As in [3] we will consider a two-input NAND as the most representative average gate.

We will distinguish between interdigitated structures and more complex meshes. In the first case, we will consider the current flowing in a single line and in the second one the maximum current in each resistive element of the grid.

A. Line of an Interdigitated Structure: Average Current

Let us consider a single line connected to N gates. If we are considering a scalable homogeneous block containing about 50

kgates according to [17], N will be about $\sqrt{50\,000} \approx 224$. The maximum current $i(t)$ drawn from V_{dd} and flowing at the end of the line is the sum of all N gate currents $i_j(t - \Theta_j)$ where Θ_j is the delay associated with the jth gate low-high switch, $j = 1, \ldots, N$. The dc component, needed for the EM estimation, of the current drawn from V_{dd} in a clock cycle $T = 1/f$ will be given by

$$\overline{I} = \sum_{j=1}^{N} \frac{1}{T} \int_0^T i_j(t - \Theta_j)dt = \sum_{j=1}^{N} \frac{1}{T} \int_0^T i_j(t)dt \quad (5)$$

where Θ_j does not appear because the delay information is cancelled by the integration. This is the definition of the average charge drawn from V_{dd} and, therefore, we can write

$$\overline{I} = \sum_{j=1}^{N} fC_j V_{dd}. \quad (6)$$

We still must introduce the statistical property of C_j and compute the average on fanout and switching probability

$$\overline{C_j} = \sum_{fo} \alpha_j(fo)C_j(fo)\Phi_j(fo) \quad (7)$$

where Φ_j is the fanout distribution and α_j is the switching probability of gate j. If we assume that gates are uniformly distributed along the line, the average capacitance does not depend on position j. The average dc value of the current will be then given by

$$I_{\text{avg}} = \alpha N f V_{dd} \sum_{fo} C(fo)\Phi(fo) = \alpha N f V_{dd}\overline{C}. \quad (8)$$

B. Element of a Mesh: Average Current

We will follow the cutting boundary method suggested by Zarkesh-Ha in [23] for the computation of IR drop. If we have N^2 gates uniformly distributed along a mesh the average branch current is given by $1/4\, Ni_g$ where i_g is the average gate current

which can be computed using (8) as I_{avg}/N. Therefore, the result is the same as in (8) multiplied by 1/4.

C. Line of an Interdigitated Structure: RMS Current

The mean square current at the end of the line is

$$
i^2 = \sum_{j=1}^{N} \frac{1}{T} \int_0^T i_j^2(t - \Theta_j) dt
$$

$$
+ 2 \sum_{j=1}^{N} \sum_{k=j+1}^{N} \frac{1}{T} \int_0^T i_j(t - \Theta_j) i_k(t - \Theta_k) dt. \quad (9)
$$

The first term represents the self-contribution of all square gate currents, while the second one is the sum of the cross contributions. In the first term, the delay Θ_j is cancelled by the integration. Averaging with respect to the fanout and integrating the square of the triangle, the first term is

$$
i_{\mathrm{self}}^2 = \alpha N f \sum_{fo} \int_0^T i^2(t, fo) dt \Phi(fo)
$$

$$
= \frac{2}{3} \alpha N f V_{dd} \sum_{fo} C(fo) I_{\max}(fo) \Phi(fo)
$$

$$
= \frac{2}{3} \alpha N f V_{dd} \overline{CI_{\max}} \quad (10)
$$

where $\overline{CI_{\max}}$ is the average over all fo values using the distribution $\Phi(fo)$. The second term in (9), is more difficult to compute since we have to take into account the waveform delays Θ_j and Θ_k and the correlation between waveforms. However, the insertion of the correlation would require an accurate knowledge of the circuit under exam. Moreover, further improvements of the model are not justified in light of the approximation made when considering an equivalent gate, a uniform distribution of gates along the grid and an average switching probability. Therefore, we will consider all gates as noncorrelated. We will first apply the average operator on noncorrelated delays Θ_j and Θ_k with distribution $f_{\Theta_j}(\theta_j)$ and $f_{\Theta_k}(\theta_k)$. We have to compute the triple integral

$$
\iiint i_j(t - \theta_j) i_k(t - \theta_k) f_{\Theta_j}(\theta_j) f_{\Theta_k}(\theta_k) dt d\theta_j d\theta_k \quad (11)
$$

where the integration domain is $[0, T]$ for all variables and integrating first in time t we get

$$
\iint R_{jk}(\theta_j - \theta_k) f_{\Theta_j}(\theta_j) f_{\Theta_k}(\theta_k) d\theta_j d\theta_k
$$

$$
= \int R_{jk}(\tau) F_{jk}(\tau) d\tau \quad (12)
$$

where $R_{jk}(\tau)$ is the mutual correlation of $i_j(t)$ and $i_k(t)$ and $F_{jk}(\tau)$ is the mutual correlation of $f_{\Theta_j}(\theta_j)$ and $f_{\Theta_k}(\theta_k)$. For the sake of simplicity, we will consider the delays uniformly distributed in $[0, T]$. Therefore, $F_{jk}(\tau)$ is a triangle centered in $\tau = 0$ with domain $[-T, T]$ and height $1/T$. The domain of $R_{jk}(\tau)$ is $[-\tau_k, \tau_j]$ where τ_k and τ_j are the two triangle bases. An example is given in Fig. 3. An upper bound of (12) is obtained by considering $F_{jk}(\tau)$ constant and equal to $1/T$ in
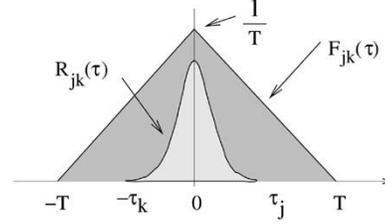

Fig. 3. R_{jk} and F_{jk} are the mutual correlations of currents i_j and i_k and of their delay distributions $f_{\Theta_j}(\theta_j)$ and $f_{\Theta_k}(\theta_k)$.

