Analysis of IR-Drop Scaling with Implications for Deep Submicron P/G Network Designs

Amir H. Ajami[1*], Kaustav Banerjee[2], Amit Mehrotra[3], and Massoud Pedram[1]
[1]Department of Electrical Engineering - Systems, University of Southern California, Los Angeles, CA 90089
[2]Electrical and Computer Engineering Department, University of California at Santa Barbara, CA 93106
[3]Computer and System Research Lab, University of Illinois, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Abstract - *This paper presents a detailed analysis of the power-supply voltage (IR) drop scaling in DSM technologies. For the first time, the effects of temperature, electromigration and interconnect technology scaling (including resistivity increase of Cu interconnects due to electron surface scattering and finite barrier thickness) are taken into consideration during this analysis. It is shown that the IR-drop effect in the power/ground (P/G) network increases rapidly with technology scaling, and using well-known counter measures such as wire-sizing and decoupling capacitor insertion with resource allocation schemes that are typically used in the present designs may not be sufficient to limit the voltage fluctuations over the power grid for future technologies. It is also shown that such voltage drops on power lines of switching devices in a clock network can introduce significant amount of skew which in turn degrades the signal integrity.*

1 Introduction

With CMOS process technology scaling down to 0.13μm and below, IR-drop is becoming an extremely important phenomenon determining the performance and reliability of ULSI designs. The IR-drop effect manifests itself in the P/G distribution networks and can adversely influence the performance of the signal nets including the clock tree [1],[2],[3]. Aggressive interconnect scaling increases the resistance per unit length of wires and the average current density. Since the supply voltage level is also reduced with technology scaling, the IR-drop effect becomes even more problematic since the ratio of the voltage drop to the ideal supply voltage level increases, which in turn degrades the switching speed of the CMOS gates and their DC noise margins. An excessive voltage drop in the power grid may also result in a functional failure in dynamic logic and a timing violation in static logic. It has been shown that a 10% voltage drop in a 0.18μm design increases the propagation delay of the gates by up to 8% [1].

As a result, the main challenge in the design of the power distribution network is to achieve a minimum acceptable voltage fluctuation across the chip (~10% of V_{dd}) while satisfying the electromigration (EM) reliability rule for the power network segments and to realize such power distribution network by consuming minimum routing area of the interconnect metal layers [4]. A critical issue in the analysis of power distribution network is the large size of the problem. Simulating all the nonlinear devices in the chip together with non-ideal power grid is not computationally feasible. Thus, the simulation is usually carried out in two separate steps. First, the non-linear devices are simulated assuming perfect supply voltages and the currents drawn by the devices are calculated. Next, the devices are modeled as independent time-varying current sources. The error incurred by ignoring this non-linearity is usually negligible. By performing these two steps, the problem of power grid analysis is reduced to solving a linear network [5]. Besides the wire-sizing technique, in order to reduce the effect of switching noise on the P/G network, decoupling capacitors are added near the switching devices over the substrate. These capacitors act as local reservoirs of charges for switching circuits and reduce the effect of the power supply glitches and ground bounce. Determining optimal values and placement of on-chip decoupling capacitors are essential steps in maintaining a robust P/G network [6],[7]. In addition to the increase in the resistance per unit length of the metal layers with technology scaling, some physical effects such as electron surface scattering and finite barrier thickness contribute significantly to the overall metal resistivity of the local thin lines. Furthermore, it has been shown that as technology feature size is reduced, the peak chip temperatures that occur on the global metal layers increase rapidly due to the self-heating effect [8]. This can cause further increase in the metal resistivity. Most of the recent research reports on the IR-drop effect have mainly focused on the methodology of efficient computation of the voltage drop values for each gate in typical P/G networks [9],[10]. In this paper various effects of the technology scaling and temperature issues in analyzing the IR-drop phenomenon are considered. It is shown that the IR-drop effect in the power/ground (P/G) network increases rapidly with technology scaling and that using well known counter measures such as wire-sizing and decoupling capacitor insertion with resource allocation schemes that are commonly used in present designs may not be sufficient to limit the voltage fluctuations over the power grid for future technologies and new guidelines should be introduced.

The remainder of this paper is organized as follows. An overview of global and local power grid distribution network topologies is described in Section 2. In Section 3 the methodology of calculating the minimum number of necessary power tracks in the global and semi-global grids in order to satisfy EM rules is discussed. In Section 4 the effects of technology scaling, including the thermal effects, and barrier and thin-film effects, on the worst-case IR-drop are studied. It is shown how thermal effects due to substrate hot spots on the global interconnect may affect the worst-case IR-drop. Section 5 examines that how the technology scaling affects the performance of the cell switching activities due to the power network voltage drops. Finally, concluding remarks and summary are presented in Section 6.

