Self-Consistent Thermal-Aware Hierarchical Power/Ground Networks Optimization

Ting-Yuan Wang, Jeng-Liang Tsai, and Charlie Chung-Ping Chen
Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI53706
wangt@cae.wisc.edu, jltsai@cae.wisc.edu, and chen@engr.wisc.edu

ABSTRACT

With the growing power density and heat-dissipation cost of modern VLSI designs, thermal and power integrity quickly become serious concerns. Although the impacts of thermal effects on transistor and interconnects performance are relatively well-known, the interactions between power-delivery and thermal are not clear. As a result, power-delivery design without thermal consideration may cause soft-error, reliability degradation, and even premature chip failures. In this paper, we propose self-consistent thermal-aware hierarchical power-delivery optimization. By simultaneously considering thermal and power integrity, we are able to achieve high power supply quality and thermal reliability. Furthermore, to deal with the large scale power-delivery optimization issue, we propose hierarchical optimization which effectively reduce the runtime and memory consumptions. Experimental results demonstrate the efficiency and optimality of our approach to produce high quality self-consistent power-delivery solution for chip-level optimization tasks.

1. INTRODUCTION

The ever-increasing demand of more functionality and higher speed has pushed the VLSI industry toward more aggressive scalings. Since this trend leads to high current density and power consumption, low-power design has become more and more important. However, the relentless push for low-power has been directed to the decrease of supply voltage which raises the maximum current density on a chip even higher. To ensure the quality of power-delivery, advanced technologies such as using new organic packaging materials, flip-chip design, and C4 bump have been developed to tackle the high current density problem. The techniques for high-quality on-chip power-delivery needs to be significantly improved in order to fulfill future requirements.

Traditional P/G network design methodologies aim at minimizing the total wire-area subject to a current-density (electromigration reliability) constraint and a voltage-drop (power-dip/ground-bounce) constraint. Two main physical-design approaches are available for improving the quality of power-delivery. Wire-sizing has been shown to be an efficient way of reducing power-dip and ground-bounce as well as improving the reliability of P/G networks [5] [6] [4] [7] [16] [17] [18]. Topology-optimization is another technique which adjust the power-delivery network topology to fit the current-supply pattern [12] [3] [15].

All the literature except [13] assume that currents drawn from each block are constant. Under this assumption, the result of topology-optimization usually comes out to be a tree topology that gives a minimal area [8]. A tree topology is less favorable in practice because it is not robust with respect to current variations. A new model which assumes the currents drawn from blocks are variables was proposed in [13]. The model assumes the first and second moments of the block currents as well as their correlation matrices are known. The objective of the optimization problem was the average power dissipated in the P/G network subject to an area constraint. This method finds a solution which is more reliable and the area is also small. An important conclusion of this work is that the optimal topology of P/G networks under reliability constraints is not a tree structure, and the RMS current density of the optimal topology in each wire is the same.

Recent studies showed that the aggressive scaling causes nonuniform temperature distribution on substrate and interconnects, which could jeopardize the chip reliability and performance. For a common flip-chip design, interconnect layers are farther from the heat sink than the substrate does, and Joule heating which is proportional to the square of RMS current density can cause high temperature on interconnects. Therefore, an optimistic design which allows maximum current density is not sufficient because temperature also affects electromigration.

The lifetime of an interconnect which is determined by electromigration effect is exponentially proportional to the inverse of interconnect temperature. In order to maintain the same interconnect lifetime under high temperature, the maximum allowable average current density need to be decreased. Therefore, a temperature constraint is also needed to account for the Joule heating effect. Hunter gives a self-consistent solution by comprehending both electromigration effect and joule heating [10].

