Power-Delivery Networks Optimization with Thermal Reliability Integrity[*]

Ting-Yuan Wang and Jeng-Liang Tsai
Electrical and Computer Engineering
University of Wisconsin-Madison
1415 Engineering Drive
Madison, WI 53706
{wangt, jltsai}@cae.wisc.edu

Charlie Chung-Ping Chen
Graduate Institute of Electronics Engineering &
Department of Electrical Engineering
National Taiwan University
Taipei 106, Taiwan
cchen@cc.ee.ntu.edu.tw

ABSTRACT

With the growing power consumption in modern high performance VLSI designs, nonuniform temperature distribution and limited heat-conduction capability have caused thermal induced performance and reliability degradation. Electromigration is the main reliability concern and will become a more limiting factor of IC designs. It must be addressed together with a thermal reliability modeling. This issue also has been recognized in the International Technology Roadmap for Semiconductors (ITRS) 2002 update as one of the difficult challenges [1]. Although the impacts of thermal effects on transistor and interconnect performance are well-studied, but still how thermal effects affect the reliability of power delivery is not very clear. As a result, traditional power-delivery designs without thermal consideration may cause soft-error, reliability degradation, and even premature chip failures. In this paper, we propose an algorithm for power-delivery networks optimization with thermal reliability integrity. By considering thermal and power integrity, we are able to achieve high power supply quality and thermal reliability. For a 56×72 mesh, our design shows that the lifetime of the optimized ground network is 9.8 years. Whereas the lifetime of the ground network designed by a traditional method without thermal integrity is only 4.1 years.

Categories and Subject Descriptors: J.6 [Computer-Aided Engineering]: Computer-aided design (CAD)

General Terms: Algorithm, Design, Reliability

Keywords: Power/Ground, Electromigration, Self-heating, Self-consistent Constraint, Optimization, Thermal, Reliability

1. INTRODUCTION

The aggressive scaling trend has been pushed to satisfy the demands for more functionality and higher speed in VLSI designs. Since this trend leads to higher power consumption, low-power design has become more and more important. However, the relentless push for low-power design has been directed towards the decrease of supply voltage which reduces noise margin. Power delivery noise is now becoming a crucial factor in determining the performance and the reliability of VLSI designs. It has been shown that a 10% voltage drop in a $0.18\mu m$ technology can increase the propagation delay of the gate by up to 8% [2]. The techniques for high-quality on-chip power-delivery design needs to be significantly improved to fulfill more strict requirements.

Power/Ground (P/G) distribution systems are designed to provide needed voltages and currents to the transistors that perform the logic functions of a chip. Due to the cost, reliability, and performance issues, traditional P/G networks design methodologies aim at minimizing the total wire area subject to an electromigration (EM) constraint and an IR-drop 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 IR-drop noise and improving the reliability of P/G networks [3] [4] [5] [6] [7] [8] [9] [10] [11] [12]. Topology-optimization is another technique which adjusts the power-delivery network topology to fit the current-supply pattern [13] [14] [15]. The major contributions of these papers are based on the improvement of area and computational speed.

In modern high speed VLSI designs, nonuniform temperature distribution on chips has become more and more serious concern [17] [18] [19]. Without thermal management, thermal problems not only lead to timing failures but also degrade chip reliability. Therefore, the constraints of traditional P/G distribution design which limit maximum current density only depending on EM effect are not sufficient and may be too optimistic. The influence of temperature on EM effect needs to be considered. For example, the current density of a wire, which satisfies the maximum current density limited by EM effect at a given temperature, has the ability to drive the temperature up because of self-heating (SH). The maximum current density needs to be decreased to satisfy the required lifetime. Therefore, a new constraint which integrates both EM and SH effects is needed.

The objective function of traditional P/G networks optimization methods is to minimize the total wire area. Smaller

[*]This work was partially funded by National Science Foundation under grants CCR-0093309 & CCR-0204468 and National Science Council of Taiwan, R.O.C., under grant NSC 92-2218-E-002-030.

ISPD'04, April 18–21, 2004, Phoenix, Arizona, USA.
Copyright 2004 ACM 1-58113-817-2/04/0004 ...$5.00.

the wire area, larger the power consumption in the network. Larger power consumption results in higher temperature in networks. It means that traditional design methods minimizing total wire area may cause the degradation of chip reliability and timing failure in hot spots. To cope up with the thermal and power integrity issues, the objective function based on the trade-off between power consumption and wire area in networks is needed.

In this paper, we propose a thermal-aware algorithm for P/G networks design. First, a new *self-consistent constraint* is defined and used to replace the EM constraint for the thermal reliability concerns. This self-consistent constraint is based on the idea of finding the current density which satisfies both EM and SH effects. This approach is to avoid the unexpected reliability failures in hot spots. Second, the objective function is based on minimizing the sum of each wire's weighted sum of average power dissipation and wire area. This approach is to address power and thermal integrity issues. The formulated optimization problem is convex and results in more reliable P/G structure with comparable wire area. This approach can solve the reliability problem of wires in hot spots by giving lower current density and wider wire widths.

The remainder of the paper is organized as follows. The problem formulation is presented in Section 2. In Section 3, thermal effects and reliability issues are discussed as well as the self-consistent constraint is defined. Section 4 presents the algorithm for thermal-aware P/G networks design. Section 5 shows the experimental results, followed by conclusion in Section 6.

2. PROBLEM FORMULATION

An example of a grid-based ground network with four ground pads connected to its four corners is shown in Figure 1. Suppose that a P/G network $G = \{N, B\}$ consists of m non-ground nodes $N = \{1, ..., m\}$ and n branches $B = \{1, ..., n\}$. The dynamic effects caused by capacitances and inductances are ignored here. The power sources are modeled as simple constant voltage sources, and the power drains drawn from circuit to G are modeled as independent time-varying current sources. The first and second moments of the current variables, as well as the correlations between the currents, are given. The metal wires and vias are modeled as resistors. Therefore, the P/G network forms a very large scale linear resistive network excited by constant voltage sources and time-varying current sources.

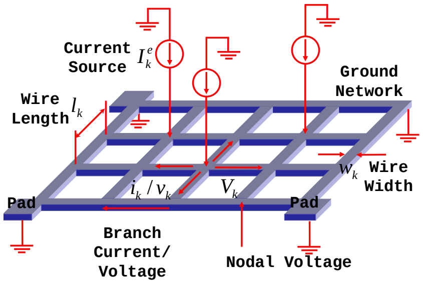

Figure 1: A ground network.

