Power Aware Placement of On-Chip Voltage Regulators

Rassul Bairamkulov and Eby G. Friedman, *Life Fellow, IEEE*

***Abstract*—In traditional power delivery networks, the on-chip supply voltage is provided by board-level converters. Due to the significant distance between the converter and the load, variations in the load current are not effectively managed, producing a significant voltage drop at the point-of-load. To mitigate this issue, modern high-performance systems utilize on-chip voltage regulators. Due to the close proximity to the load, these regulators can quickly respond to fluctuations in the input voltage or load current, providing superior power quality. Integrated voltage regulators however require significant area, limiting the number of on-chip regulators. An algorithm for distributing on-chip voltage regulators is presented in this article. The algorithm is accelerated using the infinity mirror technique, enabling the analysis of arbitrarily sized power grids. The power quality is maximized with a limited number of regulators. Practical scenarios are supported, such as limited current capacity and restricted placement. Several orders of magnitude speedup in the placement process is demonstrated while achieving up to 88% reduction in the maximum voltage drop.**

***Index Terms*—Circuit optimization, design automation, design optimization, design tools, gradient methods, power distribution networks, power system modeling, power quality, system-on-chip.**

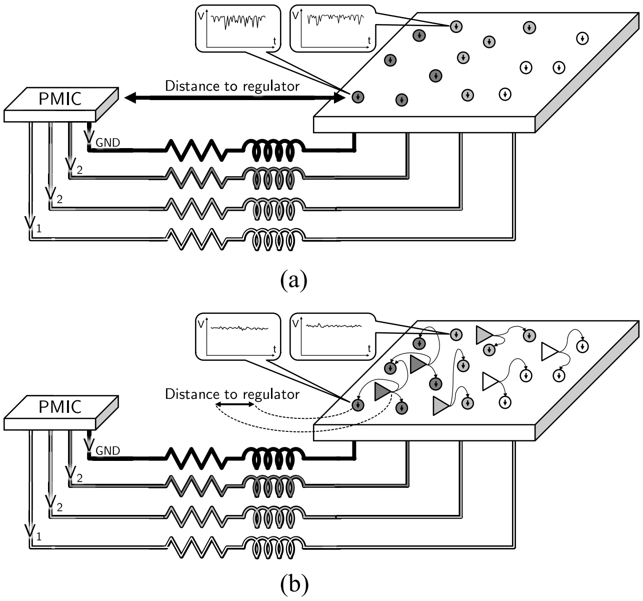


Fig. 1. Overview of power delivery systems. (a) Conventional power delivery system. The voltage converter within a PMIC provides multiple supply voltages to several power delivery systems. These networks are connected to the functional circuitry via dedicated power networks. Due to the significant distance to the regulators, fluctuations in the load current degrade the quality of the power supply. (b) Heterogeneous power delivery system with on-chip voltage regulators. The on-chip regulators are placed near the loads. A stable voltage is more effectively supplied to the functional circuits.

I. Introduction

THE PRIMARY objective of a VLSI power delivery system is to supply and maintain a nearly constant (i.e., low ripple) voltage across the load circuitry. Additional objectives include dissipating less power while limiting the current density to reduce the likelihood of electromigration [1]. In a conventional VLSI system, a power management IC (PMIC), also known as a voltage regulator module (VRM), is placed at the board level and supplies multiple voltages to the different on-chip voltage domains, as illustrated in Fig. 1(a). The primary limitation of this approach is the long physical distance between the off-chip regulator and the many billions of on-chip loads. The interconnect and I/O pins connecting the off-chip voltage converter with the load circuitry exhibit a high parasitic resistance and inductance, producing significant power noise [2]. The supply voltage is often increased to compensate for the voltage drop caused by the parasitic impedance of the power network [see Fig. 1(a)], degrading the overall energy efficiency of the system. Furthermore, the parasitic impedance between the converter and load circuitry slows the load regulation process. Considerable variations in supply voltage can be experienced by the load circuitry, potentially violating the noise margin of the many data signals.

Heterogeneous voltage regulation is a recent advancement in power delivery systems. The power efficient voltage converters within a PMIC are supplemented by area efficient on-chip fully integrated voltage regulators (FIVRs) [3], as shown in Fig. 1(b). The on-chip converters are placed in close proximity to the load devices. Since the physical distance and impedance between the on-chip regulator and devices are small, this configuration provides superior power quality despite load dependent current fluctuations. Furthermore, a local voltage domain can be created using the regulator, precisely controlling the voltage supplied to the functional circuits, achieving significant reductions in power consumption.

Manuscript received 13 December 2022; revised 4 June 2023 and 14 August 2023; accepted 20 August 2023. Date of publication 6 October 2023; date of current version 22 January 2024. This work was supported in part by the National Science Foundation under Grant CCF-2124453; in part by Synopsys; and in part by Qualcomm. This article was recommended by Associate Editor L. Behjat. *(Corresponding author: Rassul Bairamkulov.)*

Rassul Bairamkulov was with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627 USA. He is now with the Integrated Systems Laboratory, École polytechnique fédérale de Lausanne, 1015 Lausanne, Switzerland (e-mail: rassul.bairamkulov@epfl.ch).

Eby G. Friedman is with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627 USA (e-mail: friedman@ece.rochester.edu).

Digital Object Identifier 10.1109/TCAD.2023.3319285

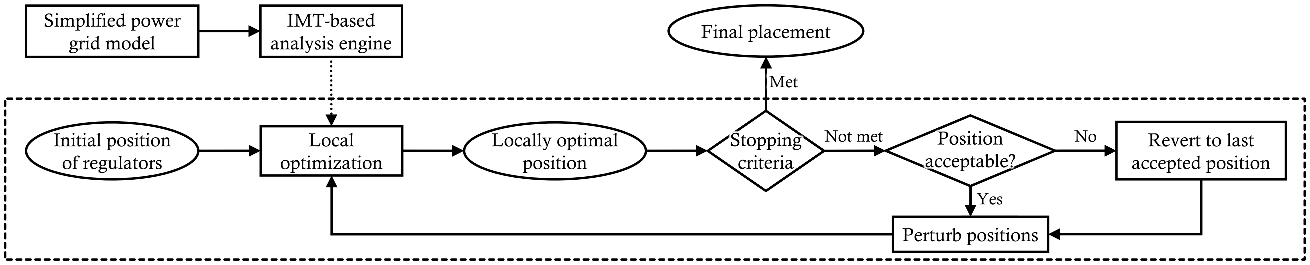


Fig. 2. Overview of the proposed optimization framework. The power grid model is initially simplified (see Sections III and IV). This model can be efficiently analyzed using the proposed grid analysis method based on the IMT (see Section II). During the BH optimization process (dashed rectangle, see Section V), locally optimal placement of the regulators is determined during each iteration. Based on the acceptance criterion (see (32)), this placement is accepted or rejected. The process repeats after randomly perturbing the regulator position until the stopping criteria are satisfied (maximum number of iterations in the case studies presented in Section VI).

A 20-fold reduction in standby power, 30% reduction in operating power, faster power gating, and significant reduction in off-chip area have been reported [3], [4], [5].

Increasing the number and enhancing the placement of the on-chip voltage regulators may greatly improve overall power integrity as compared to a single regulator, since the distance between the regulator and the load is much smaller. A typical FIVR occupies on-die area on the order of 0.1 to 2 mm^2 [5], [6]. Multiple regulators distributed within an IC can therefore occupy a significant portion of the on-chip die area. Furthermore, switching regulators often require additional infrastructural circuitry, such as extra routing layers and off-chip inductors [3]. Voltage regulators are therefore limited in quantity and should be judiciously distributed within an IC to enhance the power quality while complying with area constraints.

The literature discussing placement in the context of power delivery is relatively scarce. One of the earliest frameworks for placing voltage regulators within a grid is presented in [7]. Using a model of the local and global grids, the number and position of the LDO regulators are determined based on the estimated IR drop. The primary limitation of this method is the runtime of the algorithm. Distributing two LDO regulators within a grid with 17 000 nodes requires 49 min, and distributing 21 LDO regulators within a two million node mesh requires 2.5 days. The circuit analysis process occupies 90% of the runtime [7], limiting the scalability of the voltage regulator distribution process.

Due to the large size of power networks in modern VLSI systems, optimization based on simulation is impractical. The computational intractability of the problem has been encountered when optimizing the area of power networks [8], I/O pad locations [9], [10], [11], and decoupling capacitor allocation [11], [12]. Accelerated simulation tools are used to reduce the runtime of the circuit analysis process. A GPU accelerated multigrid analysis engine is used in [13] to explore tradeoffs associated with on-chip low-dropout regulators. The placement of 64 LDO regulators within a grid with nine million nodes requires only 2 h.

