Scaling Analysis of On-Chip Power Grid Voltage Variations in Nanometer Scale ULSI

AMIR H. AJAMI, [1*+] KAUSTAV BANERJEE[2] AND MASSOUD PEDRAM[3]

[1]Magma Design Automation, 5460 Bay Front Plaza, Santa Clara, CA 95054
[2]ECE Dept., 4151 Eng I, Univ. of California at Santa Barbara, Santa Barbara, CA 93106
[3]EE-Systems Dept., 3740 McClintock Ave, Univ. of Southern California, Los Angeles, CA 90089
Email: [1]amira@magma-da.com, [2]kaustav@ece.ucsb.edu, [3]pedram@ceng.usc.edu

Abstract - *This paper presents a detailed scaling analysis of the power supply distribution network voltage drop in DSM technologies. The effects of chip 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 voltage drop effect in the power/ground (P/G) distribution network increases rapidly with technology scaling, and that using well-known countermeasures such as wire-sizing and/or decoupling capacitor insertion which are typically used in the present design methodologies may be insufficient to limit the voltage fluctuations over the power grid for future technologies. It is also shown that such voltage drops on power supply lines of switching devices in a clock distribution network can introduce significant amount of skew which in turn degrades the signal integrity.*

Key words: technology scaling, power distribution network, voltage drop, de-coupling capacitance, thermal gradient, surface scattering, barrier thickness, clock skew

---

[*] Corresponding Author
[+] This work was done when the author was with the Dept. of EE-Systems, University of Southern California.

1 Introduction

Recent advances in CMOS process technology towards 90nm regime and below have highlighted the importance of signal integrity as one of the main challenges facing today's chip designers. With increasing operating frequencies and elevating power consumptions in VLSI circuits, the design and analysis of on-chip power distribution networks has become a critical design task [1],[2]. Aggressive interconnect scaling has increased the average current density and the resistance per unit length of wires. Since the supply voltage level is also reduced with the technology scaling, the power supply noise becomes even more pronounced because the ratio of the peak noise voltage to the ideal supply voltage level increases with each scaled technology node. The power supply noise mainly manifests itself as a voltage drop in the power distribution networks and can adversely influence the performance of the signal nets, especially the clock distribution networks [2],[3]. Any kind of voltage drop in the on-chip power distribution network should be bounded to the pre-defined device noise-margin limits. An excessive voltage drop in the power grid may result in a functional failure in dynamic logic and a timing violation in static logic. For example, it has been shown that a 10% voltage drop in a design with technology feature size of 0.18 μm, increases the propagation delay of the switching devices by up to 8% [1],[4]. 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 (about 10% of the nominal supply voltage) while satisfying the electromigration (EM) reliability rules for the power network segments and to realize such a power distribution network with minimum routing area of the interconnect metal layers [5],[6],[7].

The effective voltage drop on a chip is attributable to two factors: a) the resistive voltage drop ("on-chip *IR* drop") which is mostly due to the voltage drop due to on-chip interconnect line resistances and b)

the inductive voltage drop (or the *di/dt* noise) which is mostly caused by the pin-package inductances. The *di/dt*, also referred to as simultaneous switching noise (SSN) or ground bounce, is caused by rapid changes in the current passing through the parasitic inductors in the power network. It should be noted that solving for *IR* and *di/dt* drop separately and adding them to determine the worst-case voltage drop tends to be overly pessimistic because the worst-case corner for these two contributing factors seldom occurs. Hence, integrated package and chip level power supply network models are needed to accurately analyze the voltage fluctuations caused by each factor [8]. The main issue in the analysis of power distribution network is the huge size of the problem. Simulating all the nonlinear devices in the chip together with the entire power grid is computationally infeasible. Thus, simulation is usually carried out in two separate steps. First, all devices are simulated with an accurate non-linear simulator assuming a perfect supply voltage for each device. As a result, the current drawn by each device connected to the supply voltage is calculated. Next, devices are modeled as independent time-varying current sources. The problem of power grid analysis is thereby reduced to that of solving a linear resistive network [6],[7]. The error incurred by ignoring the non-linearity of devices is usually negligible after a few iterations [3],[7]. Notice that because the actual voltage supplied to each device is lower than the perfect voltage assumed during the non-linear simulation step, this methodology overestimates the currents drawn by each device, and therefore, tends to overrate the voltage drop. Model of the power grid in this methodology is fairly simple; it comprises of a linear network of resistances that models the power grid interconnect segments and, constant current sources that model the peak current of devices that are connected to the power grid nodes. Numerous methods have been proposed to either determine the peak voltage drop in the power supply network efficiently [3],[5],[7],[9],[10], or to reduce the peak voltage drop in the power distribution network [5],[6],[11].