TABLE II
PROJECTION OF AVG AND RMS CURRENT (MILLIAMPS) IN LINES (L) AND MESHES (M)

Year	2001	2002	2003	2004	2005	2006	2007
I_{avg} l	1.13	1.19	1.18	1.16	1.16	1.11	0.86
I_{avg} m	0.28	0.30	0.30	0.29	0.29	0.28	0.21
I_{rms} l	11.6	10.9	10.1	9.15	8.54	7.94	6.31
I_{rms} m	2.90	2.72	2.52	2.29	2.13	1.99	1.58

$[-\tau_k, \tau_j]$. The accuracy improves as long as $\tau_k, \tau_j \ll T$. This upper bound is very simple to compute and the result of (12) is

$$
\int \frac{R_{jk}(\tau)}{T} d\tau = f \int i_j(t) dt \int i_k(t) dt = f C_j C_k V_{dd}^2. \quad (13)
$$

Finally, the average on fanout fo_j and fo_k and switching probability results in $f(\alpha \overline{C} V_{dd})^2$ and reinserting this expression in the second contribution in (9) we get

$$
i_{\mathrm{cross}}^2 = 2 \frac{N(N-1)}{2} (\alpha \overline{C} f V_{dd})^2 \approx (\alpha \overline{C} f V_{dd} N)^2. \quad (14)
$$

I_{rms} will then be given by

$$
I_{\mathrm{rms}} = \sqrt{\frac{2}{3} \alpha N f V_{dd} \overline{CI_{\max}} + (\alpha \overline{C} f V_{dd} N)^2}. \quad (15)
$$

D. Element of a Mesh: rms Current

Following the same approach used for the computation of the average current, the only difference in the case of the mesh is a term $(1/4)^2$ that multiplies both i_{self}^2 and i_{cross}^2.

Now we proceed to calculate the average and rms values for the case of a functional block of around 50 kgates as a function of the roadmap year of production, in the near-term 2001–2007. The fanout distribution and the capacitance calculations are derived from [3]. Most of the input parameters of all equations are derived from the SIA roadmap and are summarized in Table I for the high performance logic devices. The thermal conductivities of low-k materials are taken from [11]. All the equations have been implemented in the GTX framework [5], [24]. The computed currents are reported in Table II.

We now have at our disposal the power distribution and current models. We will use them in a self-consistent way with the roadmap expectations in Sections IV–VI.

IV. SELF-HEATING IN POWER BUS

We begin by examining the maximum temperature in the power bus of an interdigitated structure. As a reference, we will consider the maximum frequency for high performance devices
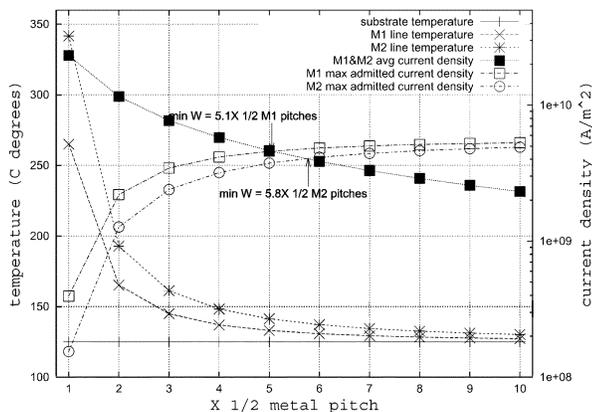
Fig. 4. Left axis: temperature in M1 and M2 interdigitated lines; right axis: average and maximum admitted current densities.

Fig. 5. Left axis: temperature in M4 and M5 grid lines; right axis: average and maximum admitted current densities.

according to Table I. We chose an average activity factor α of 0.15 according to [18]. We will see in Sections V and VI the effect of changing this value. Usually the cells of a functional block are connected to metal1 or metal2 power bus interdigitated fingers. Using (3) and data in Tables I and II we obtain the bus temperature assuming a substrate temperature of 125 °C and as a function of the bus width. The results are reported in Fig. 4 for the 2001 130-nm technology node where the bus width is expressed in multiples of half metal pitches. On the left axis substrate and metal temperatures are reported. As expected, the line temperature T_m approaches T_{sub} as the metal width increases. On the right axis the average current densities and maximum admitted current densities are reported, i.e., left and right members in (4). We can observe that 5.1 and 5.8 times the half metal1 and metal2 pitches respectively are necessary to comply with the reliability constraint. The current densities that satisfy the constraint are significantly lower than J_{ref} from Table I, 2001 technology node, that is $9.6 \cdot 10^9$ A/m^2, because the metal temperature is higher than $T_{\text{ref}} = 105\,°\text{C}$ ($\approx 135\,°\text{C}$ in both metal1 and metal2 at 5.1 and 5.8 half metal pitches, respectively).

If we have a mesh grid of power supply instead of interdigitated fingers, the most appropriate metal layers are 4 and 5 because the lower ones are normally left for local and intermediate routing. Compared to the interdigitated structure, we have a reduction of both average current (1/4 reduction due to the mesh structure as explained before) and current densities (due to the bigger pitch and aspect ratio of such layers compared to the local ones). On the other hand the bigger distance from the substrate of such layers results in a higher thermal resistance. The results for the 130-nm 2001 technology node are reported in Fig. 5 in the same conditions as in Fig. 4. We can notice that the minimum size line is enough ($1 \times 1/2$ pitch), there is not appreciable difference between M4 and M5 behavior, and self-heating is only 6 °C. These facts are explained by the 1/4 reduction in average current and root mean square currents with respect to the interdigitated case that has the twofold effect of reducing the average current density and the self heating.