In this paper, we propose a hierarchical approach to solve the P/G optimization problem under thermal considerations. The hierarchical strategy optimizes global P/G net-

DAC 2003 Anaheim, California, USA

works and local P/G networks separately. First, maximum allowable current densities that satisfy both thermal and electromigration based on the full-chip thermal profile are derived, which are then used to optimize global P/G networks. The temperature variation in a local P/G network is usually less significant and the highest temperature in the region is assumed. Under uniform temperature assumption, the optimization problem can be formulated as a convex problem which can be solved more efficiently.

The remainder of the paper is organized as follows. In Section 2 thermal effects and reliability issues are discussed; in Section 3 the P/G network design rules with thermal consideration are provided; Section 4 is our hierarchical P/G network optimization methodology, followed by the experimental results in Section 5.

2. RELIABILITY AND THERMAL EFFECTS

Thermal effects are inseparable aspects of electrical power distribution and signal transmission through the interconnects in VLSI circuits due to self-heating caused by the flow of current [1]. This results in a higher temperature in P/G network than the temperature in the substrate. Therefore, in the P/G network optimization design the current density constraint of EM effect is not sufficient without considering the temperature factor because EM is exponentially proportional to the inverse of temperature.

2.1 Electromigration

Electromigration is the transport of mass in metals under the stress of high current density. This metallization failure is the main reliability concern of high performance IC designs. The lifetime of metal interconnects is modeled by Black's equation [2]:

$$MTF = A\frac{1}{J_{avg}^2}exp(\frac{E_a}{k_B T_m}) \quad (1)$$

where MTF is the mean-time-failure, A is a constant which depends on the geometry, J_{avg} is the average current density, E_a is the activation energy, and k_B is Boltzman's constant.

If the goal of design is to achieve ten years of operation lifetime under the average current density J_o at temperature T_o, the following restriction must be satisfied :

$$\frac{exp(\frac{E_a}{k_B T_m})}{J_{avg}^2} \geq \frac{exp(\frac{E_a}{k_B T_o})}{J_o^2}. \quad (2)$$

It is observed that if the metal temperature increases by self-heating, the allowable average current density is forced to decrease in order to maintain target lifetime.

2.2 Self-Heating

Current passes through metal wires cause power dissipation $I_{rms}^2 R$ is demonstrated in Figure 1. Because the global interconnects are away from the substrate, the generated heat can not spread efficiently to the heat sinks and hence the temperature increases. This phenomena is referred to as self-heating or Joule-heating and has become more serious due to the introduction of the low-k material and the increasing number of metal layers.

For a thermally-long interconnect as shown in Figure 1, the width, length, and thickness of the metal are w_m, l_m, and t_m respectively. The thickness and the thermal conductivity of the underlying insulator are t_{ins} and κ_{ins}. The

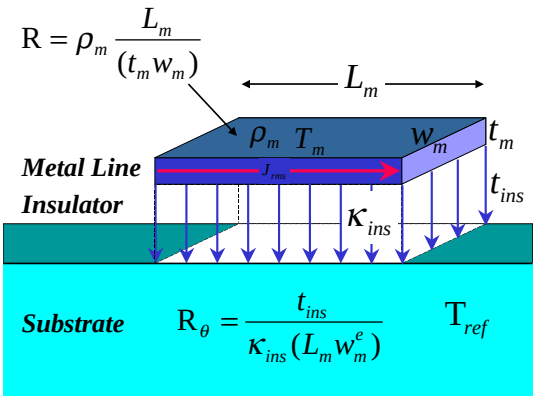

Figure 1: Structure of an Interconnect on a Chip

wire resistivity is temperature dependent and is described as follows:

$$\rho_m(T_m) = \rho_o[1 + \beta(T_m - T_{ref})] \quad (3)$$

where T_m is the metal temperature, T_{ref} is the reference (substrate) temperature, ρ_o is the wire resistivity at temperature T_{ref}, and β is the temperature coefficient of resistivity.