The current and voltage drop in branch k are denoted i_k and v_k, respectively. The nodal voltage at node k is denoted V_k, and the external current drawn from the circuit at node k is denoted I_k^e. The wire width, length, and thickness of branch k are w_k, l_k and t_k, respectively. The conductance of branch k is $g_k = w_k t_k / \rho_m l_k$, where ρ_m is the resistivity. The current density in branch k is $j_k = i_k / w_k t_k$.

The network behavior is described by a set of nodal equations $\mathcal{G}_n \mathcal{V}_n = \mathcal{I}_n$, and the branch elements are expressed as $\mathcal{G}_b \mathcal{V}_b = \mathcal{I}_b$, where $\mathcal{V}_b = [v_1, \ldots, v_n]^T$ is the vector of branch voltage drop, $\mathcal{V}_n = [V_1, \ldots, V_m]^T$ is the vector of nodal voltage, $\mathcal{I}_b = [i_1, \ldots, i_n]^T$ is the vector of branch current, $\mathcal{I}_n = [I_1^e, \ldots, I_m^e]^T$ is the vector of external current drawn from the circuit, $\mathcal{G}_b = diag(g_1, \ldots, g_n)$ is the branch conductance matrix, and $\mathcal{G}_n$ represents the nodal conductance matrix. The relation between $\mathcal{G}_n$ and $\mathcal{G}_b$ is

$$\mathcal{G}_n = \mathcal{A} \mathcal{G}_b \mathcal{A}^T = \sum_k^n \frac{w_k t_k}{\rho_m l_k} a_k a_k^T, \tag{1}$$

where $\mathcal{A}$ is the incidence matrix which implies the KCL and KVL, and a_k is the k^{th} column of matrix $\mathcal{A}$.

Due to the cost issue, P/G network routing area is required as small as possible. Therefore, the objective function of P/G network design is to minimize the total wire area $A(w) = \sum_{k=1}^n l_k w_k$ subject to the following constraints.

- **IR-drop constraint**
  IR-drop is the voltage fluctuation due to the resistance of the power delivery network, which may cause timing uncertainty and affect performance. Therefore, e.g., the voltage fluctuation from ground pads to the leaf nodes must be restricted with an upper bound

$$V_{k \in N_{leaf}} \leq V_{max}, \tag{2}$$

  where N_{leaf} is the set of leaf nodes and V_{max} is the maximum voltage of ground bounce.

- **Minimum-width constraint**
  According to the semiconductor process, there is a requirement for minimum wire width.

$$w_{k \in B} \geq w_{min}. \tag{3}$$

- **Electromigration constraint**
  EM is the transport of mass in metals under the stress of high current density. This metallization failure is the main reliability concern of IC designs. Therefore, maximum current density of each wire must be limited.

This paper is motivated by a shortcoming of the traditional design method: lack of robustness with respect to thermal reliability. It assumes that all interconnects do not exceed certain temperature and use the maximum current density of the expected lifetime at this temperature for all wires. For example, the interconnect in 0.13 μm technology at 105 ^oC as shown in ITRS [1] has the maximum current density of $1.1 \times 10^6 A/cm^2$. However, if some interconnects of the resulted design have temperature higher than the assumed temperature, the lifetime can be much lower than the expected due to its exponential relation to the inverse of temperature. Setting a higher temperature is not practical as well, because the maximum current density needs to decrease exponentially, which results in a significant area increase. In addition to this, minimizing the wire area causes the resulted design to consume more power in P/G networks, and therefore causes higher temperature in networks due to SH effect. We will next present the self-consistent constraint which integrates EM effect in consistent with SH effect to guarantee the thermal reliability.

3. THERMAL RELIABILITY AND SELF-CONSISTENT CONSTRAINT

Thermal effects are inseparable aspects of electrical power distribution and signal transmission through the interconnects due to SH caused by the flow of current [18]. Therefore, in P/G networks design, the current density constraint must comprehend both EM and SH effects.

3.1 Electromigration

The lifetime of metal wires, caused by EM effect, is modeled by Black's equation [22]:

$$MTF = \frac{A}{j^2_{EM,eff}} exp(\frac{E_a}{k_B T_m}) \tag{4}$$

where MTF is the mean-time-failure, A is a constant which depends on the geometry, $j_{EM,eff}$ is the effective ac value of current density, E_a is the activation energy, k_B is the Boltzman's constant, and T_m is the wire temperature.

Detailed studies have shown that the effective ac value of current density for unipolar rectangular pulsed stressing is the average current density $j_{EM,eff} = j_{avg}$ [21] [23]. However, currents flowing in the general P/G networks are arbitrary bipolar signals. For such case, the effective ac value of current density is given by the *Average Current Recovery* (ACR) model [21] [23]:

$$j_{EM,eff} = j_{ACR} = p_+ E[j_+] - \gamma \; p_- E[j_-] \tag{5}$$

where j_{ACR} is the current density of the ACR model, p_+ and p_- are probabilities of two current directions, and γ is the recovery parameter (< 1) of the ACR model. It heuristically accounts for the degree of healing of EM void damage that occurs when the current density changes sign.

If the goal of design is to achieve, e.g., at least 10 years of lifetime under the current density $j_{EM,ref}$ at temperature T_{ref}, the lifetime restriction of wires must satisfy

$$\frac{exp(\frac{E_a}{k_B T_m})}{j^2_{EM,eff}} \geq \frac{exp(\frac{E_a}{k_B T_{ref}})}{j^2_{EM,ref}}. \tag{6}$$

The reliability constraint from EM effect is defined as:

- **Electromigration constraint**

$$\begin{aligned} |\, j_{EM,eff}\,| \quad &\leq \quad |\, j_{EM,ref}|\, exp\left(\frac{E_a}{2k_B T_m} - \frac{E_a}{2k_B T_{ref}}\right) \\ &= \quad j_{EM,max}(T_m) \end{aligned} \tag{7}$$

  where $j_{EM,max}$ is the maximum current density of wire with temperature T_m, satisfying the required lifetime under the current density $j_{EM,ref}$ at temperature T_{ref}.