We have seen that there is a minimum linewidth that can satisfy the constraint in (4). It is then plausible to ask how this minimum linewidth scales with technology nodes. The best will be
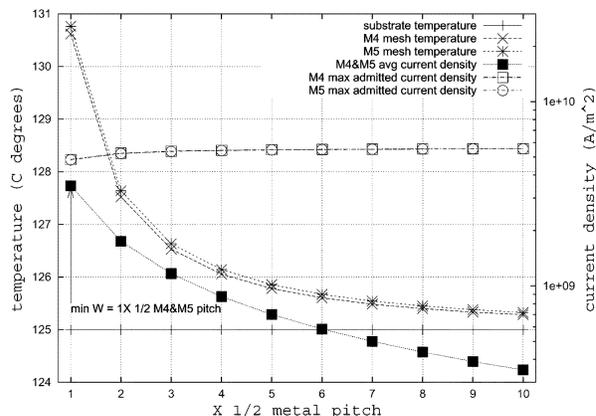
TABLE III
PROJECTION OF THE MINIMUM WIDTH FOR RELIABILITY EXPRESSED IN HALF
METAL PITCHES: $\alpha = 0.15$, $T_{\text{sub}} = 125\,°\text{C}$

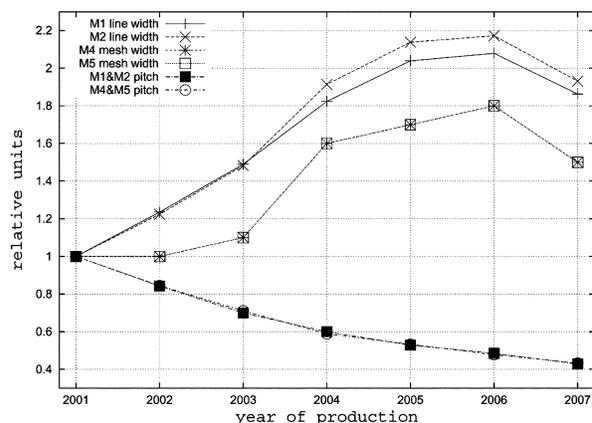| Year	2001	2002	2003	2004	2005	2006	2007
M1 line	5.1	6.3	7.6	9.3	10.4	10.6	9.5
M2 line	5.8	7.1	8.6	11.1	12.4	12.6	11.2
M4 mesh	1.0	1.0	1.1	1.6	1.7	1.8	1.5
M5 mesh	1.0	1.0	1.1	1.6	1.7	1.8	1.5

Fig. 6. Projection of the minimum width for reliability (in half-metal pitches) and metal pitches normalized to the 2001 value.

a scaling equal to or lower than the minimum metal pitch factor. Otherwise a progressively increasing area overhead in interconnects will have to be taken into account unless proper countermeasures are taken.

V. NONSCALABILITY OF POWER BUS LINE WIDTH

We have solved for the minimum width that satisfies the constraint in (4). In Table III, the minimum widths are reported in terms of half metal pitches for all technology nodes. The same values are reported in Fig. 6, normalized to the 2001 technology node. Both in the line and the mesh cases there is a progressive increase of the normalized minimum width for reliability, while we would have needed a constant value (it scales as the metal pitch) or a decreasing value (it scales more than the metal pitch).
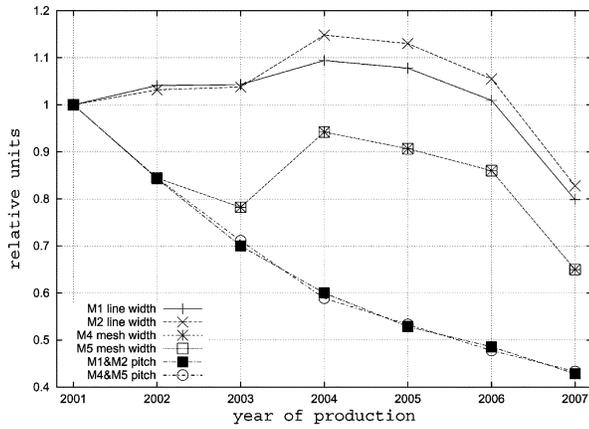
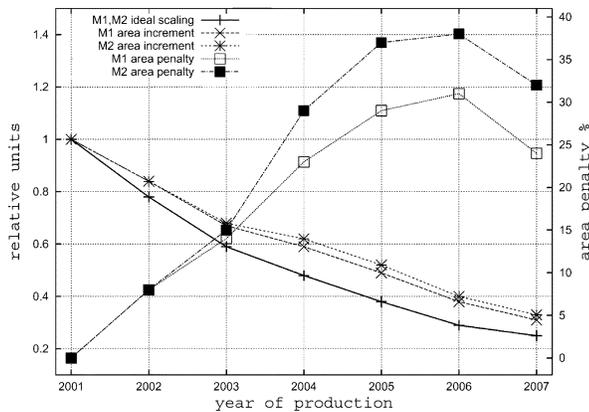Fig. 7. Projection of minimum width normalized to the 2001 value.

Fig. 8. Projection of area (left axis) in M1-M2 interdigitated fingers normalized to the 2001 value and area penalty (right axis).

In Fig. 6, the scaling of metal pitches is also reported. We can notice a strong increase in 2004 due to the increase of the thermal conductivity of the interlayer dielectric, which is due to the reduction of the dielectric constant (see Table I). The decrease in 2007 is due to the expected decrease of V_{dd} (from 0.9 to 0.7 V, i.e., -22%).

In Fig. 7 the true width is reported again not in terms of pitch but normalized to the 2001 case. The width remains relatively constant with the technology nodes. We can conclude that *"the minimum width that satisfies the EM reliability constraint does not scale with technology nodes"*.