An alternative approach based on an efficient effective resistance model is utilized in [14]. The compact model enables estimation of the grid impedance in $O(1)$ time. LDO regulators and decoupling capacitors are placed within the system while considering the impedance approaching the hot spots. The number, size, and location of the on-chip LDO regulators and decoupling capacitors are concurrently determined to improve power integrity, maximize power efficiency, and minimize area. A significant improvement in power quality is achieved while maintaining the runtime below 6 min.

Building upon the methodology described in [14], a framework for power grid analysis is presented in this article. As illustrated in Fig. 2, the placement process begins with the model setup where the power network model is simplified for efficient analysis. During the optimization process, regulators are initially randomly placed within the available whitespace. A locally optimal placement is found using a local optimization algorithm. The local optimization process repeats from other initial placements until the stopping criteria are satisfied, such as the number of stall iterations or the total number of iterations.

The major contributions of this article include the following.

1) A novel methodology for fast evaluation of the voltage drop within a power grid without explicit calculation of the effective resistance.
2) A regulator placement framework based on a basin hopping (BH) algorithm [15], supporting a large number of regulators.
3) Support of practical design constraints, such as the finite dimensions of the grid, restricted placement, and limited current, supplied by the voltage regulators.
4) Further reduction of the computational runtime of the algorithm by clustering the current sources.

The remainder of this article is organized as follows. A computationally efficient model of an on-chip power network is discussed in Section II. The power grid modeling procedure is discussed in Section III. To improve the runtime of the optimization process, load clustering is performed, as described in Section IV. The setup of the optimization process is described in Section V. In Section VI, the performance of the algorithm is evaluated in a case study with six experiments. A holistic power network design procedure is presented in Section VII, followed by the conclusions in Section VIII.

II. Accelerated Grid Analysis

Due to the size of modern integrated systems, power networks are extremely large, containing many millions to billions of nodes. Accurate evaluation of the power noise requires a transient analysis of the distributed RLC networks. A transient analysis within the objective function however utilizes prohibitive runtime, since the objective function is evaluated hundreds of times before achieving convergence. In [7] and [10], for example, optimization of a relatively small number of parameters requires several days since the transient analysis is embedded within the objective function. Furthermore, slow evaluation of the objective function severely limits the search space explored during the optimization process. A more efficient approach to evaluating power noise is therefore necessary.

In [9], the power supply pad placement process is performed while ignoring transient effects, assuming a subsequent placement of the decoupling capacitors suppresses the transient noise. In [16], transient voltage fluctuations are suppressed by reducing the effective resistance between the decoupling capacitors and loads. In [14], transient information is incorporated by including the rise time of the current signal into the objective function. Using this method, those loads with the highest frequency are prioritized during the optimization process.

A significant correlation between static and transient power noise is observed in [14], [16], and [17]. To accelerate the optimization process, the IR drop can be used as a metric for the total voltage drop. Although transient noise is not accurately considered, improved computational efficiency enables wider search space exploration. For example, in [10], using IR drops as a metric of the power noise, superior results are produced as compared to using a transient analysis. The IR drop is therefore adopted as an objective function in the proposed framework.

Standard circuit analysis tools, such as SPICE, are based on the modified nodal analysis (MNA) technique [18], In MNA, a circuit is modeled in terms of six input matrices, representing the connections and parameter values [19]

$$\begin{bmatrix} Y & B \\ C & D \end{bmatrix}\begin{bmatrix} \mathbf{v} \\ \mathbf{i} \end{bmatrix} = \begin{bmatrix} \mathbf{j} \\ \mathbf{e} \end{bmatrix} \tag{1}$$

where $\mathbf{v}$ and $\mathbf{i}$ are vectors of, respectively, the node voltages and currents through the voltage sources, Y is the matrix of nodal admittances, and B, C, D, $\mathbf{j}$, and $\mathbf{e}$ describe the current and voltage sources. The constructed matrix equation is solved for $[\mathbf{v}, \mathbf{i}]^T$.

Since MNA is based on solving a system of linear equations, this method scales superlinearly with the number of nodes within the network. Note however that VLSI systems commonly utilize global power grids consisting of two or more layers of orthogonal interconnects connected by vias, as illustrated in Fig. 3(a). Due to the regularity and symmetry of a power grid, the power network can be modeled as a resistive mesh, as depicted in Fig. 3(b). Due to the large size of the grid in practical circuits, an infinite 2-D model of the grid can be used to analyze this network. This approach supports the use of closed-form expressions for the effective resistance between two nodes within an infinite grid [20]

$$R(\mathbf{x}) = 2r\Omega_k(\mathbf{x}) \tag{2}$$

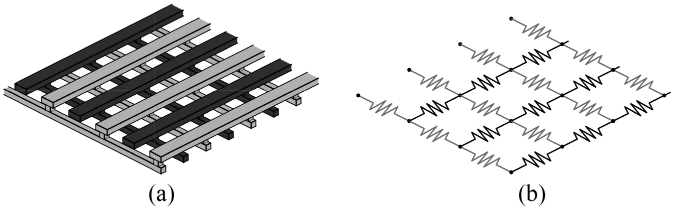

(a) (b)

Fig. 3. On-chip power grid, (a) layout of power (dark gray) and ground (light gray) distribution networks, and (b) power distribution network modeled as a resistive mesh. The ground part of the grid is typically analyzed separately or ignored.

where $\mathbf{x} = (x, y)$

$$\Omega_k(\mathbf{x}) = \frac{\sqrt{k}}{4\pi}\left[\ln\left(x^2 + ky^2\right) + 2\ln(\pi) + 2\gamma\right] + J(k). \tag{3}$$

r and x (kr and y) are, respectively, the resistance and physical distance between the nodes in the horizontal (vertical) dimension, and $J(k)$ is a polynomial function of k obtained from curve fitting (see [21] for the polynomial coefficients). Due to the finite size, however, this model exhibits a significant error near the boundaries of the grid. The infinity mirror technique (IMT), proposed in [21], overcomes this issue by modeling the boundaries of the grid with image current sources. With this approach, the effective resistance can be determined in $O(N_x N_y)$ time, where N_x and N_y denote the number of images, respectively, in the x and y dimensions. Only one to three images are sufficient to maintain the error below 1% in a practical grid [21]. Observe that the analysis runtime does not directly depend upon the size of the mesh. Based on the effective resistance, the voltage at a subset of grid nodes can be efficiently determined, as described in the following section.

A. IMT-Based Grid Analysis

The maximum voltage drop within a power network is typically observed in proximity of the load. Minimizing the voltage drop at the loads is therefore sufficient to minimize the voltage drop within the system. This feature is exploited in the proposed grid analysis algorithm. An overview of the proposed analysis procedure is illustrated in Fig. 4. Based on the effective resistance model, only a small subset of the nodes within the grid is considered. Each voltage source within the network is replaced by an equivalent current source. Based on the reduced network of current sources, the voltage at each load is determined.

Let $\ell = (\mathbf{x}(\ell), I(\ell))$ be a load located at position $\mathbf{x}(\ell)$ and drawing current $I(\ell)$ from a resistive grid of size $\mathbf{w} = (w_x, w_y)$. The set $\mathcal{L} = \{\ell_p | p \in [1, \dots, n]\}$ is a set of all loads within the network. Based on the IMT algorithm [21], the finite grid is mapped to an infinite 2-D resistive lattice by introducing image current sources mimicking the effect of the grid boundaries. The images of each load $\ell_p \in \mathcal{L}$ are described by a set of loads

$$\ell_p^* = \left\{\left(\mathbf{x}_p^{(i,j)}, I_p\right) | i \in [-N_x, \dots, N_x], j \in \left[-N_y, \dots, N_y\right]\right\} \tag{4}$$

where, for brevity, $\mathbf{x}_p^{(i,j)} = \mathbf{x}(\ell_p^{(i,j)}) = (x_p^i, y_p^j)$, $I_p = I(\ell_p)$, and

$$x_p^i = \begin{cases} w_x i + x_p, & \text{if } i \text{ is even} \qquad \text{(5a)} \\ w_x(i+1) - x_p - 1, & \text{if } i \text{ is odd} \qquad \text{(5b)} \end{cases}$$