The temperature increase in wires can be expressed as:

$$\Delta T_{SH} = T_m - T_{ref} = I_{rms}^2 R R_\theta = Q R_\theta \quad (4)$$

where $Q = I_{rms}^2 R$ is the heat flow, $R = \rho_m l_m / w_m t_m$ is the interconnect resistance, and R_θ is the equivalent thermal resistance from the metal wire to substrate.

In the steady state metal temperature increase is approximated by a quasi-one dimensional heat transport model [14] and we have:

$$R_\theta = \frac{t_{ins}}{\kappa_{ins} l_m w_m^e}. \quad (5)$$

where w_m^e is the effective thermal width and can be approximated by

$$w_m^e = [1 + 0.88\frac{t_{ins}}{w_m}]w_m. \quad (6)$$

It is accurate to 3% with $w_m/t_{ins} > 0.4$. The effective thermal width w_m^e is always greater than w_m, and approaches w_m when $w_m \gg t_{ins}$.

Next section will discuss the interaction between EM and self-heating.

3. POWER/GROUND DESIGN RULES FOR THERMAL EFFECTS

In order to have higher current-density on interconnects, the temperature tolerance need to be increased. However, temperature increase degrades the reliability of wires which requires EM current density to decrease. Therefore, there is a maximum solution of current density called self-consistent solution satisfying both effects [10] [11].

3.1 Self-Consistent Solutions

Currents on P/G interconnects are usually unipolar DC waveforms with duty cycle r. The average current density is related to peak current density by $J_{avg} = rJ_{peak}$, and the

RMS current density can be expressed as $J_{rms} = \sqrt{r} J_{peak}$. Therefore we have the following relation between J_{avg} and J_{rms}

$$\frac{J_{avg}^2}{J_{rms}^2} = r. \tag{7}$$

Now the rms current density in self-heating, Eqn. (4) and Eqn. (5), can be rewritten as

$$J_{rms}^2 = \frac{\Delta T_{SH} \kappa_{ins} w_m^e}{t_{ins} t_m w_m \rho_m}. \tag{8}$$

Therefore, the consistent solutions of the EM reliability and self-heating effect can be computed by substituting J_{avg}^2 in Eqn. (2) with equality and J_{rms}^2 in Eqn. (8) into Eqn. (7). We have

$$r = \frac{J_{avg}^2}{J_{rms}^2} = J_o^2 \frac{exp(\frac{E_a}{k_B T_m})}{exp(\frac{E_a}{k_B T_o})} \frac{t_{ins} t_m w_m \rho(T_m)}{(T_m - T_{ref}) k_{ins} w_m^e}. \tag{9}$$

This equation means that the self-consistent temperature T_m can be obtained for the given r. Once the the self-consistent temperature T_m is obtained, the corresponding J_{peak} can also be obtained either from Eqn. (4), or Eqn. (2) with equality. An example of the self-consistent solution for T_m and J_{peak} is shown in Figure 2 [10].

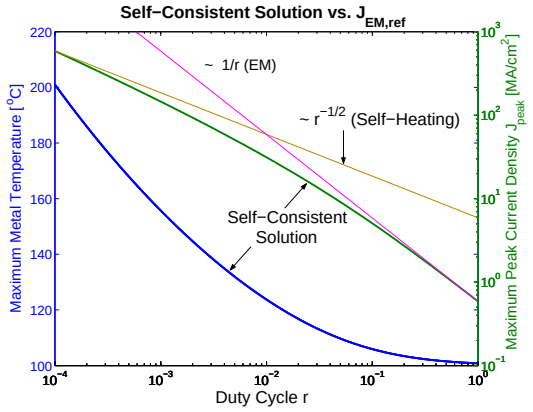

Figure 2: The self-consistent solution for the maximum allowed temperature and current density.