Wire-sizing is probably the most common method to reduce the overall peak voltage drop by reducing the resistivity of interconnect lines. Although with up-sizing of the widths of power network lines, one should be able to reduce the peak voltage drop, the amount of wire segment up-sizing in the power network is limited by the routing areas that are allocated to the power network in each routing layer (assuming that the EM rules are satisfied) [6]. In addition to the wire-sizing technique, in order to reduce the effect of switching noise on the power distribution network, decoupling capacitors are often added near the switching devices. These capacitors act as local charge reservoirs for switching circuits and reduce the effect of the power supply glitches and ground bounce. Determining the optimal values and locations of the on-chip decoupling capacitors is essential in maintaining a robust power supply network [11],[12]. Similar to the wire-sizing, the portion of the substrate area assigned to the decoupling capacitances is limited and designers should always consider the tradeoff between the reduction of the switching noise and the increase in chip area due to insertion of the decoupling capacitors.

It is well-known that through technology-scaling and as a result of reduction of the minimum interconnect widths, the resistance per unit length of the metal layers increases rapidly. However, additional physical effects such as electron surface scattering and finite barrier thickness contribute significantly to the overall metal resistivity of the local thin lines. In fact these effects become even more pronounced by scaling down of the technology feature size (especially in the Cu interconnects). Furthermore, it has been shown that as the technology feature size is reduced, the peak chip temperatures that occur on the global metal layers increase rapidly due to the self-heating effect [13]. This will cause a further increase in the metal resistivity. Any increase in the metal resistivity would be translated to an increase in the voltage drop in the power distribution network through the IR-drop effect.

Most of the recent research reports on the voltage drop effect have focused on the methodology for efficient computation of the voltage drop values for each gate in the power distribution network [3],[5],[9],[10]. In contrast, this paper studies the effects of technology scaling and temperature on the power supply network voltage drop. It is shown that the voltage drop in the power distribution 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 the present design methodologies 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 voltage drop are studied. It is also shown that how thermal effects due to substrate hot spots on the global interconnect may affect the worst-case voltage 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 Voltage Drop Analysis

Function of the power distribution network is to carry current from the chip power pads to all the cells in the design. In addition, power supply network has many 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** consists of the grid capacitive (including the decoupling capacitances) and inductive terms, *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 the grid nodes. In case of having a simplified *RC* model of the power distribution network, the grid conductance matrix **G** will be a symmetric positive definite sparse matrix. As a result, the system of (1) can be solved very efficiently by various methods like Cholesky factorization or the incomplete Cholesky conjugate gradient technique [14]. When the package model is also included in the analysis, the presence of package pin-to-pad segments necessitates the inclusion of the current through the inductors as variables in the grid conductance matrix **G**. Hence, to solve the system of equations of (1), a more general solution technique,

albeit with much lower computational efficiency, e.g., the LU-decomposition for solving the *RLC* networks, should be used. Figure **1** shows the simplified *RLC* network model which can be used to extract system (1). Time-varying current sources and decoupling capacitances are 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 the device count of the underlying functional blocks on the substrate connected to each grid node. By solving system (1), the voltages of all nodes in the global and semi-global grids become known values.

(FIGURE 1)

In standard-cell based designs, by planning a power trunk adjacent to a cell row, power can be distributed among the cells in that row (Figure **2**). 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 voltage 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 represent both the drain diffusion capacitances 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. To achieve high speed switching devices, the thin-oxide decoupling capacitors should be placed in close proximity of the

highly active switching devices. Assuming that the trunk as a purely resistive network for the time being, given 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 can be 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 voltage drop, one must use the nonlinear voltage-dependent current source 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 have been neglected in deriving (2). Using this model and by solving the linear network matrix coefficient for the power grid through (1), one can derive the voltage drop for the entire network in an iterative manner. The degree of voltage 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 on the grid. As a result, in our experimental setup, we examine the voltage drop in a local power trunk by inserting a 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 assuming a ball grid type of pin assignments in our problem setup, the power pads may be assumed to be uniformly distributed over the chip area (Figure **1**).

(FIGURE 2)

(FIGURE 3)

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 clock cycle time. To calculate the amount of necessary decoupling capacitances in order to maintain a limited voltage drop, one can use the following:

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

where P is the total chip power consumption, V_{dd} denotes the supply voltage, f is the clock frequency, C_T is the effective chip capacitance, and γ is the probability that a 0-1 transition occurs. Assuming a maximum voltage variation of 10%, the computed decoupling capacitance needed for future technologies varies in a range of 39 to 72nF/cm^2 (for 0.18 to 0.07μm, respectively) [15]. Heuristically, the amount of decoupling capacitance needed to accomplish a limited voltage swing (~10%) can be calculated as follows [16]:

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

For a metal-insulator-metal (MIM) capacitor having a dielectric with thickness T_{oxeq}=1nm, the decoupling 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 [17]. These values are presented in Table 1 for different technologies.