It can be observed that current density $j_{EM,max}(T_m)$ is sensitive to the wire temperature. Next section we will discuss the SH effect which affects the temperature.

3.2 Self-Heating

Current flowing through metal wires dissipates power and generates heat, increasing the wire temperature. This phenomena is referred to as SH or Joule heating effect and has become important due to the introduction of low-k dielectrics and the increase of 3-D thermal coupling [17]. P/G networks are away from the substrate and much longer than signal wire. The generated heat can not be spread efficiently to the heat sinks, and the temperature of P/G networks is usually higher than that of signal wires and substrate.

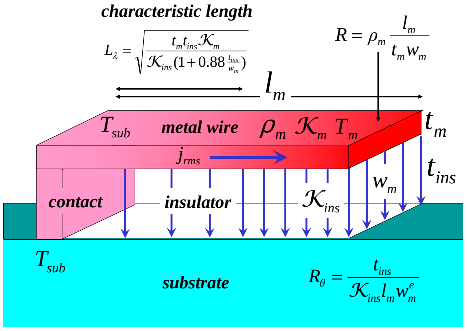
characteristic length
$L_\lambda = \sqrt{\frac{t_m t_{ins} \mathcal{K}_m}{\mathcal{K}_{ins}(1+0.88\frac{t_{ins}}{w_m})}}$
$R = \rho_m \frac{l_m}{t_m w_m}$
l_m
T_{sub}
metal wire
ρ_m $\mathcal{K}_m$ T_m
t_m
j_{rms}
t_{ins}
w_m
contact
insulator
$\mathcal{K}_{ins}$
T_{sub}
substrate
$R_\theta = \frac{t_{ins}}{\mathcal{K}_{ins} l_m w^e_m}$

Figure 2: A wire with one end connected to the substrate, showing the region at a distance greater than the characteristic length where the self-consistent solution applies. In this region, most of the heat spreads through under insulator to substrate.

Figure 2 shows a metal wire with one end connected to the substrate through a contact, in which the width, length, thickness, and thermal conductivity are w_m, l_m, t_m, and $\mathcal{K}_m$. The thickness and the thermal conductivity of the underlying insulator are t_{ins} and $\mathcal{K}_{ins}$. The wire resistivity is temperature dependent and can be described as follows:

$$\rho_m(T_m) \quad = \quad \rho_o[1 + \beta(T_m - T_o)] \tag{8}$$

where ρ_o is the wire resistivity at reference temperature T_o, and β is the temperature coefficient of resistivity. Within a short distance to the contact the temperature distribution is spatially dependent and temperature will increase from contact end temperature T_{sub} to far end temperature T_m. This distance, *characteristic length*, is defined as [24]

$$L_\lambda = \sqrt{\frac{t_m t_{ins} \mathcal{K}_m}{\mathcal{K}_{ins}\left(1 + 0.88\frac{t_{ins}}{w_m}\right)}}. \tag{9}$$

A *thermally-long* wire is a wire whose length is longer than L_λ. For local wires, most of the wire lengths may not be thermally-long. The temperature increase in these layers is not obvious. However, the temperature increase in global P/G networks compared to the substrate cannot be ignored, and that can be expressed as:

$$\Delta T_{SH} = T_m - T_{sub} = \frac{j^2_{rms} t_{ins} t_m w_m \rho_m(T_m)}{\mathcal{K}_{ins} w^e_m} \tag{10}$$

where T_{sub} is the substrate temperature, j_{rms} is the rms current density and is defined as $j^2_{rms} = E[j^2]$, and w^e_m is the effective thermal width and can be approximated by $w^e_m = w_m + 0.88 t_{ins}$. When $(w_m/t_{ins}) > 0.4$, w^e_m is accurate to within 3%. The effective thermal width w^e_m is always greater than w_m, and approaches w_m when $w_m \gg t_{ins}$.

Minimizing the wire area leads to higher current density in wires, which causes higher temperature. 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 EM and SH effects [20] [21].

3.3 Self-Consistent Solutions

For a bipolar signal, the effective duty cycle is defined as

$$r_{eff} = \frac{j^2_{EM,eff}}{j^2_{rms}} = \frac{j^2_{ACR}}{j^2_{rms}}. \tag{11}$$

Since finding realistic r_{eff} is a difficult problem by itself, we use the same r_{eff} for all branches. The maximum current density satisfying EM and SH effects can be computed by applying $j^2_{EM,eff}$ in Eqn. (6) with equality and j^2_{rms} in Eqn. (10) into Eqn. (11). The self-consistent equation is

$$r_{eff} = j^2_{EM,ref} \frac{exp(\frac{E_a}{k_B T_m})}{exp(\frac{E_a}{k_B T_{ref}})} \frac{t_{ins} t_m w_m \rho_m(T_m)}{(T_m - T_{sub}) \mathcal{K}_{ins} w^e_m}. \quad (12)$$

The self-consistent temperature T_m of the wire can be calculated for a given r_{eff}, a wave-shape parameter. Once this T_m is obtained, the corresponding j_{rms} can be derived either from Eqn. (10) or Eqn. (6) with equality. This j_{rms} satisfies not only EM effect but also SH effect.

3.4 Self-consistent Constraint

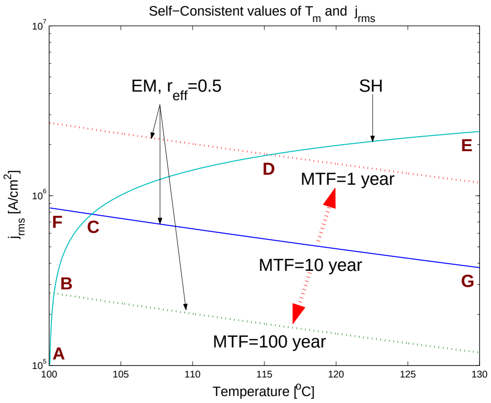

Figure 3: Graphical solutions of j_{rms} and T_m for EM and SH effects. The self-consistent solutions are on the curve A-B-C with the maximum at point C.