3.2 Design Rules for Choosing T_m and J_{peak}

The rules of choosing wire current density is illustrated as follows. Figure 3 shows the graphical solution of J_{peak} and T_m for EM and self-heating effects for the given duty cycle. Suppose the substrate is 100 ^oC and the duty cycle is 0.01, the estimated peak current density J_{peak} without considering self-heating is at point A. However, the maximum J_{peak} of self-heating which causes the temperature increase to 100 ^oC is at point D. We can see that the J_{peak} at D is much smaller than the J_{peak} at A, thus J_{peak} at D is the solution satisfying both EM and self-heating constraints. If the tolerable temperature in interconnects increases from point D to B, J_{peak} of self-heating which satisfies both constraints increases. The maximum J_{peak} solution is at point B. From point B to E, temperature on interconnects is too hot and electromigration becomes the dominating effect. For the given duty cycle, the self-consistent solution is

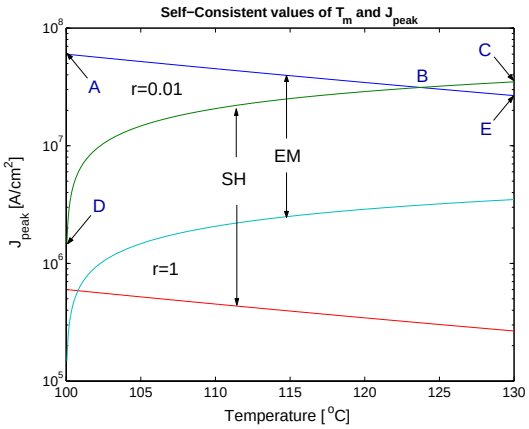

Figure 3: Graphical solution of J_{peak} and T_m for electromigration and self-heating effects.

the curve $D - B - E$. In this paper, we propose the thermal constraint which satisfy both EM and self-heating effects.

- **The thermal constraint**
  The thermal constraints give the current density which simultaneously satisfy EM and self-heating effects.

$$| J_k | \leq J_{self-consistent} \tag{10}$$

Suppose that the maximum current density in wires is about $1MA/cm^2$ at $105^oC$for $0.13\mu m$ according to ITRS [9], and the target operating lifetime is ten years. The initial wire width and duty cycle are also given. The maximum current density and temperature on interconnects for a given substrate temperature profile can be calculated. This solution is at the point B as shown in Figure 3. Without considering the temperature, the designer may overestimate the maximum allowable current-density that could lie on region $A - B$, or underestimate the maximum current-density that is on $B - E$. However, this solution is based on the initial wire widths which give an optimistic estimation. The solutions of wire-widths and current density of the P/G network wire-sizing problem may not meet the ten-year reliability requirement. The wire-widths can be used to calculate the more accurate self-heating curve $D - B - C$ and a new set of self-consistent values is used to optimize the P/G network again. Several iterations may be required to get a correct and area-saving solution.

4. HIERARCHICAL POWER/GROUND NETWORK OPTIMIZATION

In this section we first introduce general constraints of P/G network optimization. To ensure the reliability of the final network and reduce the runtime for the optimization, we explore the characteristic of the thermal profile and use different formulations for global and local P/G networks.

4.1 General Power/Ground Network Optimization

Figure 4 shows an example of a grid-based ground network with four ground pads connect to its four corners. A P/G network $G = \{N, B\}$ consists of n nodes $N = \{1, ..., n\}$ and b branches $B = \{1, ..., b\}$. A time dependent current source