One of the consequences of this crisis is that the percentage of functional block area covered by the power supply distribution system will increase with technology nodes and reduce the wiring resources. In the case of interdigitated M1 and M2 wiring the increment in area can be easily estimated because the cells need to be spaced in the vertical direction of the same amount of wire width increase. Assuming as in [25] that the cell aspect ratio is 16/4 metal pitches (which is coherent with the area factors for chip area estimation in the roadmap document), we calculated the increase in area. The results are reported in Fig. 8 where the actual area normalized to the 2001 year is reported and compared to the ideally scaled area. The area penalty (%) due to the extra spacing overhead is also plotted on the right axis.

TABLE IV
PROJECTION OF THE MINIMUM WIDTH FOR RELIABILITY EXPRESSED IN HALF
METAL PITCHES:$\alpha = 0.5$, $T_{\mathrm{sub}} = 125\,^\circ\mathrm{C}$

Year	2001	2002	2003	2004	2005	2006	2007
M1 line	15.1	19.1	23.2	26.8	30.1	30.6	27.4
M2 line	16.6	20.8	25.0	30.5	34.0	34.6	31.0
M4 mesh	2.5	3.2	3.5	5.0	5.3	5.7	4.5
M5 mesh	2.5	3.2	3.6	5.1	5.4	5.8	4.6

TABLE V
PROJECTION OF THE MINIMUM WIDTH FOR RELIABILITY EXPRESSED IN HALF
METAL PITCHES:$\alpha = 0.15$, $T_{\mathrm{sub}} = 85\,^\circ\mathrm{C}$

Year	2001	2002	2003	2004	2005	2006	2007
M1 line	2.6	3.2	3.8	5.3	5.8	5.9	5.3
M2 line	3.2	3.9	4.7	6.9	7.6	7.7	6.8
M4 mesh	1.0	1.0	1.0	1.0	1.0	1.0	1.0
M5 mesh	1.0	1.0	1.0	1.0	1.0	1.0	1.0

A. Effect of α and T_{sub}

The results reported in previous figures and tables have been obtained with an average value of $\alpha = 0.15$ and $T_{\mathrm{sub}} = 125\,^\circ\mathrm{C}$. It is then of interest to understand what happens when we consider different switching activities and substrate temperatures as for the various building blocks of a heterogeneous SoC like the one sketched in Fig. 1. For instance, if we set $\alpha = 0.5$ which is the maximum value for static logic gates, while keeping the same substrate temperature, we can expect a worsening in the minimum width. On the other hand $\alpha = 0.15$ and $T_{\mathrm{sub}} = 85\,^\circ\mathrm{C}$ represents an optimistic case. In Tables IV and V are reported the results for these two cases. As expected, a higher switching activity at the same high substrate temperature results in wider power busses. Lower values of α and T_{sub} imply narrower lines. However the trend in Tables IV and V is the same of Table III and of Fig. 6 if we normalize values to the 2001 technology node. We can observe that for the case of M4 and M5 meshes at $\alpha = 0.15$, $T_{\mathrm{sub}} = 85\,^\circ\mathrm{C}$ the inferior limit is the 1.0 half metal pitch (lithographic limit) meaning that the minimum width for EM would have been lower than 1.0. Nevertheless its trend is the same being lower than 1 but increasing at the same pace of M4 and M5 for higher values of α and T_{sub}. In the roadmap document, the recommended temperature for high-performance logic devices is 85°C for all technology nodes. Therefore one could argue that with the results in Table V the scaling is ensured for M4–M5 meshes. However the activity ratio can be higher than $\alpha = 0.15$ and approach $\alpha = 1$ for dynamic logic gates that are frequently used in high-speed datapaths of high-performance integrated circuits. Moreover the mesh we used for the computation is very dense while rarer meshes are normally used. If we reconsider all these variables, the nonscalability of power meshes is again evident as shown later in Section V-D and as confirmed by the huge amount of data we obtained and that we cannot report here for sake of brevity.

B. Neglecting the Self-Heating

It is interesting to verify how strong is the impact of self-heating in determining the minimum width. If we disable the computation of T_m and assume $T_m = T_{\mathrm{sub}}$ we get the values in Table VI to be compared with Table III. We can notice that M1 and M2 values are equal as well as M4 and M5

TABLE VI
PROJECTION OF THE MINIMUM WIDTH FOR RELIABILITY EXPRESSED IN HALF METAL PITCHES: $\alpha = 0.5$, $T_{\text{sub}} = 125\,^{\circ}\text{C}$ AND NO SELF-HEATING

Year	2001	2002	2003	2004	2005	2006	2007
M1 line	4.2	5.4	6.5	7.1	8.1	8.2	7.3
M2 line	4.2	5.4	6.5	7.1	8.1	8.2	7.3
M4 mesh	1.0	1.0	1.0	1.2	1.2	1.3	1.1
M5 mesh	1.0	1.0	1.0	1.2	1.2	1.3	1.1

TABLE VII
IR DROP IN PERCENTAGE OF V_{dd} IN M1 LINE AND M4 GRID FOR SWITCHING ACTIVITIES 0.15 AND 0.5

Year	2001	2002	2003	2004	2005	2006	2007
M1 0.15	5.4	5.9	6.4	5.7	6.5	6.6	8.2
M1 0.5	6.0	6.4	7.0	6.5	7.3	7.5	9.4
M4 0.15	2.1	2.8	2.8	2.6	2.9	3.1	3.7
M4 0.5	2.8	2.9	3.2	2.8	3.1	3.2	4.0

because they have same pitch and aspect ratio and because the different height above the substrate t_{ox} does not have any effect, the self-heating computation being disabled. The trend is similar but the impact of neglecting self-heating is on the order of -15 to $-\%25$ underestimation of the minimum width for M1, -24 to $-\%35$ for M2 and 0 to $-\%30$ for M4 and M5. As anticipated in Section II, our self-heating model uses a worst case approach since we neglect the effect of lower metals and vias that help removing heat. Therefore results obtained neglecting self-heating are best case lower bounds. Nevertheless the trend remains the same as in Sections II–IV.