Figure 3 shows the graphical solutions of EM, Eqn. (6), and SH, Eqn. (10), of a wire with $r_{eff} = 0.5$ for illustrative purpose. Here it is assumed that the substrate temperature is 100 ^oC and the required MTF is 10 years. In order to satisfy the required MTF, the current density and wire temperature must satisfy EM effect in Eqn. (6), including the line F-C-G for 10 years of MTF and the area below the line F-C-G for MTF more than 10 years. For SH effect, the current flowing through the wires and causing temperature increase follows Eqn. (10), which is the curve A-B-C-D-E. Note that the MTF of SH curve decreases from 100 years at point B to 10 years at point C and 1 year at point D. The simultaneous solution of Eqn. (6) and Eqn. (10) is the curve A-B-C and has maximum values of j_{rms} and T_m at point C, which is the self-consistent solution of Eqn. (12).

The value of $j_{EM,max}(T_m)$ in EM constraint satisfying 10 years of MTF is line F-C-G. If the value of given $j_{EM,max}$ for the traditional design methods without considering wire temperature T_m is bigger than the self-consistent solution, the MTF will be shorter than the expected 10 years due to the temperature increase caused by SH effect with such current density. Then thermal reliability cannot be satisfied. If the value of given $j_{EM,max}$ is smaller than the self-consistent solution, the MTF is better than expected 10 years due to the temperature caused by SH effect with this current density. However, the resulted design has larger area because the current density of such wires can be higher.

In this paper, EM constraint is replaced by a self-consistent constraint on the P/G networks optimization. For a given duty cycle, all points lie on curve A-B-C are self-consistent solutions, and point C has the maximum value.

- **Self-consistent constraint**
  The current density satisfying EM and SH effects has

$$|\ j_{rms}\ | \quad \leq \quad j_{sc}, \quad (13)$$

where j_{sc} is the maximum current density at point C and j_{rms} must be any point on curve A-B-C.

4. THERMAL-AWARE P/G DESIGN

In [12], an interesting idea was proposed based on minimizing a weighted sum of average power consumption in P/G networks and total wire area, $P(w) + \mu A(w)$, where μ is a positive constant controlling the relative importance of both terms. Here we give a physical meaning of μ as *ohmic power density*. The physical meaning of optimizing $P(w) + \mu A(w)$ implies that the smaller the wire area, the larger the power dissipation in networks. Therefore, the traditional design sacrifices power dissipation to save wire area, having higher temperature in networks. This may lead to the degradation of chip reliability and timing failure.

4.1 Thermal-Aware P/G Network Optimization

Our method is based on minimizing the sum of each wire's weighted sum of average power dissipation and wire area, $\sum_{k=1}^{n} [P_k \ \ (w) + \mu_k A_k(w)]$, where $P_k(w)$ and $A_k(w)$ are the average power dissipation and wire area of wire k, respectively. With this approach, each wire can be assigned a different ohmic power density, μ_k, depending on the maximum current density restriction. This problem is subject to the following constraints: IR-drop constraint, Eqn. (2); minimum-width constraint, Eqn. (3); and self-consistent constraint, Eqn. (13). Note that the EM constraint is replaced by the self-consistent constraint due to the thermal issues as discussed in Section 3. Specifically, we consider the following thermal-aware based optimization problem.

Problem TOP *(Thermal-aware Optimization Problem)*

$$\begin{aligned} &minimize: && \sum_{k=1}^{n} [P_k(w) + \mu_k A_k(w)] \\ &subject\ to: && w_k \geq 0 \end{aligned} \quad (14)$$

It is proved that the total power term in the objective function can be expressed as

$$\sum_{k=1}^{n} P_k(w) = \sum_{k=1}^{n} E[\, i_k v_k] = E[\sum_{k=1}^{m} I^e_k V_k] = E[\mathcal{I}^T_n \mathcal{G}^{-1}_n(w) \mathcal{I}_n]. \quad (15)$$

Then the expected value of the power dissipation can be expressed in terms of widths, w, and current drawn from circuits, $\mathcal{I}_n$, as [12]

$$E[\mathcal{I}^T_n \mathcal{G}^{-1}_n(w) \mathcal{I}_n] = Tr \Gamma \mathcal{G}^{-1}_n(w), \quad (16)$$

where $\Gamma = E[\mathcal{I}_n \mathcal{I}^T_n]$ is the second moment of the $\mathcal{I}_n$, and Tr is the trace of the matrix. Since $Tr\Gamma\mathcal{G}^{-1}_n(w)$ is a differentiable convex function of w, this problem is a convex optimization problem so it can be globally solved efficiently.

The resulting design of problem TOP has the property that the optimal solution of each wire k is either zero width (not used) or has the *rms* current density $j_{k,rms} = \sqrt{\mu_k / \rho_m t_k}$. The necessary and sufficient conditions for the objective function to have optimal solution, subject to the constraint $w_k \geq 0$, are

$$\frac{\partial}{\partial w_k}\left(\sum_{k=1}^{n}[P_k(w)+\mu_k A_k(w)]\right)=0 \qquad (17)$$

for each wire k with $w_k > 0$, and

$$\frac{\partial}{\partial w_k}\left(\sum_{k=1}^{n}[P_k(w)+\mu_k A_k(w)]\right)\geq 0 \qquad (18)$$

for each wire k with $w_k = 0$. After similar derivation as in [12], we have the following results

$$\frac{\partial}{\partial w_k}\left(\sum_{k=1}^{n}[P_k(w)+\mu_k A_k(w)]\right)=-\frac{t_k}{\rho_m l_k}E[v_k^2]+\mu_k l_k. \quad (19)$$

According to Eqn. (17), it implies

$$v_{k,rms}=l_k\sqrt{\frac{\rho_m\mu_k}{t_k}} \quad \text{or} \quad j_{k,rms}=\sqrt{\frac{\mu_k}{t_k\rho_m}} \qquad (20)$$

for each wire $w_k > 0$.

This property gives a hint for P/G design to satisfy self-consistent constraint. For example, to have wire k with rms current density $j_{k,rms}$, which satisfies the self-consistent constraint in Eqn. (13), this can be designed by assigning the ohmic power density

$$\mu_k = \rho_m t_k j_{k,rms}^2 \ . \qquad (21)$$

Then the resulting optimal solution for each non-zero wire has rms current density $j_{k,rms}$. Note that the current density of each wire need not necessarily be same, this property is different from that in [12] and more practical. With the design of each wire having the same current density, some of the resulted wire widths may be too small, and some of the nodes may have big voltage drop.