3 Methodology of the Power Distribution Network Planning

A key concern for the power supply 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 for VLSI interconnects. When current flows through the interconnect metal, an electric wind is set up opposite to the direction of current flow. These electrons upon colliding with the metal ions, transfer sufficient momentum and cause displacement of the metal ions from their lattice sites, thereby, creating vacancies. These vacancies condense to form voids that result in increase of interconnect resistance or even open circuit condition. The EM lifetime reliability of metal interconnects is modeled by the well-known Black's equation [18], stated as:

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

where *TTF* denotes the time-to-failure (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. Typically, exponent *n* is 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 characteristic goal is to achieve a 10-year lifetime at 100 °C, for which equation (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 account for the interconnect self-heating [19]. Based on the technology roadmap values provided by the ITRS'01 [21], 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 has been shown that interconnects at different metal layers experience different temperatures and global interconnects get hotter than the local interconnect lines [13]. The metal self-heating is mainly due to the interconnect power dissipation caused by the current flow in the interconnect lines. Although interconnect self-heating constitutes only a small fraction of the total power dissipation in the chip, the temperature rise of the interconnect lines due to the self-heating can be significant. This is due to the fact that interconnects are located far away from the substrate and the heat sink, and are separated by several layers of insulating materials that have lower thermal conductivities compared to the substrate. In fact, full-chip thermal analysis using finite element simulations has shown that the maximum temperature in the chip increases rapidly with scaling due to increased self-heating of the interconnects [20], despite the fact that per ITRS'01 roadmap guidelines [21], chip power density (power per unit area) remains nearly constant over a wide range of technology nodes. Based on the values of j_0 and T_{ref} given in ITRS'01 roadmap 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'01 roadmap guidelines [21]. The maximum allowable current density j_m for global/semi-global tiers at the maximum temperature has been calculated using (6) and is provided 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'01 roadmap 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:

$$\#Tracks = \frac{1}{w}\left(\frac{1}{ar} \times \frac{P/V_{dd}}{N_{pad}\, j_{\mathrm{m}}}\right)^{0.5} \qquad (7)$$

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. From Table 2, it can be seen that by going from global to semi-global lines, the power grid gradually gets denser which is expected due to a decrease in the line pitch.

(TABLE 1)

(TABLE 2)

4 Effects of Technology Scaling on the Voltage Drop

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

In VLSI 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 begins to have a considerable contribution to the resistivity compared to the contribution due to the bulk scattering. The surface scattering-governed resistivity ρ of a thin-film metal can be expressed in terms of the bulk resistivity ρ_0 as [22]:

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

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. The dominance of the surface effect depends on the parameter k. 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 [23].

Another effect, which is responsible for increased resistivity, is the presence of a finite cross-sectional area consumed by the higher resistivity metal barrier material which encapsulates the 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 is reduced, or equivalently, the effective resistivity of the metal line of the same drawn

dimension increases. Since barrier thickness cannot scale as rapidly as the interconnect dimension, it would increasingly occupy higher fraction of the interconnect cross section area while restricting the current flow only to the lower resistivity Cu [23]. The effective resistivity and temperature coefficient ratios for the global, semi-global and local tier metal for various technology nodes are given in Table 3.

(TABLE 3)

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 surface scattering and barrier thin-film, interconnect resistance can be re-written as:

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

The incorporation of such technological constraints in addition to the technology scaling effects on metal resistivity, leads to a more realistic line resistance per unit length than the predicted bulk resistivity. 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 should only be considered for the local lines. On the other hand, the global/semi-global tiers are the hottest lines inside the chip [13]. As a result, for global/semi-global lines the effect of line temperature should be considered.

4.2 Voltage 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 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 via or a stack of vias 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 voltage drop in the worst case scenario. 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 will be severe. Ideally, we need to have less than 10% voltage drop in order to ensure correct circuit functionality. Using the minimum-sized tracks will result in a huge and unacceptable 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 total routing area is allocated to the power network [6]. A maximum allocation of 5 to 10% of the routing area to the power network is a common 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. The data in this figure is for the case 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 can be calculated by (4) and is reported in Table 1. It is 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. Figure 5 shows 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 sub 130nm technology nodes, 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)

(FIGURE 5)

(FIGURE 6)

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

4.3 Voltage Drop in Local Power Network

One can impose the worst-case voltage drop figures from the previous section on the two sides of each power trunk to examine the voltage drop in the local power supply network (Figure **3**). Figure 7 shows the

worst-case voltage drop increase as a function of various technology nodes 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. By using data of Figure 5 and Figure 7, Figure 8 summarizes the total voltage drop increase as a function of technology node in the presence of surface scattering, barrier thickness, and temperature effects on interconnect resistivity. Note that the worst-case voltage 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 projected by Figure 5 and Figure 8.

(FIGURE 7)

(FIGURE 8)

4.4 Effect of Hot Spots on the Worst-case Voltage Drop

In reality, the magnitudes of the current sources connected to the power grid are not uniformly distributed over the chip area. 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.