2 Topology of Power Distribution Networks for IR-drop Analysis

Function of the power network is to carry current from the power chip pads to all the cells in the design. The power network has complex and tight electrical specifications, making its design a challenging task. It often consists of a top-level grid network that distributes current from the chip power pads (which are uniformly distributed over the chip area) to the local power trunks and low-level distribution structures that distribute the current from these trunks to the cells. The top-level grid itself is made of global and/or semi-global wire lines that are connected together through vias or stack of vias. Initially, the number and width of the orthogonal metal lines in the global/semi-global power grids are determined based on the EM rules. Simulating the power grid requires solving a set of differential equations that are formed through a typical approach like the *modified nodal analysis* (MNA) as follows:

$$\mathbf{G} \cdot x(t) + \mathbf{C} \cdot \dot{x}(t) = u(t) \qquad (1)$$

where **G** is the grid conductance matrix, **C** is the grid capacitive (including the decoupling capacitances) and inductive matrix, $x(t)$ is the time-varying vector of grid node voltages and currents through the inductors, and $u(t)$ is the vector of time-varying current sources attached to grid nodes. In this work, an *RC* model of the MNA has been used for simplicity. Due to the fact that the grid matrix is very sparse, we used an iterative conjugate gradient method to solve this linear system where it also exploits the symmetry and positive definitivity of the grid matrix. Figure 1 shows the RC network model used for extracting system of (1). Time-varying current sources and decoupling capacitances are

[*] The author is currently with Magma Design Automation, Cupertino, CA.

Proceedings of the Fourth International Symposium on Quality Electronic Design (ISQED'03)
0-7695-1881-8/03 $17.00 © 2003 IEEE

connected to each intermediate node in the global/semi-global grid. The amount of the current and decoupling capacitance can be derived by examining the power consumption profile and device count of the underlying functional blocks on the substrate connected to each grid node. By solving the system of (1), the voltages of each node at the global and semi-global grids are known.

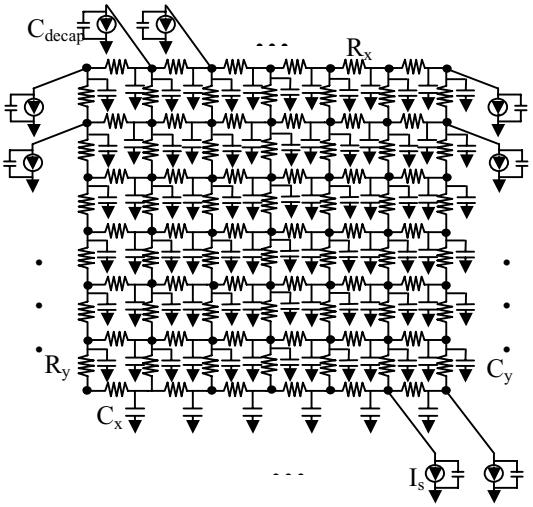

Figure 1: RC model of a power bus network. Each intermediate node is connected to underlying circuit blocks modeled as time-varying current sources I_s's and on-chip decoupling capacitances C_{decap}'s.

In standard-cell based designs, by putting a power trunk adjacent to a cell row, power can be distributed among the cells in that row (Figure 2). We do not discuss the ground distribution network in this paper since its analysis is similar. Accordingly, a number of cells (~ 50 to 100 cells) that belong to the same row of the same functional block in the design are connected to a single power trunk. The power trunks are usually routed in Metal1 and are connected together on one side by using a strip of Metal2, making a *comb*-like structure as shown in Figure 2. To achieve better results both in terms of the local IR-drop and EM reliability, one can use out-of-block extensions and connect both sides of the power trunks together. Without loss of generality, we use inverters to represent the cells that are powered by the local power trunk. A circuit model of the local power trunk is depicted in Figure 3 where we assume that *N-2* identical inverters are connected to a power trunk. Capacitors Cd_i's include both the built-in junction decoupling capacitors, and the add-on (thin-oxide) decoupling capacitances. The total on-chip n-well decoupling capacitor is determined by the area, depth and perimeter of each n-well. In high performance switches, thin-oxide decoupling capacitors should be placed in close proximity of the highly active switching devices. Assuming the trunk as a resistive-only network for the time being, by having voltages V_1 and V_N and modeling the current drawn by each inverter as a current source I_i, voltage V_i at each intermediate node in Figure 2 is calculated as follows:

$$I_{ei+1} = I_{ei} - I_i \ , \ I_{e_i} = \sum_{i=1}^{N-2} \frac{\sum_{j=i+1}^{N-1} R_j}{\sum R_i} I_i \ , \ V_{i+1} = V_i - \frac{R_i}{\sum R_i}(V_1 - V_N) - R_i I_{ei} \quad (2)$$

Note that the resistive voltage drop derived in (2) is the worst-case scenario since the current sources I_i's are dependent on the magnitude of V_{i+1}'s (Figure 3). For calculating the actual IR-drop, one must use the nonlinear voltage-dependent source current by using I_{ds} of the switching device and repeatedly solve (2) until the solution converges. Notice that effects of the decoupling capacitors and interconnect capacitance per unit length has been neglected in deriving (2). Using this model, by solving the linear network matrix coefficient for the power grid through (1), one can solve for the IR-drop for the entire network in an iterative manner. The degree of IR-drop is design-dependent and varies based on the location of cells connected to the grid, the switching activity of each cell, and the location of power pad connections to the grid. As a result, in our experimental setup we examine the IR-drop in a local power trunk by inserting reasonable number of inverters in it. To emphasize the worst-case scenario, it is assumed that all inverters connected to the grid segment switch at the same time. It is also assumed that by using a ball grid type of pin assignments in our problem setup, the power pads are uniformly distributed over the chip area.