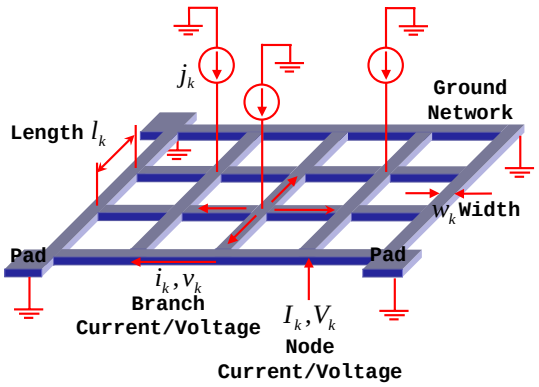

Figure 4: Ground Network

drains/injects a current from/to G which models a circuit block near the vicinity. The currents are assumed to be time dependent and the current variations are slow, i.e., the dynamic effects caused from capacitances and inductances are ignored. Each wire is modeled as a resistor. Therefore, the P/G network forms a very large scale linear network. The current and voltage in branch k are denoted as i_k and v_k. The nodal voltage at node k is V_k, and the external current source drawn from the block at node k is denoted I_k^e. The conductance of the k branch is $g_k = w_k/\rho_s l_k$, where ρ_s is the sheet resistivity, w_k and l_k are the width and length of branch k. The current density in branch k is $j_k = i_k/w_k$.

The network behavior is described by a set of nodal equations $G_n V_n = I_n$, each branch element is expressed as $G_b V_b = I_b$, where $V_b = [v_1, \ldots, v_k, \ldots, v_b]^T$ is the vector of branch voltage, $V_n = [V_1, \ldots, V_k, \ldots, V_n]^T$ is the vector of nodal voltage, $I_b = [i_1, \ldots, i_k, \ldots, i_b]^T$ is the vector of branch current, $I_n = [I_1^e, \ldots, I_k^e, \ldots, I_n^e]^T$ is the vector of nodal current source, $G_b = diag(g_1, \ldots, g_b)$ is the branch conductance matrix, and G_n represents the nodal conductance matrix. The relation between G_n and G_b is $G_n = AG_bA^T$, where A is the incidence matrix which implies the KCL and KVL.

- **Objective function** A general objective function of P/G network optimization is

$$f_A = \sum_{k \in B} l_k w_k. \tag{11}$$

Due to the reliability and technology process requirements, the objective function is subjected to the following constraints:

– **Voltage-drop constraint**
IR-drop is the voltage fluctuation due to the resistance of the on-chip P/G network, which may cause timing uncertainty and affect performance. For example, the voltage fluctuation from ground pads to the leaf nodes must be restricted with an upper bound:

$$V_{i \in N^e} \leq V_{max}, \tag{12}$$

where N^e is the set of leaf nodes where the external currents are drawn from block circuits.

– **Minimum-width constraint**
The minimum-width of wires is determined by the technology and is formulated as:

$$w_{i \in B} \geq w_{min}. \tag{13}$$

– **Electromigration constraint**
Traditional method uses a single upper bound of average current density for every wire and the constraint is as follows:

$$|\ (j_k)_{avg}\ | = \frac{|\ (i_k)_{avg}\ |}{w_k} \quad \leq \quad (j_{max})_{avg}, \tag{14}$$

where $(j_{max})_{avg}$ is the maximum allowed average current density.

– **Thermal constraint**
The thermal constraint proposed in 3.2 gives each wire a maximum current density which simultaneously satisfies EM and self-heating constraints.

$$|\ j_k\ | \quad \leq \quad j_{k,self-consistent} \tag{15}$$

4.2 Hierarchical Power/Ground Design

Due to the significant increase of transistor number, P/G networks with millions of wire segments are commonly seen in ASIC designs. Therefore, a hierarchical approach is more suitable for solving such a large scale problem. We propose a hierarchical P/G network design methodology which is illustrated in Figure 5. In our design methodology the global P/G network is grid-based, and there is only one point connecting a local network to the global mesh. First we split the budget among global and local P/G networks and optimize global P/G network first. After global P/G network optimization is done the grid voltages which acts as the ground voltages of the local P/G networks are found. We then take the rest of the IR-drop budget for local P/G networks and perform local optimizations. Different optimization strategies and problem formulations for global and local P/G networks are used to further increase the optimization speed.