C. Effect of the IR Drop

We now proceed to verify whether the minimum width for EM reliability is lower or higher than the minimum width needed to limit the IR drop to an acceptable value (e.g., 10% of V_{dd}). We can therefore try to evaluate the IR drop with simple formulas using the width obtained in Sections V-A and V-B. We will follow the approach of [23] for the mesh and we will make an adaptation for the interdigitated structure. The maximum mesh IR drop in the case of a peripheral wire-bond system is located in the mesh center and is $\Delta V_{dd} \approx I_{\text{tot}}R/16$ where I_{tot} is the sum of all average gate currents $(N^2 i_{\text{gate}})$ and R is the resistance of each mesh branch. Since the current we already computed in each branch is $I_{\text{avg}} = 1/4 N i_{\text{gate}}$, the mesh IR drop is $\Delta V_{dd} = N I_{\text{avg}}R/4$.

In an interdigitated finger the maximum IR drop is given by the sum of all drops along the finger due to each cell contribution. It is straightforward to prove that the IR drop is given by

$$Ri_{\text{gate}}(1 + 2 + \cdots + N) = Ri_{\text{gate}}\frac{N(N+1)}{2}. \quad (16)$$

Since the line average current is $I_{\text{avg}} = N i_{\text{gate}}$ the resulting IR drop will be $\Delta V_{dd} = (N+1)I_{\text{avg}}R/2$.

Considering the widths reported in Tables III and IV for the two cases of α, we obtain the corresponding values of ΔV_{dd} reported in Table VII and expressed in percentage of V_{dd} for the case of M1 line and M4 grid. Small values of IR drop are obtained for the mesh case while the fingers have higher values, still lower than 10%. Therefore it seems that the EM constraint is more compelling than the IR drop for the local power distribution within the block. For higher hierarchy levels this could not be true as it is discussed in Section V-D. It should be observed that both local and global power distribution levels concur in defining the overall IR drop such that even a small value in lower levels may reduce the IR drop budget significantly. A more sophisticated model which simultaneously considers local
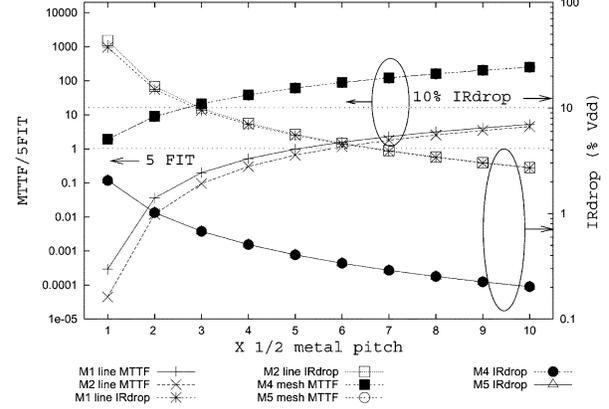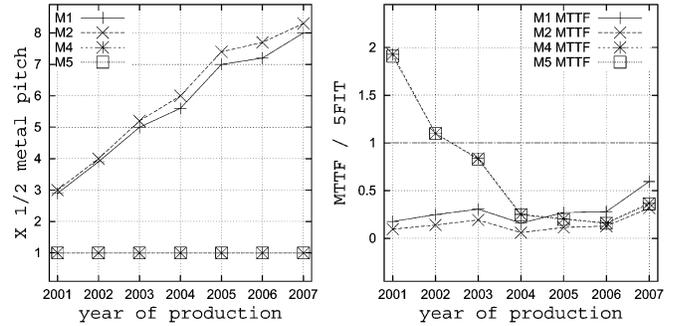

Fig. 9. IR drop and MTTF/5FIT for the 2001 year.

Fig. 10. Linewidth for 10% IR drop and corresponding MTTF/5FIT for all technology nodes.

and global wires should be used in order to evaluate the reciprocal impact of both EM and total IR drop. The results may be strongly design dependent and our high-level analysis does not clearly predict if one effect dominates the other.

Results in Table VII are obtained by setting the reliability constraint. Oppositely, if we set a constraint on IR drop of 10%, following the previous analysis we expect that the 5 FIT MTTF will not be respected. In Fig. 9 IR drop and normalized time-to-failure MTTF/5FIT are plotted as a function of the line expressed in half metal pitches. As expected, the 10% IR drop width line is smaller than the 5 FIT width for all cases. In Fig. 10 the line sizing that satisfies the constraint of 10% IR drop is plotted for all roadmap years as well as the corresponding MTTF/5FIT. We can notice that the width values in terms of half metal pitch increase, meaning that also the IR drop constraint leads to a non scalability of power ground rails. However widths are always smaller than data in Table III and the MTTF is smaller than 5FIT except for the mesh in the first two years. Therefore this confirms that the EM constraint is stricter for local power distribution.
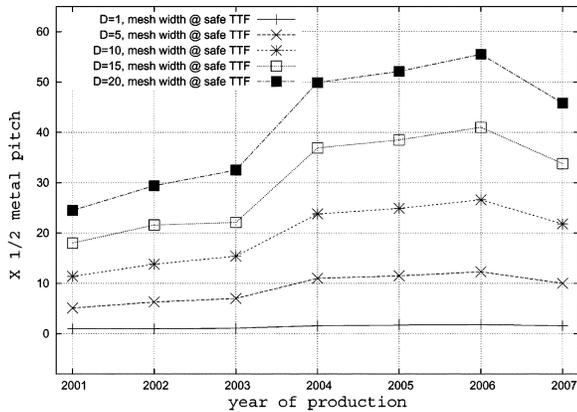
Fig. 11.   Minimum width in half pitches for the M4 lines of the mesh for different mesh densities.
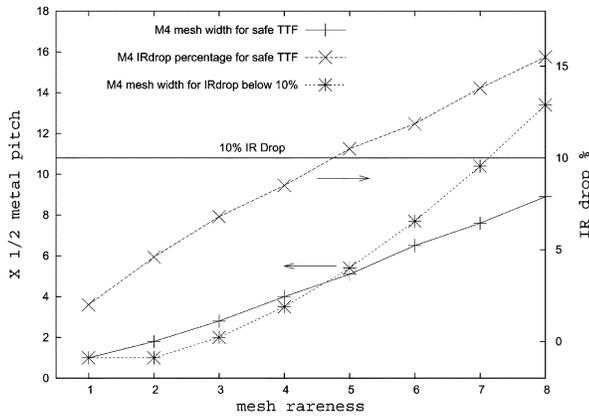

Fig. 12.   130-nm technology nodes: comparison between minimum width for IR drop and EM.