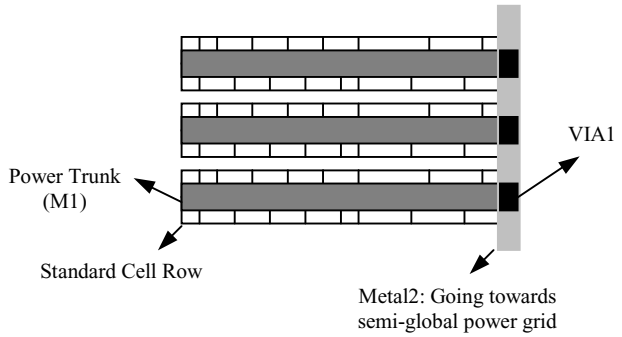

Figure 2: A local power distribution network for a typical standard cell design consisting of power trunks in a *comb*-like structure connecting to the semi-global power grid through metal2.

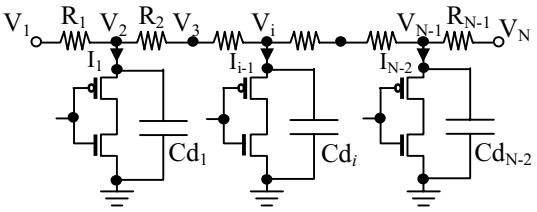

Figure 3: Magnified view of a trunk segment, containing a resistive network with inverters connected to the intermediate nodes.

To alleviate the large transient current flowing through the inductance of the global/semi-global grid and limit the voltage drop, we must place decoupling capacitances throughout the chip. Nominally, the stored charge on these capacitors will supply the required transient current for 10% of the clock period. The charge will be replenished during the remaining 90% of the CPU clock cycle time. To calculate the amount of necessary decoupling capacitances in order to maintain a limited voltage drop, one can use the following:

$$P = \gamma(C_T V_{dd}^2 f) \quad (3)$$

where P is the total chip power consumption, γ is the probability that a power transition occurs, V_{dd} is the supply voltage, f is the clock frequency and C_T is the effective chip capacitance. Assuming a maximum voltage variation of 10%, the computed decoupling capacitance needed for future technologies varies in a range of 39-72 nF/cm^2 (for 0.18 to 0.07µm respectively) [11]. Heuristically, the amount of decoupling capacitance needed to accomplish a limited voltage swing (~10%) can be deduced as [12]:

$$C_{decap} = \frac{9P}{fV_{dd}^2} \quad (4)$$

For a metal-insulator-metal (MIM) capacitor having a dielectric with T_{oxeq}=1nm, the capacitance is about 34.5 fF/µm^2. Using this value and (4) one can calculate the amount and the area of total decoupling capacitance needed in each technology [13]. These values are indicated in Table 1 for different technologies.

3 Methodology for Power Network Planning

A key concern for the P/G network design is the large amount of current that flows through the interconnect lines which gives rise to EM-induced failures. EM is the transport of mass in the metal under an

applied current density and is widely regarded as a major failure mechanism of VLSI interconnects. When current flows through the interconnect metal, an electronic wind is setup opposite to the direction of current flow. These electrons upon colliding with the metal ions, impact sufficient momentum and displace the metal ions from their lattice sites creating vacancies. These vacancies condense to form voids that result in increase of interconnect resistance or even open circuit conduction. EM lifetime reliability of metal interconnects is modeled by the well-known Black's equation [14], given by:

$$TTF = A \cdot j^{-n} \cdot exp(\frac{Q}{k_B T_m}) \qquad (5)$$

where *TTF* is the time-to-fail (typically for 0.1% cumulative failure). *A* is a constant that is dependent on the geometry and microstructure of the interconnect line, and *j* is the average current density. The exponent *n* is typically 2 under nominal conditions, *Q* is the activation energy for grain-boundary diffusion (~0.5eV for 0.1µm Cu), k_B is the Boltzmann's constant, and T_m is the metal temperature. The typical goal is to achieve a 10-year lifetime at 100 °C, for which (5) and accelerated testing data produce a design rule value for the acceptable current density, j_0, at the reference temperature, T_{ref}. However, this design rule value does not comprehend self-heating [15]. Based on the technology roadmap values provided by the ITRS [16], the values for the maximum allowable current density, j_0, at a specific temperature, T_{ref}, for different technologies are given in Table 1. On the other hand it is well know that interconnects at different metal layers experience different temperatures and global interconnects get hotter than the local interconnect lines [8]. Based on the values of j_0 and T_{ref} given in ITRS for different technologies, we can easily calculate the new values for acceptable amount of current density j_m such that the EM lifetime rule still remains satisfied at a new temperature T_m by using the following relationship (which can be deduced from (5)):

$$j_m = j_0 (\exp(\frac{Q}{k_B}(\frac{T_m - T_{ref}}{T_m T_{ref}}))^{\frac{-1}{n}} \qquad (6)$$