However, the maximum allowed current density of some wires calculated from self-consistent solution may be too large to be practical. There are two potential problems for such situations. First, the resulted wire widths can be very small, and the scaling factor of minimum-width violation is very large. Then the wire area after scaling becomes very large. This is not practical. Second, the voltage drop of wire k, $\Delta v_k = \rho_m j_k l_k$, is big if j_k is large. This is the source of worst case IR-drop violation, and the voltage profile may have a lot of spikes. A simple technique by given an upper bound of maximum allowed current density can avoid these problems. This upper bound can be estimated by $j_{ub} = V_{max}/\rho_m L$, where L is the longest length of the node to the ground pads.

4.2 Scaling Method

In the problem TOP, we only consider the wire width constraint $w_k \geq 0$. However, the optimization problem still needs to satisfy IR-drop and minimum-width constraints. Scaling method will be used to satisfy the IR-drop and minimum-width constraints.

Suppose a set of wire widths w solves the problem TOP for a set of ohmic power density $\mu = \{\mu_1, \ldots, \mu_n\}$. Then λw solves the problem TOP for μ/λ^2, where $\lambda > 0$ is the scaling factor. This scaling method only changes the current density of each wire from j_k to j_k/λ. The current flowing through each wire remains the same, and KCL is still satisfied. According to this scaling rule, the violation of minimum-width or IR-drop constraint can be fixed easily. For example, the maximum allowed voltage drop is V_{max}, but the maximum voltage drop of the solution is V_{sol}. Then the scaling factor $\lambda = V_{sol}/V_{max}$ will be used to reduce the voltage of each node. The minimum width violation can be fixed by the same method. For example, the constraint of minimum allowed wire width is w_{min}, but the minimum width of the wire of the solution is w_{sol}. The scaling factor $\lambda = w_{min}/w_{sol}$ will be used to increase the wire widths. This is the *full scaling* method.

However, a few nodes of IR-drop violation or a few wires of minimum-width violation can cause huge increase of wire area if the full scaling is used. We propose a technique called *local scaling* to alleviate such situation. For a simple case of IR-drop violation, e.g., as shown in Figure 4-(A) for a ground network, node 1 with voltage V_1 violates the IR-drop constraint, and it connects to nodes $2, 3, 4$, and 5 with wires w_{12}, w_{13}, w_{14}, and w_{15}, respectively. We try to keep the current flowing through each wire to be the same so that KCL is still satisfied. The rule is to change the number of wire widths as small as possible when we reduce the voltage V_1. Therefore, V_1 can be reduced to $V_1' \leq V_{max}$ by increasing the widths of the connected wires. In this procedure, we simply update the wire widths $w_{1i}' = w_{1i}\frac{V_i - V_1}{V_i' - V_1}$ for $i = 2 \ldots 5$. Since this update does not change the branch currents, the voltages V_2, V_3, V_4, and V_5 remains the same. This method can be applied to a group of nodes violating IR-drop constraint. For example, as shown in Figure 4-(B), the voltages of V_1, V_2, V_3, and V_4 violate IR-drop constraint. First, we find the neighbor nodes $(5 \ldots 12)$ of these violated nodes and keep their voltages unchanged. Then we can reduce the voltages of these violated nodes by changing the widths of wires bounded by points $5 \ldots 12$. The local scaling method can be applied to the case of minimum-width violation too. For example, as shown in Figure 4-(C), wire w_{12} violates the minimum-width constraint. If the wire width of w_{12} is increased, voltages V_1 and V_2 will be changed too. In such situation, the wires connected to nodes 1 and 2 have to be scaled. This idea is similar to the case of IR-drop where V_1 and V_2 are violated. A group of wires violated can be fixed with the same method as shown in Figure 4-(D).

We can find the solution efficiently by solving a small-size constrained optimization problem. The objective function is to minimize the area of corresponding wires. The voltage of nodes inside neighbor nodes, V_i, must satisfy $V_i \leq V_{max}$. The width of wires inside neighbor nodes, w_{ij}, must have $w_{ij} \geq w_{min}$. The branch currents also need to keep unchanged. There is a possibility that we can not find a solution. It can be solved by extending the region of neighboring nodes. In general, local scaling takes negligible runtime and increaes the total wire area by less than 0.01%.

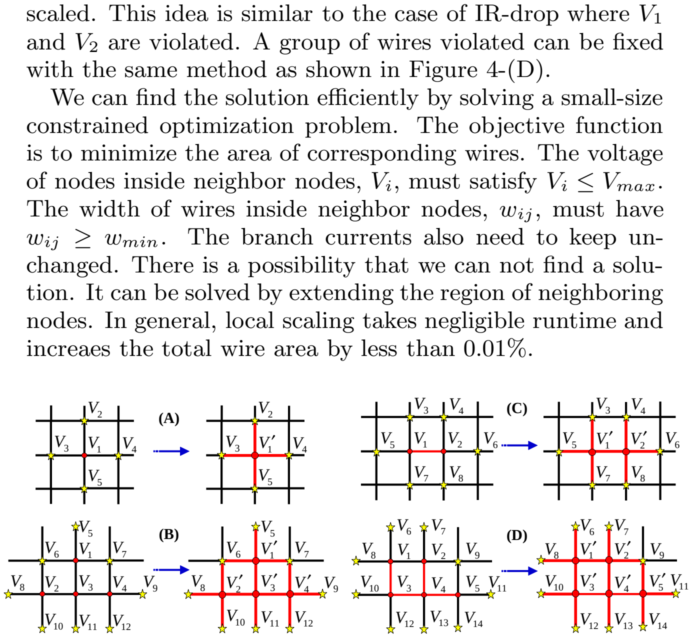

Figure 4: Local scaling method.

4.3 Algorithm

Algorithm 1 shows the thermal-aware P/G networks design algorithm. Suppose that the currents drawn are variables. The first and second moments of the current, as well as the correlations between currents, are known. Note that the current that causes the worst case IR-drop is included. The initial wire widths and duty cycle are given. According to these information, the substrate temperature profile can be simulated, e.g., by 3D Thermal-ADI [25]. From the substrate temperature profile and initial wire widths, maximum current density of each wire that satisfies self-consistent constraint is calculated, Eqn. (6) and Eqn. (12). Assigning the current density of each wire according to the self-consistent constraint. Then, we solve the problem TOP and a set of wire solutions is obtained. The new substrate temperature profile is simulated again according to the new set of wire solutions. Since the assigned current density of each wire is according to the initial wire widths and substrate temperature profile, this procedure repeats until the difference of current density of two consecutive iteration reaches error tolerance δ. After finding the solution, we still need to check the violations of IR-drop or minimum-width constraint. The violated nodes and wires will be fixed by scaling method.