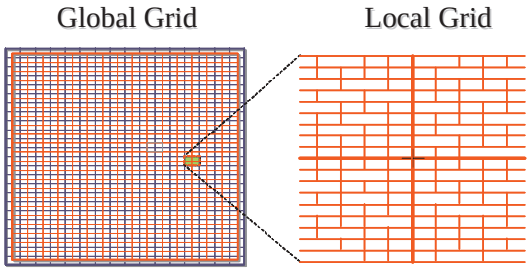

Figure 5: Illustration of the hierarchical P/G networks optimization strategy.

4.3 Optimize Global Power/Ground Networks

Figure 6 shows the thermal profile of a $48W$ $11.3mm \times 14.4mm$ chip. As seen in the figure, substrate temperature varies from 30 ^oC to 135 ^oC and we must impose the thermal constraint to ensure the reliability. The global network optimization problem is to minimize the total wire area, Eqn. (11), subject to Eqn. (12), Eqn. (13), and Eqn. (15).

The self-consistent current densities cannot be known without knowing wire-widths and an iterative method is used. First we assume all wires are at minimum-width and obtain

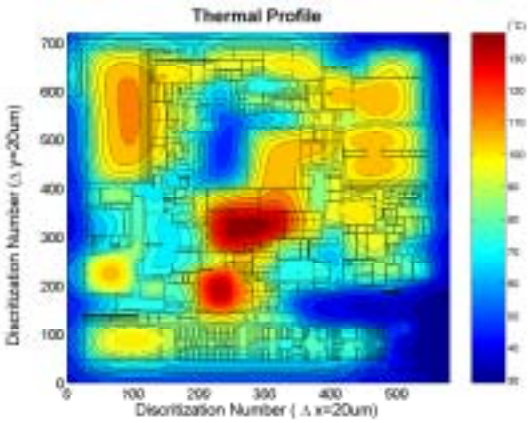

Figure 6: The full-chip temperature profile.

the self-consistent current density for each wire, and solve the optimization problem. Wire-widths are taken back to re-calculate correct self-consistent current densities. If the RMS current densities after optimization are all below these new self-consistent current densities, MTFs of all wires meet the design target. If not, the self-consistent current densities of (15) are updated and the optimization process are executed again. Note that usually only a small number of segments will violate the constraint, and costs of following iterations are low.

4.4 Optimize Local Power/Ground Networks

Observe in Figure 6 that sizes of hot spots are usually large compare with sizes of local P/G networks, thus we assume the temperature on a local P/G network to be the highest temperature in its region without paying too much area penalty for the pessimistic assumption. A convex formulation of the P/G optimization strategy was proposed and we make some changes for our design methodology. We assume local P/G stripes are required as shown in Figure 5 but vertical segments are allowed to be removed. If circuit elements in a local network draw a large current from the network, it requires more vertical segments to make the network robust. For low power demand blocks, we need fewer vertical segments and the routing resource can be allocated to signal nets. In [13] the objective function of the P/G networks optimization problem was reformulated as

$$\begin{aligned} mim: \quad & f_{power} + \mu f_{area} \\ s.t. \ : \quad & w_k \geq 0, \end{aligned} \tag{16}$$

where $f_{power} = E[I_n^T G_n^{-1} I_n] = \sum I_{rms}^2 G_b^{-1}$ is the power consumption on interconnects, and μ is the power density. The objective function has been shown to be a differentiable convex function of w. The optimal solution has the property that each wire is either zero width or has the same constant rms current density $(j_k)_{rms} = \sqrt{(\mu/\rho)}$. We can replace $(j_k)_{rms}$ with the self-consistent current density and set $\mu = \rho j_{self-consistent}^2$ in the objective function and the optimal solution can be found efficiently.

The IR-drop budget from the P/G pads to the local networks leaf nodes has been determined in the global P/G optimization phase. However, the minimum-width constraints and the IR-drop constraints are not included in the problem. Therefore, the method to compensate the violations is described as follows. If the nodal voltage V_k violates the voltage drop constraint V_{max} by a factor $\lambda = V_k/V_{max}$ or the width w_k violates the minimum-width constraints by a factor $\lambda = w_{min}/w_k$, the new solution can be obtained by setting μ as $\frac{\mu}{\lambda^2}$.