3.1 Power Network Electromigration Rule Satisfaction

Table 1 lists the different parameters for future CMOS technologies based on ITRS guidelines [16]. The maximum allowable current density j_m for global/semi-global tiers at the maximum temperature has been calculated using (6) in Table 2. With the knowledge of total power consumption and power supply voltage, one can calculate the maximum current drawn from the power supply. Dividing this value by the number of the power pads, which is usually half of the given value of the P/G pads in the ITRS guideline (and is usually 2/3 of the total number of I/O pads in today's IC technologies), one can approximately calculate the average current drawn from each power pad. Note that a ball grid type of I/O packaging has been assumed in this analysis. The maximum current drawn from each power pad is a limiting factor on the EM rule for the power grid interconnects in the area surrounded by that pad. In general, the minimum number of the minimum-width gridlines required in a global power network in order to satisfy the EM rules can be approximately calculated as follows:

$$\#Tracks \simeq \frac{1}{w}(\frac{1}{ar} \times \frac{P/V_{dd}}{N_{pad} j_m})^{0.5} \qquad (7)$$

where *w* is the minimum width of each power track at the corresponding metal layer (i.e. global or semi-global and it is usually half of its defined pitch), *ar* is the aspect ratio, *P* is the total power consumption, V_{dd} is the supply voltage, N_{pad} is the number of power pads, and j_m is the maximum allowable current density to satisfy the EM rule at the corresponding layer (i.e. global or semi-global). Using (7) and Table 1, the minimum number of minimum-width gridlines needed for different technologies in order to satisfy EM rules in global and semi-global tiers is shown in the Table 2. It can be observed from Table 2 that, by going from global tiers to semi-global ones, the power grid gradually gets denser which is expected due to the decrease in the pitch.

Node (µm)	**0.18**	**0.13**	**0.1**	**0.07**
j_0(A/cm^2)	5.8E5	9.6E5	1.4E6	2.1E6
Chip size (mm^2)	450	450	622	713
V_{dd} (V)	1.8	1.5	1.2	0.9
Frequency (MHz)	1000	1700	3000	5000
P (W)	90	130	160	170
On_Chip C_Decap (nF)	250	305	333	377
T_{max} (°C)	120	140	150	175
# of P/G pads	1536	2018	2018	2560
Global pitch (nm)	1050	765	560	390
Semi-global pitch (nm)	640	465	340	240
Global layer line *ar*	2.2	2.5	2.7	2.8
Semi-global line *ar*	2.0	2.2	2.4	2.5
R-local (KΩ/m)	76.23	125.96	219.56	435.5

Table 1: Technology parameters used in this work based on ITRS data for Cu interconnects.

Node (µm)	**0.18**	**0.13**	**0.1**	**0.07**
T_{max} global (°C)	120	130	162	170
T_{max} semiglobal (°C)	117	126	150	160
j_m/j_0	0.74	0.62	0.36	0.33
#global tracks @ 105 °C	526	796	1451	2346
#global tracks @ T_{max}	705	1281	3964	7230
#semiglobal tracks @ 105 °C	1559	2448	4429	6940
#semiglobal tracks @T_{max}	2050	3600	10016	18385

Table 2: Minimum number of (minimum-width) power tracks needed to be routed on the power grid at global/semi-global tiers for different technology nodes (in order to satisfy the EM rules) for T=105 °C and T_{max}.

4 Effects of Technology Scaling on the IR-drop Effect

4.1 Effects of Thin-Film, Barrier Thickness and Interconnect Temperature

In ULSI interconnects, metal resistivity begins to increase as the minimum dimension of the metal line becomes comparable to the mean free path of the electrons. This is because surface scattering starts having a considerable contribution to the resistivity compared to the contribution due to the bulk scattering. Resistivity ρ of a thin-film metal can be expressed in terms of the bulk resistivity ρ_0 as [17]:

$$\frac{\rho_0}{\rho_{thin\_film}} = 1 - \frac{3}{2k}(1-p)\int_1^{\infty}(\frac{1}{x^3} - \frac{1}{x^5})\frac{1-e^{-kx}}{1-pe^{-kx}}dx \qquad (8)$$

where $k=d/\lambda_{mfp}$, *d* is the smallest dimension of the film (width in our case), λ_{mfp} is the bulk mean free path of electrons and *p* is the fraction of electrons which are elastically reflected at the surface. For Copper p=0.47 and λ_{mfp}=421Å at 0 °C. Moreover, since the temperature alters the mean free path of the electrons, the temperature coefficient of resistivity, α, of the thin film is also different from its bulk temperature coefficient α_0. Another effect, which is responsible for increased resistivity, is the presence of barrier material for Cu interconnects. Since the resistivity of the barrier material is extremely high compared to Cu, it can be assumed that Cu carries all the current. Therefore, the effective area through which the current conduction takes place reduces, or equivalently the effective resistivity of the metal line of the same drawn dimension increases. This becomes more of a problem as metal lines scale since it is very difficult to scale the thickness of the barrier material. The effective resistivity and temperature coefficient ratios for

the global, semi-global and local tier metal for various technology nodes are given in Tables 3.