Algorithm 1: *Thermal-Aware P/G Networks Design*

```
Given
    Effective duty cycle $r_{eff}$;
    Current sources $\mathcal{I}$;
    Initial $w^0 = \mathbf{1}$;
Simulate substrate temperature profile $T^0_{sub}$;
Iteration $q \leftarrow 1$;
repeat
    Calculate self-consistent current density
        $j^q_{sc} \leftarrow (w^{q-1}, T^{q-1}_{sub})$;
    Solve Thermal-Aware Power Grid Optimization
    {
        Assign current density for each wire
            $j^q_k \leftarrow min(j^q_{sc,k}, j_{ub})$;
        Assign ohmic power density
            $\mu_k \leftarrow \rho_m t_k (j^q_{k,rms})^2$;
        Optimize problem TOP
            $min \;\; E[\mathcal{I}^T_n \mathcal{G}^{-1}_n(w)\mathcal{I}_n] + \sum_{k=1}^{n} \mu_k l_k w_k$
            $s.t. \;\; w_k \geq 0$
    }
    $w^{q-1} \leftarrow w^q$;
    Obtain substrate temperature profile $T^q_{sub}$;
    $T^{q-1}_{sub} \leftarrow T^q_{sub}$;
    $q \leftarrow q + 1$;
until $\parallel j^q_k - j^{q-1}_k \parallel_\infty \;\; \leq \;\; \delta$
Wire widths $w = w^q$;

if ( IR-drop or minimum-width constraint is violated )
    Local Scaling Section 4.2;
Final wire widths $w^* = w$
```

4.4 Computational Issues

The objective function of problem TOP is a smooth convex function with constraint $w_k \geq 0$. We adapt the logarithmic barrier method by augmenting the objective function with a logarithmic barrier function. The inequality constrained optimization problem TOP can be translated into an unconstrained problem TOP-LBM.

Problem TOP-LBM *(TOP with Logarithmic Barrier Method)*

$$\phi(w) = Tr\Gamma\mathcal{G}_n^{-1}(w) + \sum_{k=1}^{n}\mu_k l_k w_k - \beta\sum_{k=1}^{n}\log w_k \qquad (22)$$

Here $\beta > 0$ is referred to as barrier parameter [26]. This function is defined for each wire $w_k > 0$, and it is smooth and convex on its domain. The minimizer of problem TOP-LBM is suboptimal of problem TOP with an accuracy of at least $n\beta$. Therefore, to solve problem TOP to an accuracy of ϵ, we need to have $\beta = \epsilon/n$. Instead of solving problem TOP, we can efficiently solve problem TOP-LBM repeatedly for a sequence of decreasing β until the accuracy is reached.

For the unconstrained minimization problem TOP-LBM, a limited memory algorithm L-BFGS-B is used for solving this problem [27] [28]. L-BFGS-B is a quasi-Newton algorithm capable of handling bounds on the variables. It is useful for solving a large problem in which the Hessian is dense like our problem. L-BFGS-B only needs the gradient, and for Eqn. (22), it is

$$\frac{\partial\phi(w)}{\partial w_k} = -\frac{t_k}{\rho_m l_k}a_k^T\mathcal{G}^{-1}(w)\Gamma\mathcal{G}^{-1}(w)a_k + \mu_k l_k - \frac{\beta}{w_k}. \qquad (23)$$

The numerical issues are in the following discussions. The initial width of each wire is set to be $w_k^0 = 1\mu m$. According to the initial w, we can calculate the substrate temperature and maximum current density of self-consistent constraint. Then the ohmic power density μ_k can be assigned, Eqn. (21). For the first iteration in Algorithm 1, we scale the wire widths to make power term $Tr\Gamma\mathcal{G}^{-1}$ and weighted area term $\sum_{k=1}^{n} u_k l_k w_k$ equal. After optimization, the scaled wire widths are translated back to non-scaled wire widths. The initial value of β is set as $0.05(Tr\Gamma\mathcal{G}^{-1})/n$ suggested by [12]. The β value is decreased by a factor of 10 until $n\beta$ is smaller than the required accuracy. Another practical concern is that in the case of optimal solution, some wire widths are equal to zero or very small. An implementation lower bound, e.g., $0.1w_{min}$ is set to remove the wires below this bound. This implementation technique can improve the runtime efficiency and avoid small wire widths. The problem of small wires are discussed in the previous section.

5. EXPERIMENTAL RESULTS

The proposed thermal-aware P/G networks method is implemented with C++ language and executed on a 1.6 GHz Intel Pentium 4 with 640 MB memory. We apply the proposed optimization algorithm on the P/G networks design of an industrial test chip with size $11.3mm \times 14.4mm$ and power consumption $48W$. The P/G grid is 56×72 with wire length $200\mu m$. The large scale network problem can be extended by the multigrid technique, e.g. [11]. The ground nodes are placed every 4 nodes in the x and y directions. The minimum wire width constraint is set by $w_{min} = 0.6\mu m$. The power supply is $1.5V$ and the maximum voltage drop of IR-drop constraint is $V_{max} = 150mV$. The sheet resistivity is $0.027\ \Omega$ at temperature 120^oC, and the temperature coefficient of resistivity is $\beta = 6.8 \times 10^{-3}K^{-1}$. We assume that wires have 10 years of operation lifetime under the current density $j_{EM,ref} = 10^6 A/cm^2$ with $T_{ref} = 105\ ^oC$. The first and second moments of the currents are estimated from circuit simulation [12]. The effective duty cycle is 0.5. We test the thermal-aware P/G networks optimization method and compare with traditional methods which consider EM constraint. The simulation results are shown in Table 1.