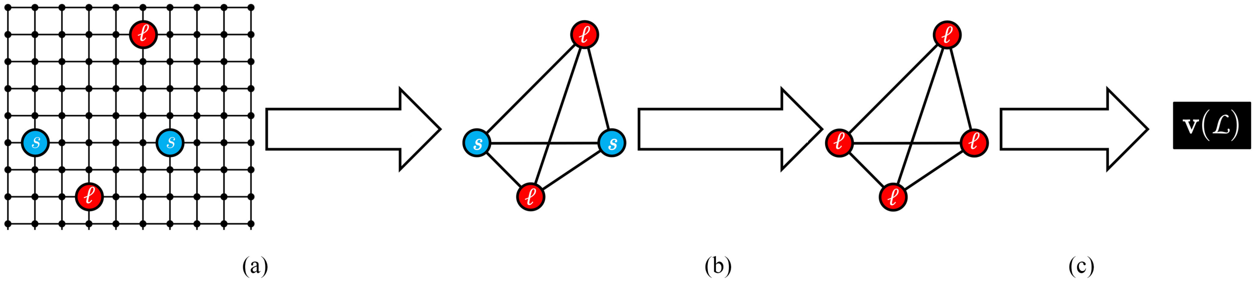


Fig. 4. Overview of the IMT-based grid analysis process. The grid network is reduced into a network consisting of only regulators s and loads ℓ. The regulators modeled as voltage sources are replaced with equivalent current sources. Upon completion, the voltage at each load $\mathbf{v}(\mathcal{L})$ is determined. (a) Reduction. (b) Replacement. (c) Result.

$$y_p^j = \begin{cases} w_y j + y_p, & \text{if } j \text{ is even} \quad \text{(6a)} \\ w_y(j+1) - y_p - 1, & \text{if } j \text{ is odd.} \quad \text{(6b)} \end{cases}$$

The electric potential at node $\mathbf{u} = (x_{\mathbf{u}}, y_{\mathbf{u}})$ in response to a unit load $\hat{\ell} = (\mathbf{x}, 1)$ with respect to a ground node at infinity is

$$\phi(\mathbf{u}, \mathbf{x}) = \sum_{\hat{\ell} \in \hat{\ell}*} \Omega_k\Big(\mathbf{u} - \mathbf{x}\big(\hat{\ell}\big)\Big). \tag{7}$$

By selecting arbitrary ground node $\mathbf{g}$, the voltage at node $\mathbf{u}$ becomes

$$v^{\mathbf{g}}(\mathbf{u}, \mathbf{x}) = \phi(\mathbf{u}, \mathbf{x}) - \phi(\mathbf{g}, \mathbf{x}). \tag{8}$$

Due to the principle of superposition, the voltage at node $\mathbf{u}$ is the weighted sum of the potentials caused by each current source within a grid

$$V^{\mathbf{g}}(\mathbf{u}) = \sum_{\ell_p \in \mathcal{L}} I_p v^{\mathbf{g}}\big(\mathbf{u}, \mathbf{x}_p\big). \tag{9}$$

Using (7)–(9), the voltage at each load $\ell \in \mathcal{L}$ is described by only considering the location of the current injection, effectively reducing the grid into a smaller network, as shown in Fig. 4(a). If a grid contains only current sources, the voltage at any node within a grid can be determined using (9). The power network however includes voltage regulators that maintain a constant voltage while changing the current supplied to the network. Any voltage source can therefore be transformed into a current source supplying equivalent current into a network, as shown in Fig. 4(b).

Finding the current injected by each voltage source requires additional processing. Suppose m voltage regulators are connected to a network. The set of voltage regulators $\mathcal{S}$ within the network is

$$\mathcal{S} = \big\{s_q | q \in [1, \ldots, m]\big\} \tag{10}$$

where

$$s_q = \big(\mathbf{x}_q, I_q\big). \tag{11}$$

The target voltage at each node $\mathbf{x}_q$, $q \in [1, \ldots, m]$ is known *a priori*, producing a vector $\mathbf{v}(\mathcal{S}) \in \mathbb{R}^m$ of target voltages

$$\mathbf{v}(\mathcal{S}) = [V_1, \ldots, V_m]^T. \tag{12}$$

To determine the current injected by each voltage regulator, an arbitrary node $\mathbf{g}$ is initially designated as ground. Without loss of generality, suppose $\mathbf{g} = \mathbf{x}_m$, producing set $\mathcal{S}^{\mathbf{g}} = \mathcal{S} \setminus s_m$. The target voltages are therefore adjusted, yielding a vector $\mathbf{v}^{\mathbf{g}}(\mathcal{S}) \in \mathbb{R}^{m-1}$

$$\mathbf{v}^{\mathbf{g}}(\mathcal{S}) = \big[V_1^{\mathbf{g}}, \ldots, V_{m-1}^{\mathbf{g}}\big]^T \tag{13}$$

where

$$V_q^{\mathbf{g}} = V_q - V_m. \tag{14}$$

The voltage $V_r^{\mathbf{g}}$ is determined by superimposing the effect of the supply and load currents

$$V_r^{\mathbf{g}} = \sum_{q=1}^{m} I\big(s_q\big) v^{\mathbf{g}}\big(s_r, s_q\big) + \sum_{p=1}^{n} I\big(\ell_p\big) v^{\mathbf{g}}\big(s_r, \ell_p\big) \tag{15}$$

where, for brevity

$$v^{\mathbf{g}}\big(s_r, s_q\big) = v^{\mathbf{g}}\big(\mathbf{x}(s_r), \mathbf{x}\big(s_q\big)\big) \tag{16}$$

$$v^{\mathbf{g}}\big(s_r, \ell_p\big) = v^{\mathbf{g}}\big(\mathbf{x}(s_r), \mathbf{x}\big(\ell_p\big)\big). \tag{17}$$

Reformulating (15) in matrix form yields

$$\begin{bmatrix} v^{\mathbf{g}}(s_1, s_1) & \ldots & v^{\mathbf{g}}(s_m, s_1) \\ \vdots & \ddots & \vdots \\ v^{\mathbf{g}}(s_1, s_{m-1}) & \ldots & v^{\mathbf{g}}(s_m, s_{m-1}) \end{bmatrix} \begin{bmatrix} I(s_1) \\ \vdots \\ I(s_{m-1}) \end{bmatrix} = \mathbf{v}^{\mathbf{g}}(\mathcal{S}) - \begin{bmatrix} v^{\mathbf{g}}(\ell_1, s_1) & \ldots & v^{\mathbf{g}}(\ell_n, s_1) \\ \vdots & \ddots & \vdots \\ v^{\mathbf{g}}(\ell_1, s_{m-1}) & \ldots & v^{\mathbf{g}}(\ell_n, s_{m-1}) \end{bmatrix} \begin{bmatrix} I(\ell_1) \\ \vdots \\ I(\ell_n) \end{bmatrix} \tag{18}$$

or, equivalently

$$\Phi^{\mathbf{g}}\big(\mathcal{S}, \mathcal{S}^{\mathbf{g}}\big)\mathbf{i}(\mathcal{S}) = \mathbf{v}^{\mathbf{g}}(\mathcal{S}) - \Phi^{\mathbf{g}}\big(\mathcal{L}, \mathcal{S}^{\mathbf{g}}\big)\mathbf{i}(\mathcal{L}). \tag{19}$$

The system described by (19) is underdetermined with $m - 1$ equations and m unknowns. To obtain the remaining equation, note that the total current drawn by the loads is equal to the total current injected by the voltage regulators

$$\mathbf{1}_{1,m}\mathbf{i}(\mathcal{S}) + \mathbf{1}_{1,n}\mathbf{i}(\mathcal{L}) = 0 \tag{20}$$

where $\mathbf{1}_{a,b}$ is an $a \times b$ matrix with all entries equal to 1. The current $\mathbf{i}(\mathcal{S})$ supplied by the voltage regulators can therefore be determined by solving a system of linear equations

$$\begin{bmatrix} \Phi^{\mathbf{g}}(\mathcal{S}, \mathcal{S}^{\mathbf{g}}) \\ \mathbf{1}_{1,m} \end{bmatrix} \mathbf{i}(\mathcal{S}) = \begin{bmatrix} \mathbf{v}^{\mathbf{g}}(\mathcal{S}) \\ 0 \end{bmatrix} - \begin{bmatrix} \Phi^{\mathbf{g}}(\mathcal{L}, \mathcal{S}^{\mathbf{g}}) \\ \mathbf{1}_{1,n} \end{bmatrix} \mathbf{i}(\mathcal{L}). \tag{21}$$

By combining $\mathcal{L}$ and $\mathcal{S}$, the set of current injections $\mathcal{I} = \mathcal{L} \cup \mathcal{S}$ is obtained. The voltage at each load is therefore

$$\mathbf{v}^{\mathbf{g}}(\mathcal{L}) = \Phi^{\mathbf{g}}(\mathcal{I}, \mathcal{L})\mathbf{i}(\mathcal{I}) + V_m \mathbf{1}_{||\mathcal{I}||,1}. \tag{22}$$

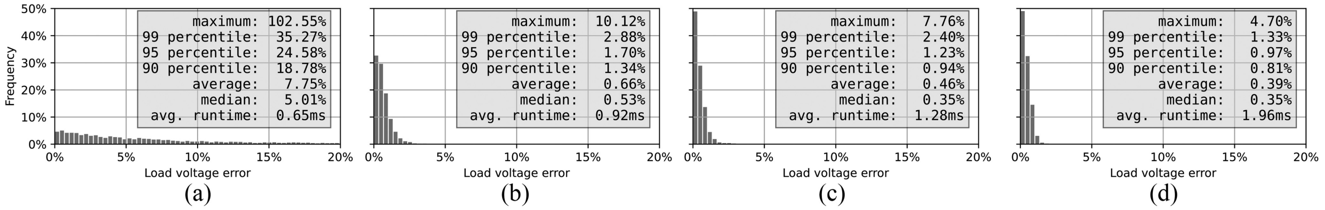


Fig. 5. Accuracy of the efficient voltage drop evaluation based on IMT with the number of images varied from (a) 0 to (d) 3. Using HSPICE [24], the voltage drop is determined, on average, in 4.19 s, three orders of magnitude slower than the IMT analysis. The error is significantly reduced by considering more images at the cost of increased runtime.