Node (μm)	0.18	0.13	0.1	0.07
(Global)ρ/ρ_0	1.066	1.090	1.125	1.186
(Semi-global) ρ/ρ_0	1.113	1.158	1.222	1.334
(Local) ρ/ρ_0	1.158	1.222	1.315	1.485
(Global) α/α_0	0.953	0.935	0.912	0.875
(Semi-global) α/α_0	0.923	0.895	0.858	0.803
(Local)α/α_0	0.902	0.867	0.82	0.752

Table 3: Effective resistivity (barrier plus thin-film) and temperature coefficient of resistivity ratios for the global, semi-global and local tiers for various technologies.

It is also well known that interconnect resistance changes linearly with its temperature. This relationship can be written as $R=r_0(1+\beta.\Delta T)$ where r_0 is the unit length resistance at reference temperature and β is the temperature coefficient of resistance (1/°C). By including the effects of scattering and thin-film, this can be re-written as follows:

$$R = r_0(\frac{\rho}{\rho_0})_{thin\_barrier\_eff}(1+\beta\frac{\alpha}{\alpha_0}\Delta T) \qquad (9)$$

As we will see later, in order to reduce the maximum voltage drop, the global and semi-global tier metals usually have widths that are many times larger than the minimum width. As a result the barrier-thickness effect can only be considered for the local lines. On the other hand, the global/semi-global tiers are the hottest lines inside the chip [8]. As a result, for global/semi-global lines the effect of line temperature should be considered.

4.2 IR-drop in Global/Semi-Global Power Network

Based on the minimum required number of the power gridlines as calculated from (7) for both the global and semi-global grids at each technology node, we can build the system of linear equations (1) for combined global/semi-global power grids and solve it to find the voltage at each node. Nodes at the semi-global power grid distribute the power to the local power trunks through a via or a stack of via and/or metal2. Hence, by finding the worst-case voltage drop over the nodes at the semi-global level and accounting for the drop over the vias, one can find the voltage at the power pin of the drivers in the local blocks. In this way one can quantify how severely the global and semi-global power grids can affect the IR-drop in the worst case. To examine the effect of decoupling capacitors, two cases are analyzed as detailed next.

Case I) No decoupling capacitors: Using the number of the tracks provided in Table 2, the resulting voltage drop values would be drastic. Ideally, we need to have less than 10% voltage drop in order to ensure a correct functionality of the circuit. Using the minimum sized tracks would result in a huge voltage drop. As a result, an optimization procedure should be used that (while satisfying the EM rules), attempts to minimize the voltage drop such that a fixed percentage of the routing area gets allocated to the power network. Maximum allocation of 5 to 10% of the routing area to the power network is a usual policy in current technologies. Figure 4 shows the worst-case voltage drop in different technologies for 5% and 10% allocation of the routing area to the power network, respectively, without considering the on-chip decoupling capacitances.

Case II) Uniformly distributed decoupling capacitors: From Figure 4, it can be observed that even with wire sizing up to the allowed budget of the routing area, the voltage drop will be more than the maximum allowable margin of 10%. Insertion of on-chip decoupling capacitors near the switching devices on the substrate decreases the peak magnitude of the voltage drop. The total decoupling capacitor calculated by (4) and reported in Table 1. It can be assumed that the decoupling capacitors are uniformly distributed over the substrate surface. Figure 5 shows the worst-case voltage drop in global/semi-global grids while using projected amount of on-chip decoupling capacitor and 10% of routing area for the power network. From Figure 5, it can be observed that by using the suggested on-chip decoupling capacitors, the worst-case voltage drop reduces to an acceptable margin for 0.18μm technology. However, as the technology node scales towards the sub 130nm regime, the maximum voltage drop violates the 10% voltage swing rule. In the above experiments the area of the total on-chip decoupling capacitors is 5% of the total substrate area.

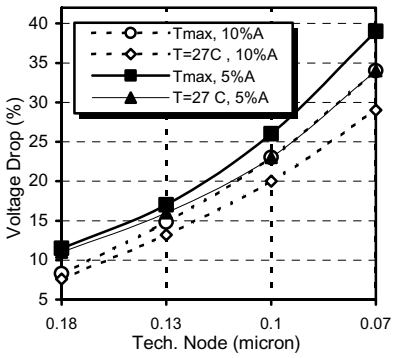

Figure 4: Worst-case voltage drop increase as a function of technology node for combined global/semi-global power grids considering the effects of self-heating, while allocating 5% and 10% of the routing area to the power network, respectively.