The existence of such hot spots in 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 is thereby exposed to the thermal non-uniformities of the substrate. It has been shown that thermal non-uniformities on the substrate surface may seriously affect the performance of global interconnect lines [25]. Having the power consumption profile of the cells and/or blocks on the substrate, one can easily determine the substrate thermal profile [24]. To derive the thermal profile of a long global interconnect passing over the substrate, one can use the following [25]:

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

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 constants which can be derived by using the physical dimensions of the interconnect line and the insulator and thermo-electrical properties of the interconnect. It is also well known that resistivity of a metal line has a linear relationship with its thermal profile. As a result, due to the non-uniformity of the substrate temperature, resistance profile of the power network would distribute non-uniformly. Specifically, resistances of those segments right above the hot spots are 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 voltage drops at nodes in the proximity of the hot spots become worse. To model a hot spot, a Gaussian distribution thermal profile with a constant peak at T_{max}, i.e. $T(x) = T_{\max} \cdot \exp(-(x-\mu)^2 / 2\sigma^2)$, is assumed. Figure 9 shows the variation of the worst-case voltage drop as a function of the magnitude of thermal gradient of a hot spot over the substrate surface. It is seen that by neglecting the thermal effects of hot spots on the resistivity of the global layers, the worst-case voltage drop of hot devices cannot correctly be predicted, and

consequently, the amount of the inserted decoupling capacitance proposed by existing heuristic rules may be insufficient.

(FIGURE 9)

5 Effect of the Voltage Drop on the 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 [26]:

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

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, with 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 the power supply voltage variations as a function of technology node. As an example, Figure 10 shows 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 technology feature sizes. This delay difference will appear as skew among the devices in a clock circuitry. It is seen that technology scaling can introduce a considerable amount of skew into a clock tree by changing the switching speeds of the clock buffers through non-uniform voltage drop over the power grid network.

(FIGURE 10)

(FIGURE 11)

6 Conclusion

In this paper we highlighted the growing importance of the voltage 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 were taken into consideration for power supply voltage drop analysis. It was shown that barrier thickness and scattering effects must carefully be considered in local interconnect lines, while for global/semi-global tiers, temperature plays an important role in increasing the metal resistivity. Severe performance degradation and/or functional alterations due to power network voltage drop suggests that voltage drop issue is going to become an increasingly important factor in determining the interconnect design policies and signal integrity guidelines. It was shown that new resource allocation guidelines for power supply network metal area and on-chip decoupling capacitors should be provided for future technologies in order to limit the maximum voltage swing in the power networks. It was 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. Finally, it was

observed that non-uniform voltage drops on power supply lines of switching devices in a clock distribution network, introduces a significant amount of skew which in turn degrades the signal integrity.

7 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. Dharchoudhury, R. Panda, D. Blaauw, and R. Vaidyanathan, "Design and analysis of power distribution network on PowerPC microprocessor," *Proc. of Design Automation Conference*, 1998, pp. 738-743.

[4] M. Iwabuchi, N. Sakamoto, Y. Sekine, and T. Omachi, "A methodology to analyze power, voltage drop and their effects on clock skew/delay in early stages of design," *Proc. of Int'l Symposium on Physical Design*, 1999, pp. 9-15.

[5] 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.

[6] 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 Conference*, 1999, pp. 78-83.

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

[8] H.H. Chen, and J. S. Neely, "Interconnect and circuit modeling techniques for full-chip power supply noise analysis," *IEEE Trans. On Components, Packaging, and Manufacturing*, vol 21-3, pp. 209-215, 1998.

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

[10] R. Chaudhry, D. Blaauw, R. Panda, and T. Edwards "Current signature compression for IR-drop analysis," *Proc. of Design Automation Conference*, 2000, pp. 162-167.

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

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

[13] S. Im and K. Banerjee, "Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs," *Technical Digest Int'l Electron Device Meetring*, 2000, pp. 727-730.

[14] D. Luenberger, *Linear and Nonlinear Programming*, 2nd edition, Addison-Wesley Pub. Company, 1984.

[15] 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.

[16] 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.

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

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

[19] 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.

[20] S. Rzepka, K. Banerjee, E. Meusel, and C. Hu, "Characterization of self-heating in advanced VLSI interconnect lines based on thermal finite element simulation," *IEEE Trans. on Components, Packaging and Manufacturing Technology-A*, vol. 21-3, pp. 406-411, 1998.

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

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

[23] K. Banerjee, S.J. Souri, P. Kapur, and K.C. Saraswat, "3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration." *Proc. of the IEEE, Special Issue, Interconnections- Addressing The Next Challenge of IC Technology*, vol. 89-5, pp. 602-633, 2001.

[24] C.H. Tsai and S.M. Kang, “Cell-Level Placement for Improving Substrate Thermal Distribution,” *IEEE Trans. on Computer Aided Design*, vol 19-2, pp 253-265, 2000.

[25] 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.