The IMT technique accelerates the regulator placement process in two ways. First, using the proposed fast grid analysis method, the voltage within a grid can be determined in $O(N_x N_y n(m+n))$ time, where m and n denote the number of, respectively, voltage regulators and loads [21]. Note that the proposed method does not depend upon the size of the mesh, enabling the analysis of arbitrarily large grids. The voltage for a subset of nodes is determined in milliseconds, many orders of magnitude faster than SPICE. Observe however that the runtime increases quadratically with the number of loads. To further accelerate the evaluation of the objective function, a load clustering operation is performed during the model setup, as described in Section IV.

Second, the algorithm enables evaluation of $\mathbf{v}(\mathcal{S})$ with noninteger $\mathbf{x}_q$, since the effective resistance can be evaluated for noninteger coordinates. Integrality relaxation is a commonly used method in integer programming, where the objective function is initially approximated with continuous variables [22]. With the relaxed integrality requirement, efficient continuous optimization algorithms can be applied, as described in Section V-A.

The accuracy of the IMT-based analysis process increases with the number of images N_x and N_y. The accuracy of the circuit analysis process is evaluated by randomly placing 20 voltage sources and 20 loads (current sources) within a 500 × 500 grid. The parameter k is varied between 1 and 6, according to the value of k applied in the `ibmpg` benchmark circuits [23] used in the case studies. The number of images is increased from zero to three. The load voltage is evaluated 1000 times using the proposed grid analysis method and HSPICE [24]. Due to the large size of the grid, the average runtime of the HSPICE analysis is 4.19 s. In contrast, the IMT-based analysis method requires, on average, less than 2 ms, providing three orders of magnitude speedup in voltage drop analysis. The accuracy of the load voltage drop evaluation is illustrated in Fig. 5. Observe that a single image is sufficient to reduce the average error below 1%, and two images reduce the error below 1.23% in 95% of evaluations. In the case studies, N_x and N_y are set to two to balance the accuracy with the runtime, as described in Section VI.

B. Limited Regulator Current

The amount of current reliably delivered by a regulator is a strong function of the regulator area [25]. For example, LDO regulators with wider power transistors can supply more current to the loads. Since on-chip regulators occupy silicon area, other circuitry may constrain the placement and size of the regulator. The maximum current supplied by an LDO is therefore limited by the size of the regulator. Furthermore, even if the size of the regulator is unlimited, other factors, such as electromigration [26], limit the maximum current that can be sourced by a regulator.

To consider this limitation during the optimization process, the fast grid analysis algorithm described in Section II is extended to support the limited current capacity of a regulator. Let $I_{\max} : \mathcal{S} \mapsto \mathbb{R}$ be a function mapping each regulator s to the maximum current $I_{\max}(s)$ that can be supplied.

Suppose after solving (21), the estimated current of subset $\mathcal{S}^* \subset \mathcal{S}$ exceeds the corresponding maximum current. Vector $\mathbf{i}(\mathcal{S})$ therefore does not realistically represent the current supplied by each regulator. This result however indicates that the regulators in $\mathcal{S}^*$ operate at maximum capacity, i.e., $I(s) = I_{\max}(s) \forall s \in \mathcal{S}^*$. Since the current supplied by these regulators is known, these nodes can be treated as loads. Transferring $\mathcal{S}^*$ into $\mathcal{L}$ yields

$$\mathcal{S}_1 \leftarrow \mathcal{S} \setminus \mathcal{S}^* \tag{23}$$

and

$$\mathcal{L}_1 \leftarrow \mathcal{L} \cup \mathcal{S}^*. \tag{24}$$

Note that a different ground node $\mathbf{g}$ should be selected if $\mathbf{g} \in \mathcal{S}^*$.

The system of (21) is transformed into

$$\begin{bmatrix} \Phi^{\mathbf{g}}\left(\mathcal{S}_1, \mathcal{S}_1^{\mathbf{g}}\right) \\ \mathbf{1}_{1,||\mathcal{S}_1||} \end{bmatrix} \mathbf{i}(\mathcal{S}_1) = \begin{bmatrix} \mathbf{v}^{\mathbf{g}}(\mathcal{S}_1) \\ 0 \end{bmatrix} - \begin{bmatrix} \Phi^{\mathbf{g}}\left(\mathcal{L}_1, \mathcal{S}_1^{\mathbf{g}}\right) \\ \mathbf{1}_{1,||\mathcal{L}_1||} \end{bmatrix} \mathbf{i}(\mathcal{L}_1). \tag{25}$$

If none of the currents in $\mathbf{i}(\mathcal{S}_1)$ exceeds the current limit, the process is completed, and the voltage at any node can be determined. Otherwise, the process is repeated until all of the regulator currents satisfy the constraints.

To converge, this recursive procedure requires the size of $\mathcal{S}_1$ to be greater than zero; i.e., at least one regulator should operate within the current capacity of that regulator during each iteration. This condition requires the total current drawn by the loads to not exceed the combined current capacity of the regulators

$$\sum_{\ell \in \mathcal{L}} (I(\ell)) \le \sum_{s \in \mathcal{S}} (I_{\max}(s)). \tag{26}$$

During subsequent iterations, the total current drawn by the extended set of loads $\mathcal{L}_1$ is

$$\sum_{\ell \in \mathcal{L}_1} (I(\ell)) = \sum_{\ell \in \mathcal{L}} (I(\ell)) - \sum_{s \in \mathcal{S}^*} (I_{\max}(s)) \tag{27}$$

TABLE I
PARAMETERS OF THE EQUIVALENT GRIDS USED TO MODEL THE ibmpg BENCHMARK CIRCUITS [23]

	Pitch		Resistivity, mΩ		Dimensions	
	x	y	x	y	x	y
ibmpg2	48	72	4.000	16.25	170	115
ibmpg3	864	1,296	0.714	2.407	354	236
ibmpg4	48	24	35.00	32.50	284	571
ibmpg5	12	12	10.00	21.67	1,772	1,772
ibmpg6	280	280	0.286	0.464	3,630	3,644

while the total current capacity of the reduced set of regulators $\mathcal{S}_1$ is

$$\sum_{s\in\mathcal{S}_1}(I_{\max}(s)) = \sum_{s\in\mathcal{S}}(I_{\max}(s)) - \sum_{s\in\mathcal{S}^*}(I_{\max}(s)). \tag{28}$$

Combining (26)–(28) yields

$$\sum_{\ell\in\mathcal{L}_1}(I(\ell)) \le \sum_{s\in\mathcal{S}_1}(I_{\max}(s)). \tag{29}$$

Equation (29) indicates that the set $\mathcal{S}$ cannot be reduced to an empty set in subsequent iterations provided that (26) is initially satisfied. Condition (26) is therefore sufficient to ensure the convergence of the procedure.

III. Power Grid Model

Although practical power networks are typically grid structured, significant deviations, such as missing vias or variable interconnect pitch, do exist [27]. Furthermore, a global mesh may span more than two layers, complicating the two layer model. To consider practical grids, a power network should be converted into an equivalent resistive mesh while preventing excessive deviations from the original grid.

To simplify the structure of the network, a 3-D-to-2-D grid regularization technique is proposed in [28]. By ignoring the via impedance, multiple grid layers are initially collapsed into a single layer based on location. The 2-D network is mapped into a 2-D grid with a fixed pitch, yielding a resistive mesh with a fixed pitch. The resulting grid exhibits an error of less than 1% as compared to SPICE.