5. EXPERIMENTAL RESULTS

We use the same set of current sources that generates the thermal profile for our ground network optimization. We assume that wire segments are $400\mu m$ in the global network and $20\mu m$ in the local network. The size of the global network is 29×36, and 20×20 in a local network. The wire-widths of the global network is from $1\mu m$ to $40\mu m$ and $0.3\mu m$ to $3\mu m$ in the local network. The unit square resistance is 0.022Ω and the current density on an interconnect with 10 years MTF is $10^6 A$ per square at 105 ^oC. Global network optimization can be completed in two hours and the average runtime for local network optimization is about 10 minutes.

Figure 6 shows the optimized ground network with only electromigration constraints. The marked segments are those have actual MTF less than 10 years and the histogram of the 100 least reliable segments are in Figure 8. The lowest MTF is only 5.9 years while the original design goal is 10 years. It is not hard to see by comparing Figure 6 and Figure 7 that those premature failures tend to happen at hot spots of the chip. Without considering thermal effects, the lifetime of a chip can be 40% less than the original design specification! Observe Figure 8 we find many wire segments

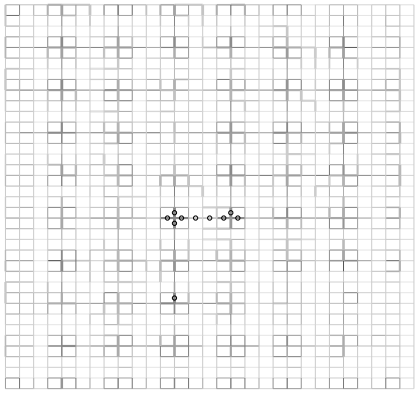

Figure 7: Global network.

are potentially over-designed. If the voltage-drop constraint is not dominant, a thermal-aware design methodology can relax the current density constraint at cooler areas and saves area. Figure 9 shows an optimized local network with low current flow. For a low current flow network voltage-drop constraints are usually not dominant, and only the segments near the center of the network will carry large currents. The result shows that for these networks a significant amount of wire segments can be removed to save routing resources. Figure 10 shows an optimized local network with high current flow. In this case voltage-drop constraints as well as current density constraints affect the topology. However, there are still 15% segments that can be removed.

6. REFERENCES

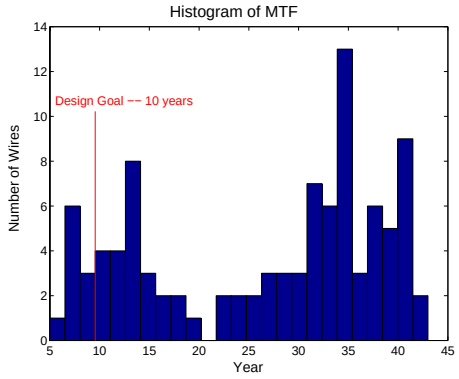

Figure 8: The MTF histogram of the 100 least reliable segments without thermal constraints