[26] K.Toh, P. Ko, and R. Meyer, “An empirical model for short-channel MOS devices”, *IEEE. Journal of Solid-States Circuits*, vol23, pp. 2950-957, 1988.

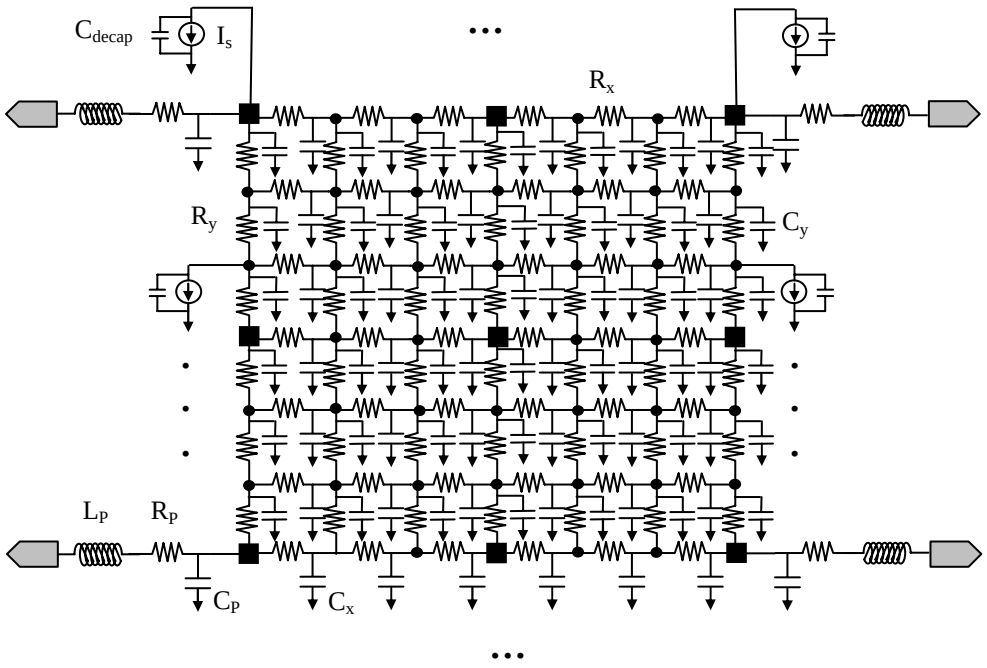

Figure 1: *RC* model of a top-level global/semi-global power grid distribution network. Each intermediate node is connected to underlying circuit blocks modeled as time-varying current sources, I_s's, and on-chip decoupling capacitances C_{decap}'s. The square nodes are the power pad pins connected by to the package pins through the *RLC* model of the package (R_P, L_P, C_P). Assuming a ball grid type of pin assignments in the package, the power pads are uniformly distributed over the chip area in our problem setup.

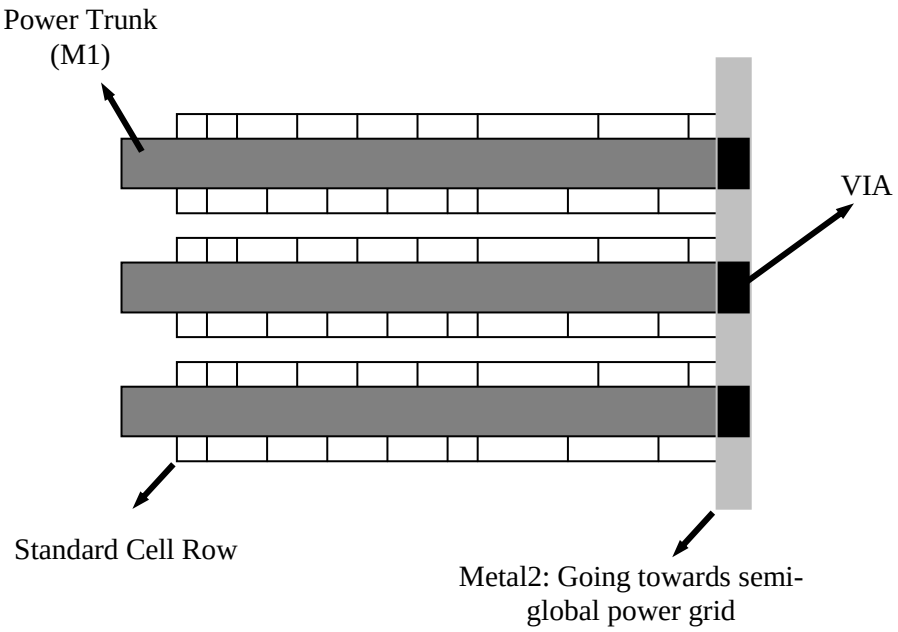

Figure 2**: A local power distribution network for a typical standard cell design consisting of power trunks in a comb-like structure connecting to the semi-global power grid through metal2.**