A similar approach is followed here. By examining each benchmark circuit, a dominant wire pitch and resistance are observed. Consider, for example, the ibmpg2 power network [23]. The dominant resistivity and pitch of the interconnects in the x dimension are, respectively, 72 units and 0.635 milliohms per unit length, as depicted in Fig. 6(a) and (b). Similarly, the dominant pitch in the y direction is 48 units with a resistivity of 4 milliohms per unit length, as shown in Fig. 6(c) and (d). The resulting simplified grid has dimensions $\mathbf{w} = (170, 115)$ and $k = 4.2$. The parameters of a simplified grid for each benchmark circuit are listed in Table I.

IV. Load Clustering

The functional circuits within an IC are typically distributed across the entire power network. A large number of load currents within each functional block is therefore connected to the power grid. As described in Section II, the runtime of the proposed method increases with the number of loads. Individually considering each load incurs a significant computational penalty. Recall however that a power grid is a smooth system, i.e., a small variation in position correlates with a small variation in voltage [29]. Multiple loads can therefore be merged into a single load if located sufficiently close to each other.

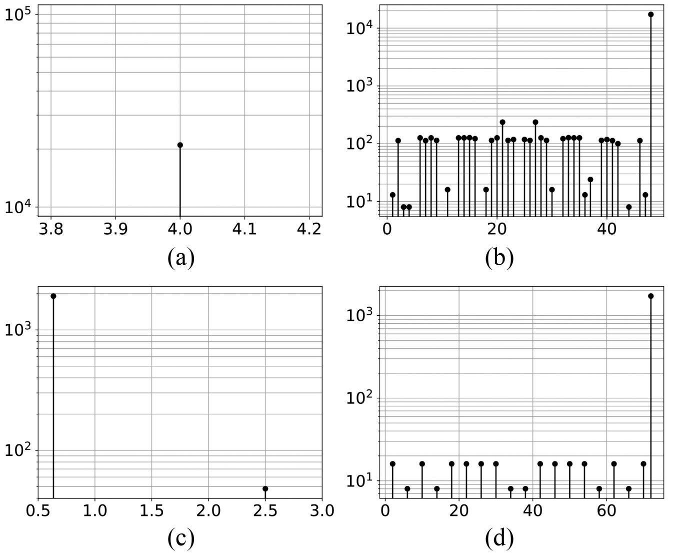
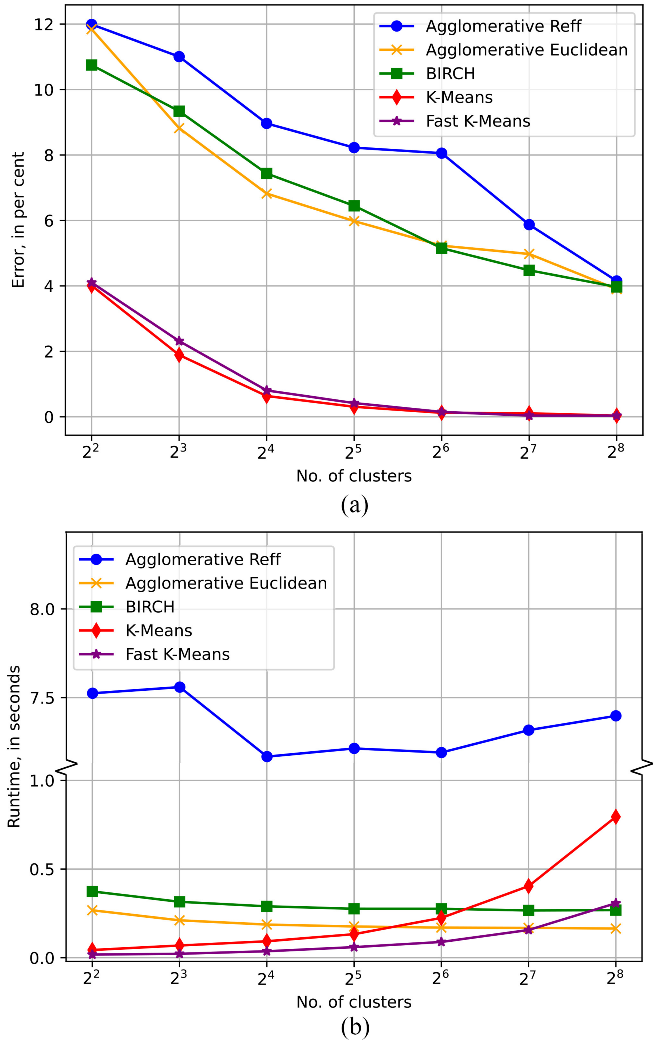


Fig. 6. Prevalence of resistivity and pitch within the ibmpg2 benchmark circuit, (a) resistivity along the x dimension, (b) pitch along the x dimension, (c) resistivity along the y dimension, and (d) pitch along the y dimension. The equivalent grid is constructed based on the dominant resistivity and pitch within the network.

At the global level, this procedure is accomplished by clustering. The clustering algorithm divides the set of current loads into separate groups based on the location and size. Loads within the same cluster are replaced by a single current source at the centroid of the cluster. The current drawn by the new load is the sum of all currents drawn by the loads within the cluster.

Four clustering algorithms are considered for load clustering, including agglomerative clustering [30], K-means [31], fast K-means [32], and BIRCH [33]. Agglomerative clustering is a bottom-up clustering algorithm that constructs clusters by recursively merging smaller clusters. The advantage of K-means clustering is the support of weighted clustering, enabling the size of the load to be considered during clustering. The fast K-means algorithm is an accelerated version of the K-means algorithm, achieving faster clustering by processing smaller subsets of the points [32]. The advantage of the BIRCH algorithm is speed, enabling a large number of clusters to be more efficiently created.

To evaluate the error produced by these algorithms, a 51 × 51 power grid containing 2,601 loads (one per node) of random size is considered. The number of clusters is varied from 4 to 256. The relationship between the error (in per cent) and the number of clusters is shown in Fig. 7(a). Observe that the error gradually decreases with the number of clusters. Both the K-means and fast K-means clustering algorithms exhibit superior performance. The smaller error can



(a)

(b)

Fig. 7. Comparison of algorithms for load clustering. (a) Error in estimated minimum voltage within a 51 × 51 grid and, (b) computational runtime (in minutes) of the algorithms (note the broken y-axis). A two orders of magnitude reduction in the number of loads is possible with only a minor effect on the accuracy of the estimated minimum voltage.

be explained by considering the load current, enabling more accurate clustering of the loads. Weighted clustering supported by the K-means algorithms greatly improves the accuracy of the voltage estimation, outperforming algorithms without this feature.

The computational runtime of each algorithm is illustrated in Fig. 7(b). Observe that with a small number of clusters, the runtime of the K-means clustering algorithm is smaller than the runtime of the agglomerative clustering and BIRCH algorithms. The K-means algorithms however scale superlinearly with the number of clusters. The number of clusters in the case studies described here is however sufficiently small to tolerate the superlinear complexity of the K-means algorithms. Considering the superior accuracy and reasonable runtime, the fast K-means algorithm is preferable for the case studies described here.

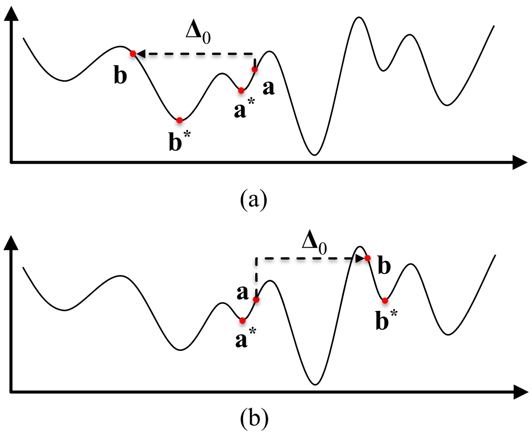
V. Optimization Setup

Constrained global optimization is applied to determine the optimal location of the regulators. The optimization process aims to minimize the maximum power noise within the network. The objective function is the voltage drop as a function of the position of the voltage regulators

$$v_{\text{drop}}(\mathcal{S}) = -\min\left(\mathbf{v}^{\mathbf{g}}(\mathcal{L})\right)|_{\mathcal{S}} \tag{30}$$

where $\mathcal{S} = \{s_q | q \in [1, \ldots, m]\}$ is the set of voltage regulators within the network. Each voltage regulator s_q has an associated location $\mathbf{x}_q = (x_q, y_q)$ and current capacity $I_{q,\max}$

$$s_q = \left(\mathbf{x}_q, I_{q,\max}\right). \tag{31}$$



(a)

(b)

Fig. 8. Example of BH procedure applied to 1-D function f. The local optimization algorithm is applied to the initial point $\mathbf{a}$, yielding point $\mathbf{a}^*$. The next point $\mathbf{b}$ is obtained by adding a random vector $\mathbf{\Delta}_0$ to $\mathbf{a}$, and the local optimization algorithm is repeated, yielding point $\mathbf{b}^*$. (a) If the value of the objective function is improved (i.e., $f(\mathbf{a}^*) < f(\mathbf{b}^*)$), new hops are performed from point $\mathbf{b}$. (b) Otherwise, the probability of accepting point $\mathbf{b}$ is calculated using (32).