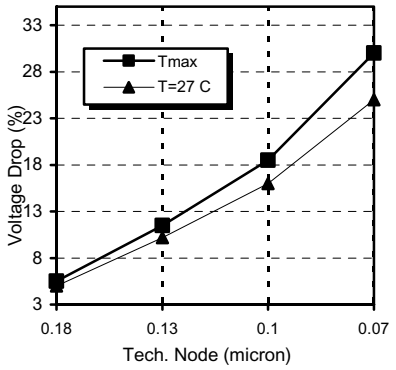

Figure 5: Worst-case voltage-drop increase as a function of technology node for combined global/semi-global power grids considering the effects of self-heating, while allocating 10% of the routing area to the power network and assuming uniformly distributed on-chip decoupling capacitors.

Observations made in two previous cases show that for future technologies, assigning 10% of the routing area to the power network and 5% of the substrate area to the decoupling capacitors are not going to be sufficient in order to limit the maximum voltage drop to the desired value of 10%. As a result, for future technologies, new resource allocation limitations should be determined. Figure 6 shows the minimum required percentage of the allocated resources to ensure a worst-case 10% voltage drop for future technologies.

4.3 IR-drop in Local Power Network

Using the worst-case voltage drops from previous section, we can impose them on the two sides of each power trunk to examine the IR-drop in the local power trunks (Figure 3). Figure 7 shows the worst-case voltage drop increase as a function of technology node in the local power trunks. Note that the actual voltage applied at the two ends of the power trunk are at V_{dd}-$V_{dd'}$ where $V_{dd'}$ can be extracted from Figure 5 for different technologies. To extract the total worst-case voltage drop, one must combine the results of the two previous steps. Using Figures 5 and 7, Figure 8 summarizes the total voltage drop increase as a function of technology node in the presence of barrier effect/thin-film and

temperature effects. Also note that the worst-case IR-drops for different technologies are based on the assumption of uniformly distributed power pads all over the chip area, which is the emerging trend in the industry. By using the periphery-only power pad distribution scheme, the worst-case voltage drop is going to be much more severe than the results extracted by Figure 5 and 8.

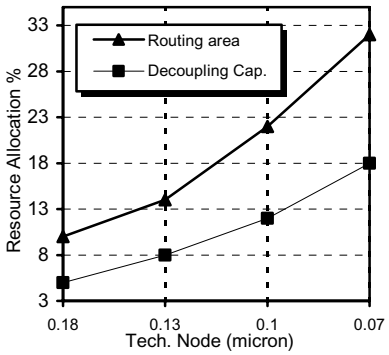

Figure 6: Minimum required percentage of the allocated resources (global layer routing area and substrate area) to ensure a worst-case 10% voltage drop for future technologies, considering the maximum temperature on global/semi-global interconnects.

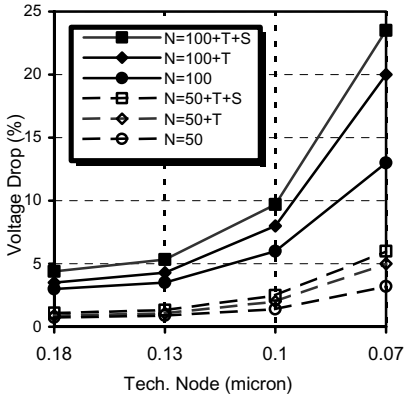

Figure 7: Worst-case voltage-drop increase as a function of technology node in the presence of maximum interconnect temperature (T) and surface scattering/barrier effects (S), in the local power trunk lines. Notice that in this graph (V_{dd}-$V_{dd'}$) is the actual voltage over the two sides of the local power trunks, and *N* is the number of standard cell connected to the power trunk.

4.4 Effect of Hot Spots on the Worst-case IR-drop

In reality, the magnitudes of the current sources connected to the power grid are not uniformly distributed. Due to different switching activities and/or sleep modes of various functional blocks, the distribution of current sources over the power network is generally non-uniform. As a result, one should distribute the decoupling capacitances non-uniformly according to the activity profiles of the different blocks over the substrate. Existence of such non-uniformly distributed switching activities on the substrate results in substrate thermal gradients and in extreme cases leads to the creation of hot spots.

However, the existence of such hot spots along the substrate surface introduces non-uniform temperature profiles along the lengths of the long global interconnects. More specifically, the power distribution network spans over the entire substrate area and it is exposed to the thermal non-uniformities of the substrate surface. It has been shown that any kind of thermal non-uniformities on the substrate surface would affect globally interconnect performance [18]. Having the power consumption profile of the blocks over the substrate, one can easily determine the substrate thermal profile. To derive the thermal profile of a long global interconnect passing over the substrate, one can use the following [18]:

$$\frac{d^2T_{line}(x)}{dx^2} = \lambda^2 T_{line}(x) - \lambda^2 T_{sub}(x) - \theta \qquad (10)$$