D. Mesh Density

In this work we assumed to have a very dense power grid because we have $N \times N$ grid lines and $N \times N$ gates per functional block. Such assumption could be questionable. However it can be thought as a best case power distribution. A more realistic intermediate case is represented by a less dense mesh. However, if we decrease the density, instead of having one gate per grid point we will have a cluster of gates served by each point. The amount of current carried by each line will be higher approximately of the same number of gates in the cluster and the linewidth needed to comply with the constraint larger of about the same ratio. The trend with technology nodes will be the same. This is evident in Fig. 11 where the minimum width in half pitches is reported for different mesh densities. Index D is such that the number of vertical or horizontal lines is N/D and is then a mesh "rareness" index. The less dense mesh can also be thought as a "global" level in the hierarchy of the power distribution network, while the full mesh is the "local" power. If we evaluate at the same time the minimum width for reliability and the width for 10% IR drop we see that beyond a certain density the IR drop constraint is more compelling. For the 2001 year this occurs at D between 4 and 5 as shown in Fig. 12. In the same figure the two IR drop curves obtained by a sizing strategy for reliability or for IR drop avoidance are also plotted.

Year	2001	2002	2003	2004	2005	2006	2007
M1 0.15	2	3	3.9	5.3	5.7	7	6.1
M1 0.5	6.2	9.1	11.9	15.5	16.8	20.2	17.7
M4 0.15	1	1	1	1	1	1	1
M4 0.5	1	1	1.2	1.9	2	2.5	2

Based on these results, it is likely that for global power wires, the reliability constraint will be met if the IR drop limit is used for design. However, due to the unavoidable simplifications made to carry out this analysis, it is clear that both phenomena concur at setting the constraints and that they should be simultaneously taken into account in the power distribution design phase.

E. Area-Array Bonding Packaging

In the previous derivation, we assumed that the mesh or the fingers are connected to two ideal rings of V_{dd} and gnd surrounding the blocks of logic gates. Such model represents both the case of a peripheral wire-bonding package and an area-array bonding, if the block size is lower than the array pad pitch. If the latter packaging is used and the block size is larger, this can be partitioned in smaller subblocks whose boundary is defined by the pad pitch. The previous formulation applies then to the subblocks by simply recomputing the smaller number of gates within the boundary. By taking the roadmap data for packaging and area-array pad pitch, we observe that our block of 50 kgates is larger than such pitch. Therefore, it can be partitioned according to the new subblock size. We obtained the new projection of the minimum width for reliability, reported in Table VIII, for the case of M1 fingers and M4 mesh, activity factors 0.15 and 0.5. For the fingers case, the width is smaller but still does not scale with technology nodes. For the mesh case and low activity factor, the minimum pitch is sufficient for all nodes. However, for the highest factor, the nonscalability is again evident.

VI. POSSIBLE COUNTERMEASURES

We now look for countermeasures that permit to have the same scaling of metal pitches for the minimum power bus width. Since the results are very sensitive to the actual metal temperature because T_m is at the exponent in (4), we could try to slightly reduce the T_{sub} admitted for each technology node. Another possibility consists in increasing the aspect ratios. However, there will be other issues such as the increased coupling capacitance. Finally, we can analyze what happens if we assume that a clock-gating style is used at the functional blocks level. This will reduce the switching activity.

Decreasing T_{sub}: Variations of -3 to -5% (in absolute temperature, -10 to -15% in Celsius degrees starting from 125 °C) per technology node can compensate for the variation of the other parameters having an effect on the minimum width. In Fig. 13 is reported the substrate temperature projection for the case $\alpha = 0.15$ and initial $T_{\text{sub}} = 125$ °C. We notice, that it is possible to have an increase of temperature in the last node because of the same reason why the minimum width decreases in
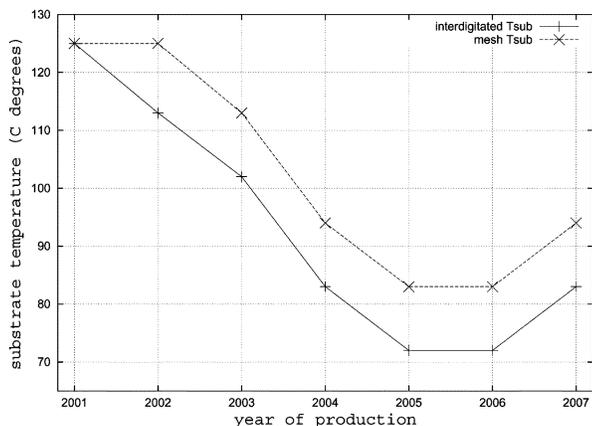
Fig. 13. Projection of the substrate temperature needed to have the same scaling in metal pitch and minimum power bus width.
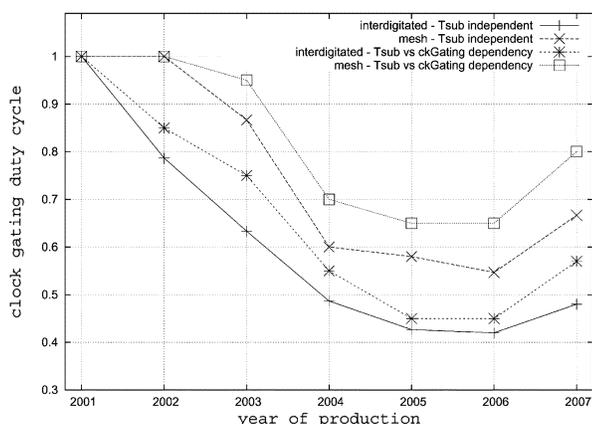

Fig. 14. Projection of the clock gating duty cycle needed to have the same scaling in metal pitch and minimum power bus width.