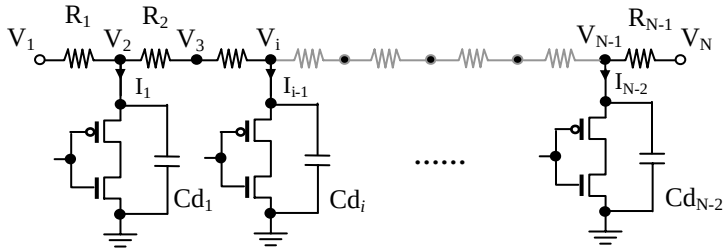

Figure 3**: Magnified view of a trunk segment, containing the resistive network of local power supply network to a series of inverters connected to the intermediate nodes and the decoupling capacitances C_d's.**

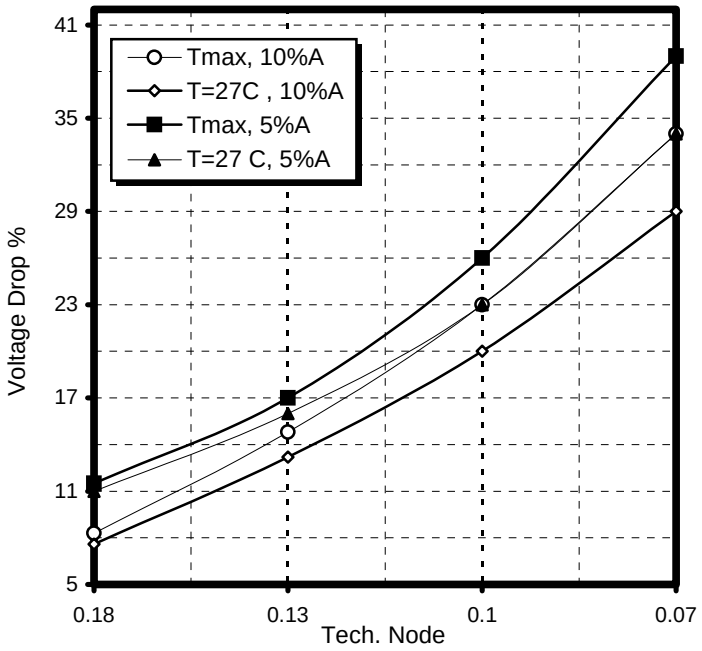

Figure 4: Worst-case percentage of voltage drop increase as a function of technology node for combined global/semi-global power grids considering the effects of self-heating, while allocating 5% and 10% of the routing area to the power distribution network, respectively.

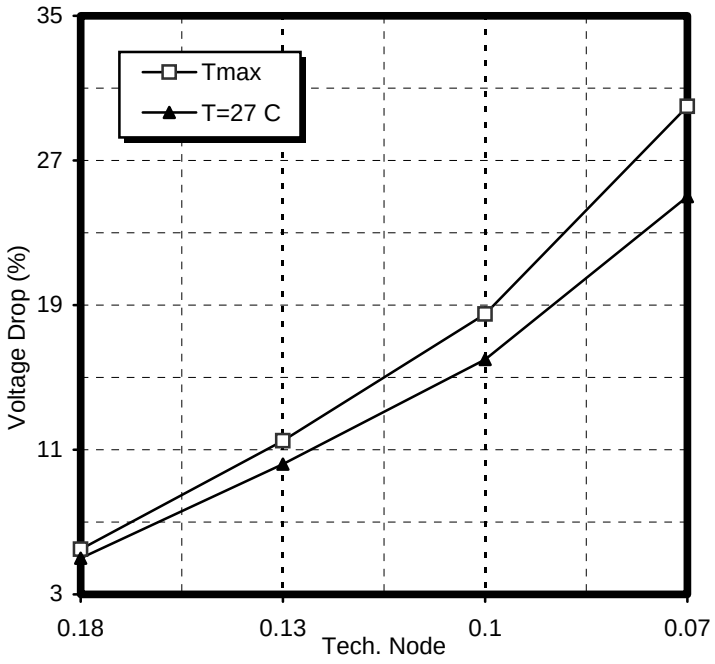

Figure 5: Worst-case percentage of voltage-drop increase as a function of technology node for combined global/semi-global power grids considering the effects of self-heating, while allocating 10% of the routing area to the power network and assuming uniformly distributed on-chip decoupling capacitors.