where $T_{line}(x)$ and $T_{sub}(x)$ are the interconnect thermal profile and substrate thermal profile along the length of the interconnect, respectively, and λ and θ are two constant which can be derived using the physical dimensions of the interconnect line and the insulator and thermo-electrical properties of the interconnect metal. It is also well known that resistivity of a metal has a linear relationship with the thermal profile of the line. As a result, due to the non-uniformity of the substrate temperature, resistance profile of the power network would distribute non-uniformly. Specifically, resistance of those segments on top of the hot spots is going to be higher than the rest of the power network segments. It is expected that by considering the actual temperature-dependent resistivity of the global interconnect, the IR-drop value at those nodes in the proximity of the hot spots should worsen. To model a hot spot, a constant peak Gaussian distribution thermal profile with a constant standard deviation ($T(x) = T_{\max} \cdot \exp(-(x-\mu)^2 / 2\sigma^2)$) is assumed here. Figure 9 shows the variations of the worst-case IR-drop as a function of the magnitude of thermal gradient of a hot spot over the substrate surface. It can be seen that by neglecting the thermal effects of hot spots on the resistivity of the global layers, one can not predict the worst-case voltage drop of hot devices correctly, and consequently the amount of the inserted decoupling capacitance proposed by current heuristics is not sufficient.

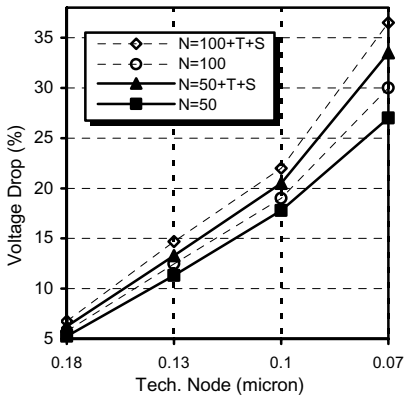

Figure 8: Total worst-case voltage-drop increase as a function of technology node in the presence of maximum interconnect temperature (T) and surface scattering/barrier effects (S), while allocating 10% of the routing area to the power network and 5% of the substrate for decoupling capacitor.

5 Effects of the IR-drop on the Cell Performance and Clock Skew

Performance of each cell connected to the local trunk segment is strongly dependent on the fluctuations of power supply voltage (V_{dd}). For deriving the sensitivity of the gate delay as a function of the changes in V_{dd}, we can use a simple short-channel model for transistors in the saturation region. The I_{ds} can be expressed as follows:

$$I_{ds} = w \cdot v_{sat} \cdot C_{ox} \cdot (V_{gs} - V_t - V_{ds}) \qquad (11)$$

where C_{ox} is the oxide capacitance, V_{gs} is the gate to source voltage, v_{sat} is the carrier saturation velocity, V_{ds} is the drain to source voltage and V_t is the threshold voltage. The gate delay sensitivity, S^D_{Vdd}, to the power supply voltage fluctuations can be written as follows:

$$S^D_{Vdd} = \frac{V_{dd}V_T - V_T^2 + E_c L V_{dd} + E_c L V_T}{(V_{dd} - V_T + E_c L)(V_{dd} - V_T)} \qquad (12)$$

where E_C is the critical electric field, L is the channel length ($E_C L$=1.4 V) and V_T is assumed to be V_{dd}/5. As shown in Figure 10, by technology scaling, the dependency of the gate delay on the power supply voltage fluctuations becomes more severe. Notice that Figure 10 depicts the *sensitivity* of power supply voltage variations as a function of technology node. As an example, it can be seen from Figure 10 that for each 10% decrease in the power supply voltage in the 0.18 µm technology, we expect to see an 8.5% increase in the gate delay. Figure 11 shows the maximum percentage of delay difference among the devices connected to a semi-global grid for different technologies. This delay difference will appear as skew among the devices in a clock circuitry. It can be observed that technology scaling can introduce a considerable amount of skew into the clock tree by affecting the performance of the clock buffers through non-uniform voltage drop over the power grid network.

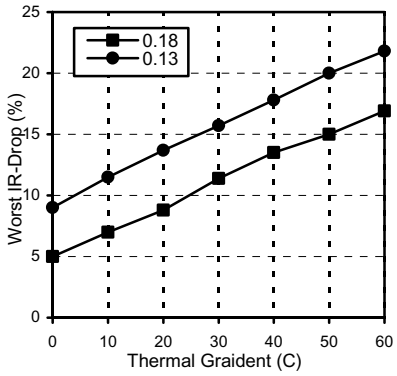

Figure 9: Worst-case voltage-drop ($\Delta V_{IR}/V_{dd}$) increase (based on Figure 5) as a function of technology node in the presence of hot spots as a function of thermal gradient magnitudes (°C) shown for two technology nodes.

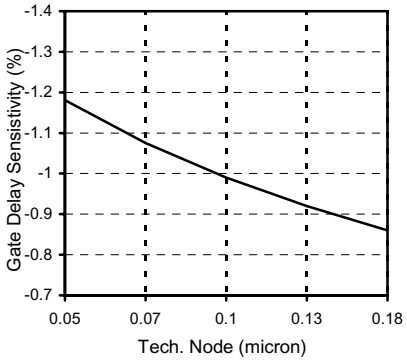

Figure 10: Sensitivity of the cell delay (S^D_{Vdd}) to the fluctuations of the supply voltage V_{dd} for different technology nodes. Y-axis values show the percentage increase in gate delay for each percent decrease in V_{dd} at the specific technology.