Thermal-Aware		Yes	No
Grid size		56×72	56×72
Total wires		8192	8192
Removed wires		1663	1857
Wire width before scaling (μm)		$0.05 \sim 34.18$	$0.05 \sim 16.85$
Minimum Lifetime ($years$)		9.8	4.1
Lifetime Violations ($> 5\%$)		0	198
Target j_{rms} ($mA/\mu m^2$)		$3.58 \sim 5.5$	5.5
Final j_{rms} ($mA/\mu m^2$)		$3.58 \sim 5.5$	$4.06 \sim 8.9$
Worst v_{rms} (mV)		97.2	113.5
Area (μm^2)		2.678×10^6	2.615×10^6
Iteration #		3	1
Runtime (sec)	Iteration 1	614.404	866.426
	Iteration 2	77.024	-
	Iteration 3	60.103	-
	total	751.621	866.426

Table 1: Comparison of the P/G networks design with and without thermal integrity.

First, the chip reliability by way of wire lifetime is shown in Figure 5. Without thermal reliability integrity, there are 198 wires violating the 10 years lifetime and the minimum lifetime is only 4.1 years. In general, these violated wires are in hot spots and are not well-designed. With thermal-aware design, the reliability of all wires satisfies the expected 10 years, and the minimum lifetime is 9.8 years. This 2% error comes from the experimental setting of the stopping criteria δ in Algorithm 1. In this case, δ is set to be $0.1 mA/\mu m^2$. The error of lifetime can be improved by decreasing the stopping criteria δ, but the number of iterations will increase.

If we allowed each wire to reach maximum current density of self-consistent solution, the lifetime of each wire should be 10 years. However, there are practical concerns for this design choice as discussed in Section 4.1 due to the large value of maximum current density. Therefore, we set an upper bound of the self-consistent current density, e.g., $5.5 mA/\mu m^2$ to avoid high current density wires. This can be observed in Table 1 for the case with thermal-aware design that wire rms current density is in the region $3.58 \sim 5.5 mA/\mu m^2$. For the case without thermal integrity, the EM constraint gives the target rms current density $5.5 mA/\mu m^2$. However, the SH effect makes the final current density for the case of without thermal integrity in the region $4.06 \sim 8.9 mA/\mu m^2$ which is much higher than the case with thermal integrity.

In Figure 6, the voltage distribution of the case without thermal integrity has higher voltage drop in the region of hot spots. On the other hand, the voltage distribution of the case with thermal integrity is more smooth. The worst value of rms voltage is 97.2 mV for the case with thermal integrity, and 113.5 mV for the case without thermal integrity. The average of the voltage distribution is 45.61 mV for the thermal-aware design, and 48.34 mV for the case without thermal integrity. From the discussion of wire reliability, current density, and voltage distribution, the power delivery quality is better with the thermal-aware design. The area of thermal-aware design is only 2.41% larger.

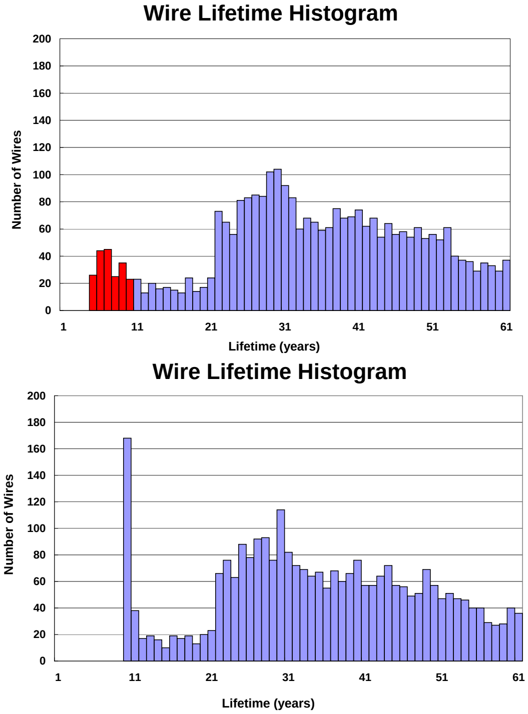

Figure 5: The wire lifetime histogram with lifetime up to 61 years for the P/G networks design without (top)/with (bottom) thermal integrity.

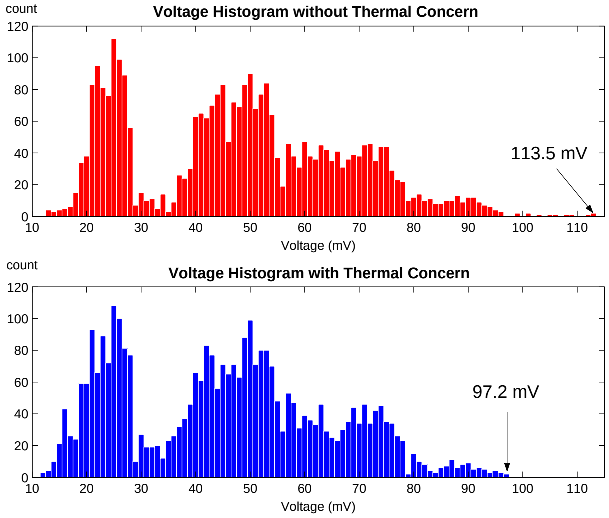

Figure 6: The voltage distribution of the P/G networks design without/with thermal integrity.

According to the optimal solution of wire widths, the substrate temperature profile is shown in Figure 7. It can be observed that the temperature varies from 30 ^oC to 135 ^oC. Therefore, it is important to consider thermal issue on P/G design to ensure the reliability. The optimal ground network for the thermal-aware P/G networks optimization design is shown in Figure 7. It can be seen that some wires are removed because these regions are reserved for routing and no currents are drawn. It also can be observed that some wires, marked in the circles, are dark due to their wide widths. Compare the layout with the full-chip temperature profile, it is obvious that these wires are in the hot spots. Since high temperature causes wires to have lower maximum current density to satisfy at least 10 years of lifetime, some of the wires in hot spots have wider widths.