The runtime of the optimization process is dominated by the evaluation of the objective function. To overcome this limitation, the IMT-based analysis method is proposed in Section II, yielding several orders of magnitude faster evaluation of the voltage drops within a network. Unlike simulation-based optimization algorithms requiring prohibitive runtime for circuit evaluation, the proposed fast analysis algorithm requires significantly less runtime. This feature not only accelerates the optimization process but also enables exploration of a larger search space, potentially producing superior results.

Each voltage regulator adds two dimensions to the optimization problem. Due to the multidimensional nature of the regulator placement, the global optimization algorithm should be capable of exploring a space involving many variables. The BH algorithm [15] has gained significant attention in material physics for its effectiveness in crystal structure prediction [34], [35], computer vision [36], and interplanetary trajectory optimization [37].

Suppose the objective n-input function $f : \mathbb{R}^n \mapsto \mathbb{R}$ has multiple local optima, as illustrated in Fig. 8. A point $\mathbf{a} \in \mathbb{R}^n$ is initially selected, and the local optimization algorithm is applied to find the local minimum at $\mathbf{a}^*$. A random perturbation vector $\mathbf{\Delta}_0$ is added to initial point $\mathbf{a}$ to obtain the new point, $\mathbf{b} = \mathbf{a} + \mathbf{\Delta}_0$. This step, often referred to as hopping, allows the algorithm to escape the local minima. The local optimization step is repeated from the new point, converging to point $\mathbf{b}^*$. At this stage, the next hop can start from $\mathbf{a}$ or $\mathbf{b}$, The starting point is selected based on the Metropolis criterion adopted from simulated annealing [38]. If $f(\mathbf{b}^*) < f(\mathbf{a}^*)$, new point $\mathbf{b}$ is selected for the hop, as depicted in Fig. 8(a). Otherwise, new point $\mathbf{b}$ is selected with probability

$$p(\mathbf{a} \to \mathbf{b}) = \exp\left(-\frac{f\left(\mathbf{b}^*\right) - f(\mathbf{a}^*)}{T}\right) \tag{32}$$

where T is a temperature parameter [38] controlling the likelihood of accepting a suboptimal point [see Fig. 8(b)].

A. Local Optimization

Although the position of the voltage regulators within the grid is a discrete variable, the IMT technique enables formulation of the optimization problem as a continuous optimization. Efficient polynomial-time convex optimization algorithms can therefore be applied to approximate the solution. The closest integer coordinate is chosen as the solution.

Two local optimization algorithms are used in this framework. For those experiments where the position of the regulators is unconstrained, no linear constraints are applied to the optimization problem. A limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) [39] solver is used in this case. The L-BFGS algorithm belongs to a class of quasi-Newton convex optimization algorithms [40]. Unlike gradient descent, requiring a relatively large number of function evaluations, the L-BFGS algorithm incorporates the function curvature information, thus converging in fewer steps. Unlike Newton's method requiring expensive calculation of the Hessian matrix (matrix of second derivatives), the L-BFGS algorithm approximates the Hessian matrix based on prior iterations, requiring less memory and runtime.

Handling blockages requires the optimization algorithm to consider the constraint functions, as described in Section V-B. Since L-BFGS cannot directly handle the linear constraints, a different optimization algorithm is required. The objective function $f(\mathbf{x})$ and a vector of constraint functions $\mathbf{g}(\mathbf{x})$ are typically represented as a Lagrangian function [40]

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) \tag{33}$$

where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers. The solution is found by finding $\mathbf{x}, \boldsymbol{\lambda}$ such that

$$\nabla L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}. \tag{34}$$

In this work, the sequential least squares programming (SLSQP) algorithm is used [41]. SLSQP is a type of the sequential quadratic programming (SQP) algorithm, where the objective and constraint functions are reformulated as a sequence of quadratic programming problems

$$\begin{aligned} \min_{\mathbf{d}^k} \quad & \nabla f\left(\mathbf{x}^k\right) + \frac{1}{2}\mathbf{d}^T B_k \mathbf{d} \\ \text{s.t.} \quad & \nabla \mathbf{g}\left(\mathbf{x}^k\right)^T \mathbf{d}^k + \mathbf{g}\left(\mathbf{x}^k\right) \leq \mathbf{0} \end{aligned} \tag{35}$$

where $\mathbf{x}^k$ is the value of $\mathbf{x}$ at iteration k, $\mathbf{d} = \mathbf{x}^{k+1} - \mathbf{x}^k$, and B_k is the Hessian matrix of f at $\mathbf{x}^k$. Similar to L-BFGS, the SLSQP algorithm avoids the expensive calculation of the Hessian matrix by the series of least squares problems, reducing the runtime of the algorithm.

Two practical constraints are considered in subsequent experiments, namely, restricted position of the regulators and limited current of the regulators.

B. Restricted Position

In practical VLSI systems, a regulator is placed within the silicon layer, requiring significant area for placement and routing. Placing a regulator can therefore significantly contribute to congestion, adversely affecting system performance. The placement of the regulators can therefore be limited to less congested regions within the layout. This limitation is described by the constraint

$$\mathbf{x}_q \in A \subseteq U \quad \forall s_q \in \mathcal{S} \tag{36}$$

where A is the set of whitespace nodes, i.e., unoccupied positions available for placing voltage regulators, and U is the set of all nodes within the grid. Constraint (36) restricts the position of the voltage regulators within a grid to those regions capable of accommodating regulators.

The information describing the congestion is typically not available during the power network design process [42]. An alternative metric for estimating congestion is therefore necessary. Since congestion at the bottom layers of an IC is driven primarily by the devices at the silicon layer [43], proximity to the load circuitry is likely correlated with routing congestion. Based on this observation, a proxy metric for evaluating routing congestion is described here.

Suppose $I : U \to \mathbb{R}$ is a function denoting the current drawn at node $(x, y) \in U$ within the power grid. The score $C : U \to \mathbb{R}$ is defined here as the total load current within distance $l_{\max}$ of the node $(x, y) \in U$

$$C(x, y) = \sum_{l(x_0, y_0) < l_{\max}} I(x_0, x_0) \tag{37}$$

where

$$l(x_0, y_0) = \sqrt{(x_0 - x)^2 + (y_0 - y)^2}. \tag{38}$$

By using the score function, those areas in close proximity to hot spots are determined. Furthermore, due to the summation of currents within a specified radius, those loads spread over significant area are given higher weight as compared to isolated loads. An example of scores in selected ibmpg benchmark circuits is shown in Fig. 9. By selecting the regions with the largest scores, a portion of the layout is marked unavailable for regulator placement, as shown in Fig. 10. The black and gray regions represent, respectively, the 85th and 70th percentiles of the current scores within the layout. In the case studies described in Section VI, these regions are excluded from the available whitespace.

C. Restricted Current

Due to physical limitations, the regulators cannot provide arbitrarily large currents. Large currents produced by a regulator can generate excessive heat, complicating the thermal management process, increasing the risk of electromigration, and potentially damaging the regulator and surrounding circuitry [44], [45]. The regulators are therefore equipped with

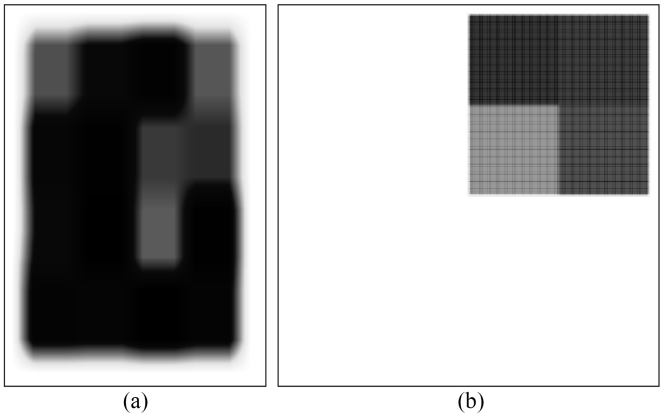


Fig. 9. Current score, (a) ibmpg2 and (b) ibmpg5. A darker color indicates those regions with a higher score.