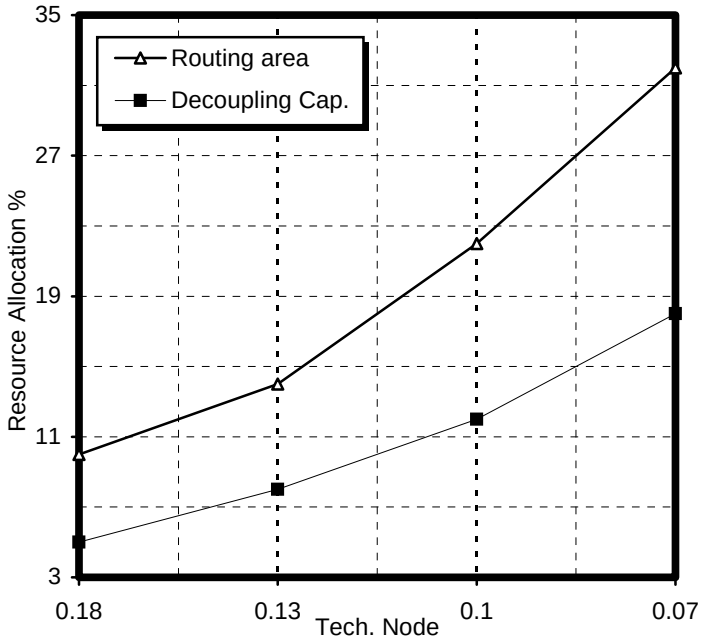

Figure 6: Minimum required percentage of the allocated resources (global layer routing area for wire sizing and substrate area for decoupling capacitances) to ensure a worst-case 10% voltage drop for future technologies, considering the maximum temperature of global/semi-global interconnects.

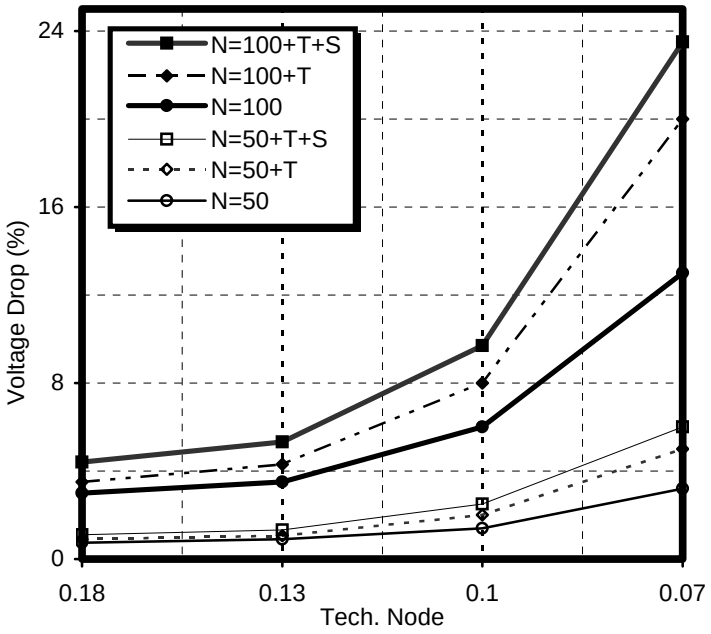

Figure 7: Worst-case percentage of voltage-drop increase as a function of technology node in the presence of maximum interconnect temperature (T) and surface scattering/barrier effects (S), in the local power trunk lines. Notice that in this graph (V_{dd} - $V_{dd'}$) is the actual voltage over the two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

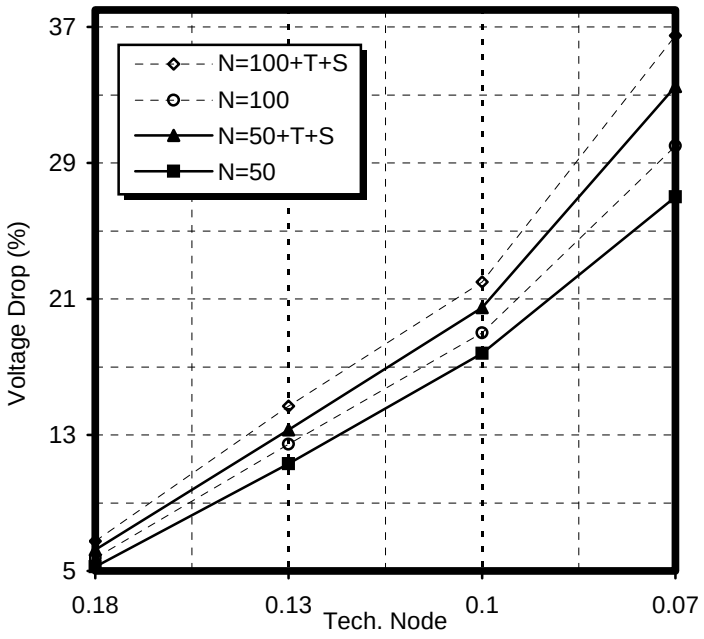

Figure 8: Total worst-case voltage-drop increase as a function of technology node in the presence of maximum interconnect temperature (T) and surface scattering/barrier effects (S), while allocating 10% of the routing area to the power network and 5% of the substrate area for decoupling capacitors.