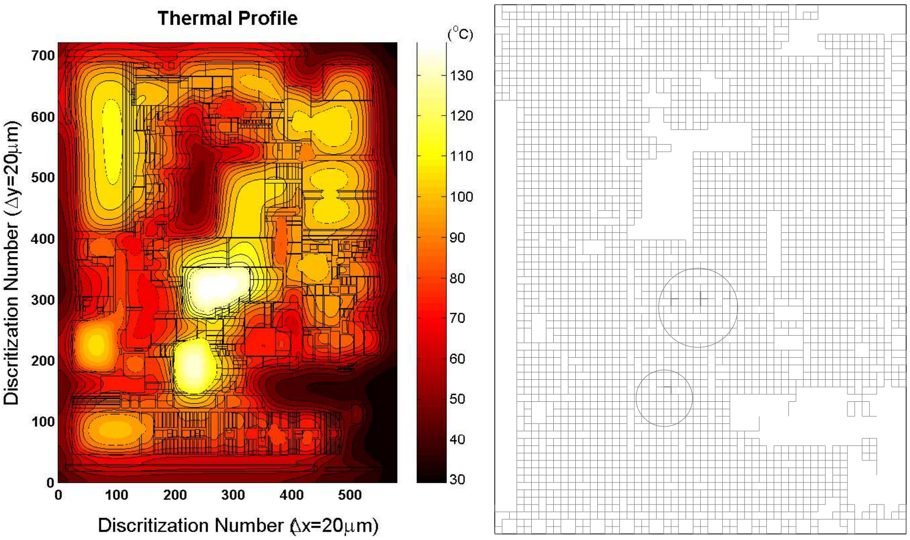

Figure 7: Full-chip temperature profile and optimal ground network for the P/G networks design.

6. CONCLUSION

In this paper we presented a P/G networks optimization algorithm which considers thermal and power integrity issues. This design method improves power delivery quality and thermal reliability.

The major advantage of the proposed algorithm is its ability to handle the thermal reliability in the P/G networks design. Unlike the traditional design methods, a self-consistent constraint is used to replaced the EM constraint. The self-consistent constraint gives a maximum allowed current density which simultaneously accounts for EM and SH effects. In addition to this, the objective function is based on minimizing the sum of each wire's weighted sum of average power dissipation and wire area. The power-area tradeoff approach can improve the thermal reliability in P/G networks through wire current density. The experimental results show the thermal-aware design improves thermal reliability and power delivery quality from the comparison of wire lifetime, voltage drop distribution, and wire *rms* current density.

7. REFERENCES

[1] http://public.itrs.net. International technology roadmap for semiconductors 2001 edition/2002 update.

[2] R. Saleh, S. Z. Hussain, S. Rochel, and D. Overhauser. Clock skew verification in the presence of IR-drop in the power distribution network. *TCAD*, 19(6): 635–644, Jun. 2000.

[3] Haihua Su, Kaushik Gala, and Sachin S. Sapatnekar. Fast analysis and optimization of power/ground networks. *ICCAD'00*, pp. 477–482.

[4] Salim U. Chowdhury and Melvin A. Breuer. Minimal area design of power/ground nets having graph topologies. *TCAS*, CAS-34(12): 1441–1450, December 1987.

[5] Salim U. Chowdhury and Melvin A. Breuer. Optimum design of ic power/ground nets subject to reliability constraints. *TCAD*, 7(7): 787–796, July 1988.

[6] S. Chowdhury. Optimum design of reliable ic power networks having general graph topologies. *DAC'89*, pp. 787–790.

[7] Robi Dutta and Malgorzata Marek-Sadowska. Automatic sizing of power/ground (p/g) networks in vlsi. *DAC'89*, pp. 783–786.

[8] X. Tan, C. J. Richard Shi, D. Lungeanu, and L. Yuan J. Lee. Reliability-constrained area optimization of VLSI power/ground networks via sequence of linear programmings. *DAC'99*, pp. 78–83.

[9] Ting-Yuan Wang and Charlie Chung-Ping Chen. Optimization of the power/ground network wire-sizing and spacing based on sequential network simplex algorithm. *ISQED'02*, pp. 157–162.

[10] Xiaohai Wu, Xianlong Hong, Yici Cai, C. K. Cheng, Jun Gu, and Wayne Dai. Area minimization of power distribution network using efficient nonlinear programming techniques. *ICCAD'02*, pp. 153–157.

[11] Kai Wang and Malgorzata Marek-Sadowska. On-chip power supply network optimization using multigrid-based technique. *DAC'03*, pp. 113–118.

[12] S. Boyd, L. Vandenberghe, A.El Gamal, and S. Yun. Design of robust global power and ground networks. *ISPD'01*, pp. 60–64.

[13] Jaewon Oh and Massound Pedram. Multi-pad power/ground network design for uniform distribution of ground bounce. *DAC'98*, pp. 287–290.

[14] H. Cai. Multi-pads single layer power net routing in vlsi circuit. *DAC'88*, pp. 183–188.

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

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

[17] Alberto Sangiovanni-Vincentelli Kaustav Banerjee, Amit Mehrotra and Chenming Hu. On thermal effects in deep sub-micron vlsi interconnects. *DAC'99*, pp. 885–891.

[18] Kaustav Banerjee and Amit Mehrotra. Global (Interconnect) Warming. *IEEE Circuits and Devices Magazine*, 17(5):16–32, September 2001.

[19] Amir H. Ajami, Kaustav Banerjee, Amit Mehrotra, and Massoud Pedram. Analysis of ir-drop scaling with implications for deep submicron p/g network designs. *ISQED'03*, pp. 35–40.

[20] William R. Hunter. Self-consistent solutions for allowed interconnect current density – part I: Implications fortechnology evolution. *TED*, 44(2):304–309, Feburary 1997.

[21] William R. Hunter. Self-consistent solutions for allowed interconnect current density – part II: Application to design guidelines. *TED*, 44(2):310–316, February 1997.

[22] James R. Black. Electromigration–a breief survey and some recent results. *TED*, ED-16(4): 338–347, April 1969.

[23] L. M. Ting, J. S. May, W. R. Hunter, and J. W. McPherson. Ac electromigration characterization and modeling of multilayered interconnects. *Proc. Int. Reliability Physics Symposium*, pp. 311–316, 1993.

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

[25] Ting-Yuan Wang and Charlie Chung-Ping Chen. 3D Thermal-ADI: A linear-time Chip Level Transient Thermal Simulator. *TCAD*, 21(12), December 2002.

[26] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*, chapter 17. Springer, 1999.

[27] R. H. Byrd, P. Lu, and J. Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16(5): 1190–1208, 1995.

[28] C. Zhu, R. H. Byrd, and J. Nocedal. L-bfgs-b: Algorithm 778: L-bfgs-b, fortran routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23(4): 550–560, 1997.