6 Conclusion

In this paper we have highlighted the growing importance of the IR-drop effects with technology scaling. The effects of temperature, electromigration reliability and interconnect technology scaling including resistivity increase of Cu interconnects due to electron surface scattering and finite barrier thickness has been taken into consideration for IR-drop analysis. Severe performance degradation and/or functional alterations due to power network IR-drop suggests that IR-drop issue is going to become an increasingly important factor determining P/G networks interconnect design policies and signal integrity guidelines. It has been shown that new resource allocation guidelines for P/G metal area and on-chip decoupling capacitors should be provided for future technologies in order to limit the maximum voltage swing in the P/G distribution networks. It is also shown that by considering the non-uniform temperature effects of the substrate hot-spots on the resistivity of global interconnects, the allocated decoupling capacitances to the hot spot region should be modified accordingly.

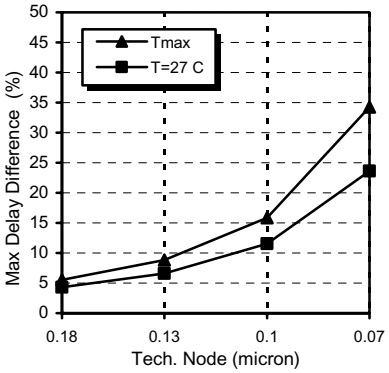

Figure 11: Maximum percentage of the delay difference among drivers connected to a local power trunk for different technologies at room temperature, and at maximum interconnect temperature.

References:

[1] R. Saleh, S.Z. Hussain, S. Rochel, and D. Overhauser, "Clock skew verification in the presence of IR-Drop in the power distribution network," *IEEE Trans. on Computer-Aided Design*, vol. 19, No. 6, 2000, pp. 635-644.

[2] A. Chandrakasan, W.J. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001.

[3] A. Dalal, L. Lev, and S. Mitra, "Design of an efficient power distribution network for the UltraSPARC-I™ microprocessor," *Proc. Int'l. Conf. on Computer Design: VLSI in Computers and Processors*, 1995, pp. 118-123.

[4] X. Tan, C.J.R. Shi, D. Lungeanu, J. Lee and L. Yuan, "Reliability-constrained area optimization of VLSI power/ground networks via sequence of linear programming," *Proc. of Design Automation Conf.*, 1999, pp. 78-83.

[5] T.Mitsuhashi and E.S. Kuh, "Power and ground network topology optimization for cell-based VLSI," *Proc. of Design Automation Conf.*, 1992, pp. 524-529.

[6] M.D. Pant, P. Pant, and D.S. Wills, "On-chip decoupling capacitor optimization using architectural level current signature prediction," *Proc. Intl. ASIC/SOC Conf.*, 2000, pp 288-292.

[7] H.H. Chen and D.D. Ling, "Power supply noise analysis methodology for deep-submicron VLSI chip design," *Proc. Design Automation Conf.*, 1997, pp. 638-643.

[8] S. Im and K. Banerjee, "Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs," *Tech. Digest IEDM*, 2000, pp. 727-730.

[9] R. Chaudhry *et. al.* "Current signature compression for IR-drop analysis," *Proc. of Design Automation Conf*., 2000, pp. 162-167.

[10] S.R. Nassif and J.N. Kozhaya, "Fast power grid simulation," *Proc. of Design Automation Conf*., 2000, pp. 156-161.

[11] P. Chahal, R.R. Tummala, M.G. Allen, and M. Swaminathan, "A Novel integrated decoupling capacitor for MCM-L technology," *IEEE Trans. on Components, Packaging, and Manufacturing*, vol 21-2, pp. 184-193, 1998.

[12] C.S. Chang, A. Oscilowski, and R.C. Bracken, "Future challenges in electronics packaging," *IEEE Trans. on Circuits and Devices*, vol 14-2, pp. 45-54, 1998.

[13] Applications of Metal-Insulator-Metal (MIM) Capacitors, *International SEMATECH*, Technology transfer 00083985A-ENG.

[14] J.R. Black, "Electromigration- A brief survey and some recent results," *IEEE Trans. on Electron Devices*, vol. ED-16, pp. 338-347, 1969.

[15] K. Banerjee, A. Mehrotra, A. Sangiovanni-Vincentelli, and C. Hu, "On thermal effects in deep sub-micron VLSI interconnects," *Proc. Design Automation Conference,* 1999, pp. 885-891.

[16] *International Technology Roadmap for Semiconductors*- ITRS, 2001.

[17] J.C. Anderson, *The use of Thin Films in Physical Investigation*, Academic Press, 1966.

[18] A.H. Ajami, K. Banerjee, M. Pedram, and L.P.P.P. van Ginneken, "Analysis of non-uniform temperature-dependent interconnect performance in high performance ICs" *Proc. Design Automation Conference*, 2001, pp. 567-572.