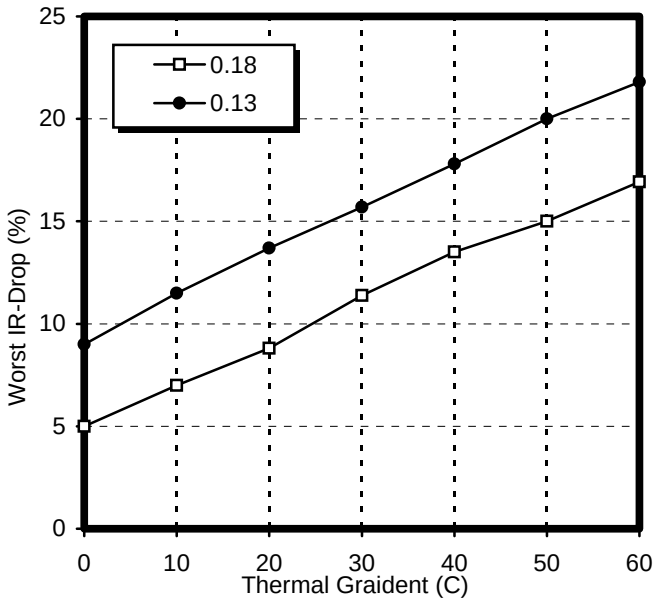

Figure 9: Worst-case voltage-drop ($\Delta V_{IR}/V_{dd}$) increase (based on Figure 5) in the presence of hot spots modeled by constant-peak Gaussian distribution as a function of thermal gradient magnitudes (°C), shown for two different technology feature sizes.

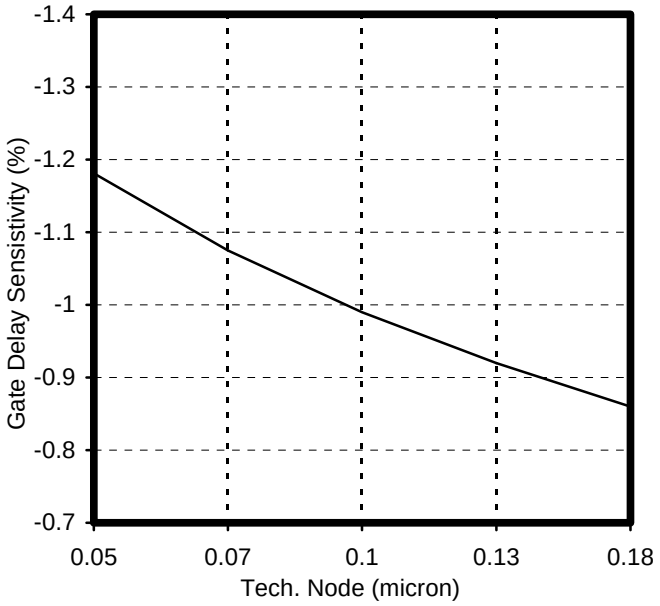

Figure 10: **Sensitivity of the cell delay ($S^{D}V_{dd}$) to the fluctuations of the supply voltage V_{dd} for different technology nodes. Y-axis values show the percentage increase in cell delay for each percent decrease in V_{dd} at the specific technology node.**

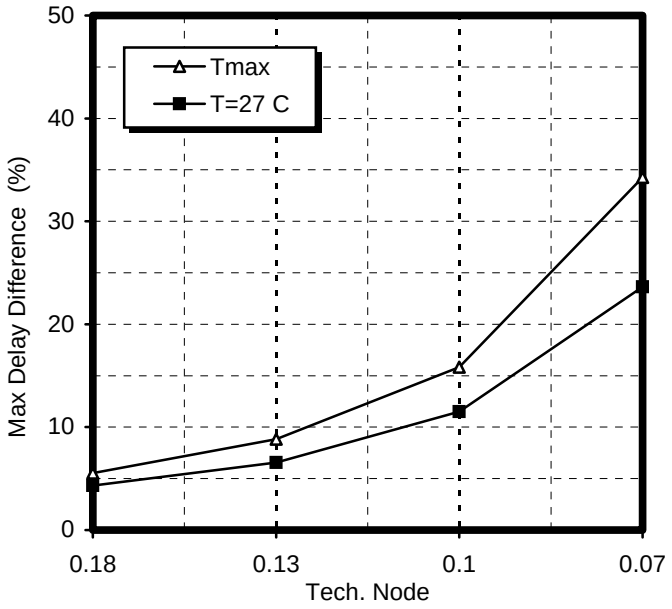

Figure 11: Maximum percentage of the delay difference among drivers in a clock tree connected to a local power trunk for different technologies at room temperature, and at maximum interconnect temperature.

Table 1: **Technology parameters used in this work based on ITRS'01 roadmap data for Cu interconnects.**

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 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}.**

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 3: Effective resistivity ratio, ρ/ρ_0, (barrier thickness plus scattering) and temperature coefficient of resistivity ratio, α/α_0, for the global, semi-global and local tiers for various technology feature sizes.

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