Fig. 6. Such a solution would require a progressively increasing employment of cooling or a more efficient heat dissipation.

Increasing Aspect Ratios: While this solution should be correct both for metal resistivity (i.e., self-heating) and current density, we observed that not only an excessively big aspect ratio would be needed (e.g., 2.4 instead of 1.6 passing from 2001 to 2002 for the interdigitated structure and 2.8 instead of 1.7 for the mesh in 2004) but it should also increase progressively (3.5 in 2003 for interdigitated and 3.0 in 2005 for mesh). Moreover, high-aspect ratios increase the coupling capacitance of signal wires and, thus, exacerbate signal integrity issues that are already one of the most difficult challenges in deep submicron technologies. Therefore, even if slightly higher aspect ratios are manufacturable, this solution is not feasible.

Clock-Gating Styles: We can model at first order the effect of clock-gating as a modification of the switching activity like

$$\alpha' = \alpha \cdot \frac{T_{\text{on}}}{T_{\text{off}} + T_{\text{on}}} = \alpha \cdot DC_{\text{ckg}} \qquad (17)$$

where DC_{ckg} is the clock gating duty cycle.

As a first approximation, the substrate temperature can be assumed independent of the clock gating duty cycle. Actually, the power saving reduces the substrate temperature. The two classes of curves in Fig. 14 represent the value of DC_{ckg} needed to have the same scaling in minimum width for EM and in metal

pitches, starting from a no clock-gating policy (i.e., $DC_{\text{ckg}} = 1$) in 2001 for the case $\alpha = 0.15$. In the first class $T_{\text{sub}} = 125°C$ *independent* of DC_{ckg}. The second class includes the substrate temperature as an increasing linear function of power dissipation [6]. As expected, the second more realistic DC_{ckg} target is higher than the first one. The temperature range for the second class of curves is 70 °C–125 °C for the interdigitated fingers and 90–125°C for the mesh. These values of DC_{ckg} can also be thought as the "additional" amount of clock-gating needed in technology nodes starting from a given DC_{ckg} already taken into account in α' in 2001.

The use of clock-gating strategies seems a good possibility to contrast the nonscalability. However, its possibility of use is strictly dependent on the application and on the design style of functional blocks.

VII. SUMMARY

In this work, we derived a high-level electrothermal model of the power-supply distribution network including the self-heating and a statistical model of average and rms current activities in homogeneous functional blocks like the ones used to build complex heterogeneous SoCs. We then used such models together with a fanout distribution and an interconnect wirelength model in order to evaluate the minimum width of power supply lines that satisfies the EM reliability constraints according to values reported in the 2001 SIA roadmap document. We have observed, under different conditions of substrate temperature and switching activity, that such minimum width does not scale with the technology nodes giving rise to a reduction in wiring resources. Possible countermeasures are suggested like a progressive reduction in substrate temperature, to be obtained with cooling or technological improvements, and an additional use of clock-gating design styles to be used where the applications requirements permit its employment.

REFERENCES

[1] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI.* Reading, MA: Addison-Wesley, 1990, ch. 9.
[2] B. Geuskens and K. Rose, *Modeling Microprocessor Performance.* Norwell, MA: Kluwer, 1998.
[3] D. Sylvester and C. Hu, "Analytical modeling and characterization of deep-submicrometer interconnect," *Proc. IEEE*, vol. 89, May 2001.
[4] J. C. Eble, "A Generic System Simulator With Novel On-Chip Cache and Throughput Models for Gigascale Integration," Ph.D. dissertation, Georgia Inst. Tech, Atlanta, GA, 1998.
[5] A. E. Caldwell *et al.*, "GTX: The MARCO GSRC technology extrapolation system," in *Proc. DAC*, 2000, pp. 693–698.
[6] K. Banerjee and A. Merhotra, "Global (interconnect) warming," *IEEE Circ. Dev. Mag.*, pp. 16–32, Sept. 2001.
[7] M. R. Casu *et al.*, "Clock distribution network optimization under self-heating and timing constraints," in *Proc. PATMOS*, 2002, pp. 198–208.
[8] D. T. Blaauw, "Static electromigration analysis for on-chip signal interconnects," *IEEE Trans. Computer-Aided Design*, vol. 22, pp. 39–48, Jan. 2003.
[9] J. R. Black, "Electromigration-a brief survey and some recent results," *IEEE Trans. Electron Devices*, vol. ED-16, pp. 338–347, Apr. 1974.
[10] W. R. Hunter, "Self-consistent solutions for allowed interconnect current density-Part I: Implications for technology evolution," *IEEE Trans. Electron Devices*, vol. 44, pp. 304–309, Feb. 1997.
[11] K. Banerjee *et al.*, "On thermal effects in deep sub-micron VLSI interconnects," in *Proc. DAC*, 1999, pp. 885–891.
[12] C. Teng *et al.*, "Item: A temperature-dependent electromigration reliability diagnosis tool," *IEEE Trans. Computer-Aided Design*, vol. 16, pp. 882–893, Aug. 1997.