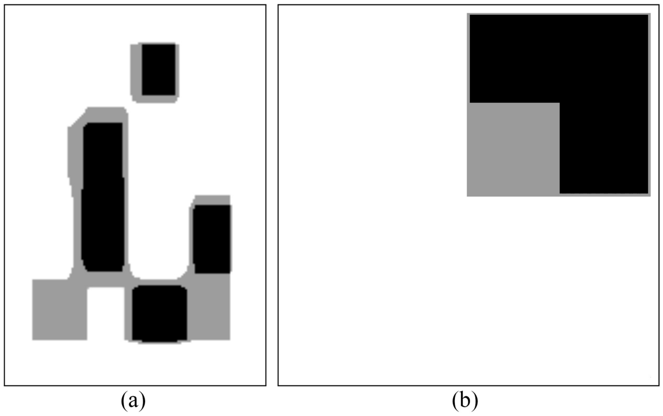


Fig. 10. Blocked regions, (a) ibmpg2 and (b) ibmpg5. In the case studies, the entire layout is initially available for placement. During the second experiment, the placement is prohibited within the black regions. During the final experiment, the placement is additionally prohibited within the gray regions.

current limiting modules preventing excessive currents [46]. In this article, the current limit for each voltage regulator is assumed to be known *a priori* and is expressed by the constraint

$$\mathbf{i}(\mathcal{S}) \leq \mathbf{i}_{\max} \tag{39}$$

where $\mathbf{i}(\mathcal{S})$ and $\mathbf{i}_{\max}$ are vectors of, respectively, the estimated and maximum currents supplied by each regulator. The proposed IMT-based analysis algorithm is extended to support the limited current capacity of the voltage regulators, as described in Section II-B.

For simplicity, an equal current limit is imposed on each voltage regulator in the case studies. The maximum current supplied by a voltage regulator s is

$$I_{\max}(s) = (1+\eta) \times \frac{\mathbf{1}_{1,n}\mathbf{i}(\mathcal{L})}{m}. \tag{40}$$

The parameter value $\eta = 0.2$ is chosen, i.e., the current supplied by any regulator is at most 20% higher than the average regulator current. This constraint ensures a more even distribution of current supplied by each regulator.

VI. Case Studies

The analysis and optimization algorithms are implemented in Python and applied to IBM power grid benchmark circuits [23]. The algorithms are run on a Linux workstation powered by a dual core 2.3-GHz Intel Core i5 processor with 16 GB of RAM. The BH optimization algorithm is used in this case study, as described in Section V. A maximum of 50 BH iterations is permitted during each experiment. Each local optimization is terminated when the improvement in the objective function does not exceed 10^{-3} volts. As described in Section II, parameters N_x and N_y are set to 2. The number of clusters in each algorithm is varied from 100 in ibmpg2 to 400 in ibmpg5 and ibmpg6. The initial distribution of regulators is generated using a quasirandom Sobol sequence [47] to evenly distribute the regulators within the available whitespace. The voltage within the power grid is evaluated using HSPICE [24] before and after the optimization process. The number of voltage regulators is varied from 5 to 100. A total of six experiments is performed. The per cent of the layout excluded from the whitespace is varied from 0 to 30%. In the first three experiments, the current supplied by each regulator is unlimited, while in the latter three experiments, the current sourced by each regulator is limited to 120% of the average current supplied by the regulators.

A. Results of Experiments

The results of the experiments are summarized in Table II. The relative improvement is

$$\texttt{imp.} = \frac{v_d^{\text{init}} - v_d}{v_{\text{drop}}^{\text{init}}}$$

where the initial voltage drop $v_{\text{drop}}^{\text{init}}$ corresponds to an even distribution of regulators within the 2-D space produced by the pseudorandom Sobol sequence [47]. The final voltage drop v_d is determined by analyzing the circuit with the regulators placed at the coordinates suggested by the placement algorithm. Consistent with expectations, additional regulators provide superior regulation, reducing the voltage drop in all of the experiments. In all of the cases, the placement algorithm improves the power quality as compared to the initial placement. Note, however, that with additional regulators, both v_d^{init} and v_d are reduced. A small initial voltage v_d^{init} makes further improvement more challenging, reducing the relative improvement.

Observe that in most experiments the smallest voltage drop is achieved using unrestricted placement. Limiting the current supplied by the regulators does not significantly affect the voltage drop. In certain cases, improved results are achieved assuming limited regulator current. This behavior can be partially explained by the effect of limiting current on the voltage drop. The inadequate current supplied by a regulator incentivizes the optimization algorithm to move other regulators closer to the hot spots, improving convergence of the algorithm.

In these experiments, restricting the position of the regulator has a more significant effect on the voltage drop. This effect can be explained by the choice of blocked region. Since

TABLE II
VOLTAGE DROP, RELATIVE IMPROVEMENT, AND RUNTIME ACHIEVED IN CASE STUDIES. v_d, `imp.`, AND t DENOTE, RESPECTIVELY, THE MAXIMUM VOLTAGE DROP, IMPROVEMENT (RELATIVE TO THE INITIAL QUASIRANDOM DISTRIBUTION), AND COMPUTATIONAL RUNTIME. THE PERCENTAGES IN THE SQUARE BRACKETS DEPICT THE IMPROVEMENT RELATIVE TO THE INITIAL (EVEN) REGULATOR PLACEMENT

		Unlimited current ($\eta = \infty$)									Limited current ($\eta = 0.2$)								
		0% blocked			15% blocked			30% blocked			0% blocked			15% blocked			30% blocked		
	m	v_d, V	imp.	t, s	v_d, V	imp.	t, s	v_d, V	imp.	t, s	v_d, V	imp.	t, s	v_d, V	imp.	t, s	v_d, V	imp.	t, s
ibmpg2	5	10.27	35%	5.21	10.11	51%	5.49	9.91	66%	29.31	9.72	55%	5.47	11.64	67%	8.21	14.59	70%	16.44
	10	4.97	68%	12.25	6.38	74%	41.51	4.99	84%	54.96	5.48	79%	14.39	4.89	84%	19.21	5.21	88%	67.73
	20	2.48	46%	23.64	3.15	56%	53.60	2.62	72%	212.74	2.77	65%	56.96	3.08	72%	276.16	2.37	82%	319.84
	50	1.09	60%	60.83	1.64	65%	402.72	2.03	67%	548.79	1.29	74%	362.18	1.76	73%	2022.67	2.17	78%	3547.33
	100	0.76	51%	394.12	0.78	65%	1459.11	1.22	59%	2491.98	0.74	67%	533.35	1.20	66%	2563.73	1.55	69%	2238.51
ibmpg3	5	24.95	48%	3.82	21.77	61%	5.76	26.64	70%	8.19	22.68	65%	4.85	24.75	75%	15.46	21.52	82%	20.94
	10	12.76	40%	4.03	13.27	53%	27.20	12.71	67%	43.39	11.87	59%	21.98	13.33	71%	23.38	10.96	81%	135.00
	20	6.14	53%	8.03	7.56	57%	55.71	6.67	75%	121.22	7.57	60%	35.58	5.62	78%	224.62	6.92	85%	302.46
	50	3.36	45%	85.31	4.76	49%	607.49	3.24	73%	477.76	4.25	52%	76.57	3.64	75%	1591.49	2.84	84%	2287.90
	100	2.23	23%	445.95	2.92	26%	1499.69	1.99	61%	2266.70	3.07	21%	299.76	1.88	64%	2308.84	2.57	71%	4250.05
ibmpg4	5	3.41	14%	0.49	3.24	44%	1.85	3.70	56%	0.65	4.21	33%	1.97	2.87	65%	12.97	3.05	73%	9.30
	10	1.83	27%	0.90	1.70	47%	10.50	2.32	47%	54.33	1.47	53%	7.46	2.10	60%	9.03	2.36	66%	108.61
	20	0.84	39%	1.64	0.84	50%	80.48	1.27	49%	157.86	0.89	59%	5.86	0.94	67%	248.96	1.98	55%	523.96
	50	0.48	18%	62.76	0.44	44%	204.10	0.56	46%	739.40	0.41	45%	79.72	0.47	59%	970.13	0.53	63%	1209.06
	100	0.26	36%	294.94	0.25	53%	1555.94	0.41	53%	3161.00	0.31	46%	241.14	0.31	60%	933.43	0.36	66%	1895.52
ibmpg5	5	0.177	63%	0.33	0.205	71%	2.58	0.349	62%	13.79	0.234	72%	2.32	0.213	85%	14.03	0.312	85%	18.52
	10	0.153	36%	0.84	0.141	53%	0.98	0.177	58%	76.87	0.167	50%	9.92	0.192	71%	61.17	0.276	67%	197.69
	20	0.058	56%	2.83	0.064	63%	133.24	0.107	55%	94.11	0.115	44%	23.13	0.109	68%	252.06	0.268	49%	534.44
	50	0.042	59%	29.45	0.044	66%	414.43	0.056	72%	551.18	0.086	42%	878.70	0.081	71%	947.97	0.234	43%	2093.12
	100	0.028	39%	309.63	0.038	54%	1156.57	0.035	71%	926.05	0.089	15%	3928.31	0.126	20%	2748.73	0.195	9%	2858.97
ibmpg6	5	2.18	8%	1.67	2.52	25%	7.28	2.70	40%	13.15	2.09	37%	3.72	2.15	53%	18.20	3.12	60%	34.61
	10	1.13	23%	2.31	1.11	46%	45.85	1.56	56%	37.46	1.14	55%	21.72	1.20	65%	46.65	1.30	70%	53.92
	20	0.66	28%	2.72	0.56	51%	84.12	1.00	51%	201.27	0.79	44%	18.19	0.67	64%	718.08	1.01	66%	411.70
	50	0.33	20%	48.88	0.35	43%	532.92	0.41	53%	1319.43	0.39	46%	86.92	0.32	66%	1061.68	0.58	57%	2595.12
	100	0.18	12%	518.17	0.20	35%	1728.70	0.25	43%	4385.17	0.23	31%	487.98	0.20	56%	4239.72	0.31	48%	2619.98