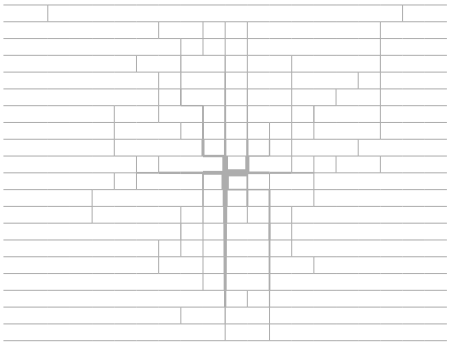

Figure 9: An optimized local network with low current flow

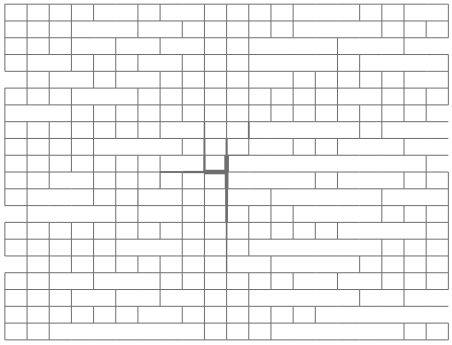

Figure 10: An optimized local network with high current flow

[1] K. Banerjee and A. Mehrotra. Global (interconnect) warming. *IEEE Circuits and Devices Magazine*, 17(5):16–32, September 2001.

[2] J. R. Black. Electromigration–a breief survey and some recent results. *IEEE Trans. on Electron Devices*, ED-16(4):338–347, April 1969.

[3] H. Cai. Multi-pads single layer power net routing in vlsi circuit. In *25th ACM/IEEE Design Automation Conference*, pages 183–188, 1988.

[4] S. Chowdhury. Optimum design of reliable ic power networks having general graph topologies. In *26th ACM/IEEE Design Automation Conference*, pages 787–790, 1989.

[5] S. U. Chowdhury and M. A. Breuer. Minimal area design of power/ground nets having graph topologies. *IEEE Trans. on Circuits and Systems*, CAS-34(12):1441–1450, December 1987.

[6] S. U. Chowdhury and M. A. Breuer. Optimum design of ic power/ground nets subject to reliability constraints. *IEEE Trans. on Computer-Aided Design*, 7(7):787–796, July 1988.

[7] R. Dutta and M. Marek-Sadowska. Automatic sizing of power/ground (p/g) networks in vlsi. In *26th ACM/IEEE Design Automation Conference*, pages 783–786, 1989.

[8] K.-H. Erhard and F. M. Johannes. Power/ground networks in vlsi: are general graphs better than trees? *Integration, the VLSI journal*, 14:91–109, 1992.

[9] http://public.itrs.net. International technology roadmap for semiconductors 2001 editiion.

[10] W. R. Hunter. Self-consistent solutions for allowed interconnect current dsnsity – part i: Implications fortechnology evolution. *IEEE Trans.Electron Devices*, 44(2):304–309, Feburary 1997.

[11] W. R. Hunter. Self-consistent solutions for allowed interconnect current dsnsity – part ii: Application to design guidelines. *IEEE Trans. Electron Devices*, 44(2):310–316, February 1997.

[12] J. Oh and M. Pedram. Multi-pad power/ground network design for uniform distribution of ground bounce. In *35th ACM/IEEE Design Automation Conference*, pages 287–290, 1998.

[13] A. G. S. Boyd, L. Vandenberghe and S. Yun. Design of robust global power and ground networks. In *ISPD '01*, pages 60–64, Sonoma, California, USA, April 2001. ACM.

[14] H. A. Schafft. Thermal analysis of electromigration test structures. ED-34(3):664–672, March 1987.

[15] Z. A. Syed and A. E. Gamal. Single layer routing of power and ground networks in integrated circuits. *Journal of Digital Systems*, 6(1):1441–1450, 1982.

[16] X. Tan, C. J. R. Shi, D. Lungeanu, and L. Y. J. Lee. Reliability-constrained area optimization of vlsi power/ground networks via sequence of linear programmings. In *36th ACM/IEEE Design Automation Conference*, pages 78–83, 1999.

[17] X.-D. S. Tan and C.-J. R. Shi. Fast power/ground network optimization based on equivalent circuit modeling. In *38th ACM/IEEE Design Automation Conference*, pages 550–554, 2001.

[18] T.-Y. Wang and C. C.-P. Chen. Optimization of the power/ground network wire-sizing and spacing based on sequential network simplex algorithm. In *2002 International Symposium on Quality Electronic Design*, pages 157–162, 2002.