[13] D. Chen, "Interconnect thermal modeling for accurate simulation of circuit timing and reliability," *IEEE Trans. Computer-Aided Design*, vol. 19, pp. 197–205, Feb. 2000.

[14] "The International Technology Roadmap for Semiconductors," SIA., 2001.

[15] P. Zarkesh-Ha *et al.*, "Prediction of net-length distribution for global interconnects in a heterogeneous system-on-a-chip," *IEEE Trans. VLSI Syst.*, vol. 8, pp. 649–659, Dec. 2000.

[16] P. Christie and D. Stroobandt, "The interpretation and application of rent's rule," *IEEE Trans. VLSI Syst.*, vol. 8, pp. 639–648, Dec. 2000.

[17] D. Sylvester and K. Keutzer, "Impact of small process geometries in system on chip," *Proc. IEEE*, vol. 89, Apr. 2001.

[18] BACPAC: The Berkeley Advanced Chip Performance Calculator. [Online]. Available: http://www.device.eecs.berkeley.edu/~dennis/BACPAC

[19] P. Kapur *et al.*, "Technology and reliability constrained future copper interconnects-Part I: Resistance modeling," *IEEE Trans. Electron Devices*, vol. 49, pp. 590–597, Apr. 2002.

[20] M. R. Casu *et al.*, "Power supply wire sizing considering self-heating in bulk-to-SOI migrated designs," in *Proc. PATMOS*, 2001, pp. 8.3.1–8.3.10.

[21] W. E. Donath, "Wire length distribution for placements of computer logic," *IBM J. Res. Develop.*, vol. 25, pp. 152–155, May 1981.

[22] J. A. Davis *et al.*, "A stocastic wire length distribution for gigascale integration (GSI)—Part I: Derivation and validation," *IEEE Trans. Electron Devices*, vol. 45, pp. 580–589, Mar. 1998.

[23] P. Zarkesh-Ha, "Global interconnect modeling for a gigascale system-on-a-chip (GSoC)," Ph.D. dissertation, Georgia Inst. Tech., Atlanta, GA, 2001.

[24] GTX: The MARCO GSRC Technology Extrapolation System. [Online]. Available: http://nexus6.cs.ucla.edu/GSRC/GTX

[25] D. M. Sylvester, "Analytical Modeling and Characterization of Deep Submicron Interconnect," Ph.D. dissertation, Univ. Calif., Berkeley, 1999.

Mario R. Casu received the electronics engineer degree (*summa cum laude*) and the Ph.D. degree from the Politecnico di Torino, Torino, Italy, in 1998 and 2002, respectively.

In 1998, he joined the Very Large Scale Integration (VLSI) Laboratory, Department of Electronics, Politecnico di Torino, first as a Ph.D. Student and now as a Researcher. In 2001, he was with ST Microlectronics Central Research and Development Crolles, France, working on SRAM development and CMOS library characterization on a partially-depleted 0.13-μm SOI technology. His research activities are in the field of SOI and bulk CMOS circuits modeling and design, ultradeep submicron issues and related computer-aided design tools and systems-on-chip performance prediction.

Mariagrazia Graziano received the Dr.Eng. degree and the Ph.D. degree in electronics engineering from the Politecnico di Torino, Torino, Italy, in 1988 and 2001, respectively.

From 1999 to 2000, she was a Visiting Scholar at University of Illinois at Chicago. Since 2002, she has been a Researcher at the Politecnico di Torino. Her research interests include the development of models and physical design algorithms for the design of CMOS high-speed and low-noise digital circuits. She is author and coauthor of more than 30 publications.

Guido Masera (M'98) received the Dr.Eng. (summa cum laude) and the Ph.D. degrees in electrical engineering from Politecnico di Torino, Torino, Italy, in 1986 and 1992, respectively.

From 1986 to 1988, he was with Centro Studi e Laboratori in Telecomunicazioni (CSELT), Torino, Italy as a Researcher, where he was involved in the standardization activities for the GSM system. In 1992, he joined the Electronics Department, Politecnico di Torino, where he was an Assistant Professor and then Associate Professor and a member of the VLSI-Lab group. His research interests include several aspects in the design of digital integrated circuits and systems, with special emphasis on high-performance architecture development (especially for wireless communications and multimedia applications) and on-chip interconnect modeling and optimization. He has coauthored more than 80 journal and conference papers in the areas of ASIC-systems-on-chip (SoC) development, architectural synthesis, very large scale integration (VLSI) circuit modeling and optimization. In the frame of National and European research projects, he has been a co-designer of several ASIC and field-programmable gate arrays (FPGA) implementations in the fields of artificial intelligence, computer networks, digital signal processing, transmission, and coding.

Gianluca Piccinini received the Dr.Ing. and the Ph.D. degrees in electronics engineering, from Politechnico di Torino, Italy, in 1986 and 1990, respectively.

Since 1998, is has been an Associate Professor with the Department of Electronics, Politecnico di Torino, Italy. His research interests include CMOS very large scale integration (VLSI) design methodologies, VLSI architectures for digital-signal processing, and VLSI circuits for communication networks. Currently, his research activity is mainly focused on the design and optimization of high-speed circuits in CMOS Technology. He is author and co-author of more than 50 publications and holds one international patent.

Maurizio Zamboni received the electronics engineer degree and the Ph.D. degree, from the Politecnico de Torino, Torino, Italy, in 1983 and 1988, respectively.

In 1983, he joined the Electronics Department, Politecnico di Torino, where he became a Researcher in 1989, and an Associate Professor in 1992. Since 1992, he has been teaching a course in applied electronics and since 1993, he has been teaching a course in integrated circuits architectures. His research activity began with the study of multiprocessor architectures, followed by work with digital IC design. His current research interests include the design of special ICs for artificial intelligence, vision, and telecommunication. He has worked with several European projects in the field of ATM networks and digital television. He is coauthor of more than 60 scientific papers and has two patents.