the congested regions carry the greatest load current, moving the regulators from these regions significantly increases the distance between the regulators and the loads, increasing the voltage drop.

B. Runtime

Three factors affect the runtime of the placement process, namely, the number of regulators and the two constraints, which require additional processing time during the placement process. A superlinear relationship is observed between the runtime and number of regulators (and consequently, the number of optimization variables). This factor is driven primarily by the cubic complexity of the SLSQP solver [48] combined with the quadratic complexity of the grid analysis process. Note that the runtime of the optimization process does not increase with grid size.

The computational time is additionally influenced by imposing constraints on the optimization process. The runtime of the IMT analysis with limited current depends upon the number of iterations to achieve $S^* = \varnothing$. In those benchmarks where the current is unevenly distributed, such as `ibmpg2` and `ibmpg5`, the initial placement unevenly distributes the current among the regulators. Those regulators in proximity to the loads supply greater current as compared to the regulators farther from the loads. A larger portion of regulators therefore initially operates beyond the current capacity. The IMT algorithm with limited current requires multiple iterations to converge, degrading the runtime. In contrast, the load current is relatively evenly distributed in `ibmpg3`, `ibmpg4`, and `ibmpg6`. The initial placement is more likely to evenly distribute the current load among the regulators, with few regulators violating the current limit. The runtime with limited and unlimited current will therefore likely be similar. Due to the stochastic nature of the BH algorithm, the directions of the random hopping may greatly affect the runtime, producing significant deviations. For example, as observed in `ibmpg3` for $m = 100$, the runtime is larger with unlimited current.

Finally, constraining the position of the voltage regulators significantly increases the runtime. As mentioned in Section V, the L-BFGS algorithm changes to SLSQP when positional constraints are applied. Since L-BFGS is best suited for unconstrained optimization, faster performance is achieved.

In addition to a relatively slower constrained optimization, the runtime is affected by the size and position of the blockages. Larger blockages typically require more constraints to be described, as in the case of `ibmpg2`, degrading the runtime. Therefore, in most cases, the runtime increases when a larger percentage of the layout is blocked. Note however that the local optimization is often less effective with larger blockages due to the limited local search. This effect is observed in `ibmpg2` ($\eta = 0.2$, $m = 100$), `ibmpg5` ($\eta = \infty$, $m = 100$), and `ibmpg6` ($\eta = 0.2$, $m = 100$). Increasing the blockage size from 15% to 30% produces more unsuccessful basinhopping iterations, quickly reaching the maximum number of iterations. In these cases, the runtime is smaller with the larger blockage, while the quality of the optimization is degraded. Increasing the temperature parameter to step over large blockages and allowing more BH iterations may help to overcome this issue.

C. Comparison With Prior Works

Two works investigating a similar problem exist in [7], [14]. In both of these works, however, reproducibility of the results is limited. In [7], low-dropout regulators are distributed within the layout. The maximum voltage drop is used as an objective function and is evaluated by a transient analysis of the power grid accelerated by a GPU. Positional constraints are imposed on the placement of an LDO. Unfortunately, the information necessary to reproduce the experiments, including the power network structure, load current, and grid resistivity, is not reported in [7]. It is however possible to compare the two metrics, namely, runtime and voltage drop improvement. Due to the significantly more expensive objective function, the distribution of two LDOs within a circuit with 17 000 nodes requires 49 min. The similarly sized `ibmpg2` requires fewer than 30 s for placing five regulators. In [7], 21 LDOs within a grid with two million nodes require 62.5 h. In contrast, distributing 20 regulators within `ibmpg5` with over three million nodes requires only 19.5 min using the IMT-based placement algorithm. In the case studies described here, the improvement in voltage drop ranges from 8% to 88%, as shown in Table II. A larger improvement is observed if the position of the regulator is restricted. This observation is consistent with [7] where an 80% to 90% improvement as compared to an even allocation of the regulators is observed. In the experiments described here, the blockages are chosen in those regions with the highest load current. The placement of the regulators therefore provides insufficient current to those areas with larger loads. This effect is particularly noticeable if the loads are unevenly distributed within the layout, as in `ibmpg5` [see Fig. 9b].

The framework proposed in [14] is tested using the `superblue5` benchmark circuit with over 550 000 nodes, requiring up to 5 min to complete the optimization process. The case study using the `superblue5` benchmark circuit has been repeated using the IMT-based placement algorithm. The resistivity and total load current of the power grid are adapted from the similarly sized `ibmpg4`. The current of the individual blocks is assumed proportional to the physical area. Note however that only the dimensions of the power grid are reported in [14], rendering a direct comparison not representative. For completeness, however, the results are compared in Table III. Note that if the number of regulators is small, the proposed algorithm exhibits significantly better runtime, while the runtime is significantly larger with 101 regulators. The faster runtime of the IMT algorithm can be explained by the fixed number of BH iterations combined with the quadratic scaling of the objective function with the number of regulators.

VII. Holistic Power Network Design

Placement of the voltage regulators is an important part of the power distribution network design process for high-performance integrated systems. Several other aspects of power management exist, including dynamic voltage and frequency scaling (DVFS), power and clock gating, and multiple voltage domains [1]. Designing higher quality power distribution networks requires a holistic co-design process considering both the static power network characteristics (e.g., number of voltage domains) and dynamic power management techniques (e.g., regulators, DVFS, gating). In this section, several strategies for incorporating these concepts into the power distribution network design process are described.

TABLE III
Comparison of Runtime of IMT Algorithm With [14] Based on the `superblue5` Benchmark Circuit. The Number of Regulators Is Adjusted to Match the Combined Number of LDOs and Decoupling Capacitors

This work		[14]	
#`reg.`	t, s	#`LDO` + `Cap`	t, s
5	0.37	5	103.48
25	2.16	25	309.28
101	381.24	101	209.41

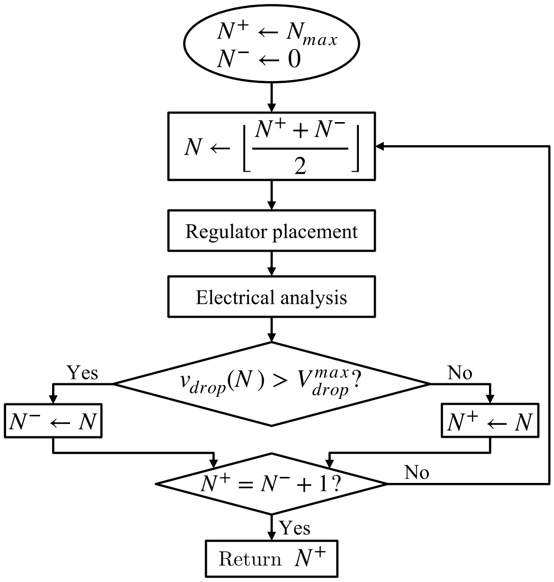


Fig. 11. Binary search procedure to determine the minimum number of regulators N_+ required to comply with the voltage drop constraint $V_{\text{drop}}^{\text{max}}$.

A. Minimizing the Number of Regulators

The optimization setup described in Section V assumes a fixed number of voltage regulators. In practical systems, however, the number of voltage regulators is not known in advance; rather, an upper limit on the power noise $V_{\text{drop}}^{\text{max}}$ is provided. A procedure to determine the minimum number of regulators is proposed in [7], where the number of regulators is incrementally increased. This procedure can be enhanced by adopting a binary search approach, as shown in Fig. 11. Suppose a maximum of N_{max} regulators are placed within the system. During the first iteration, the power network is optimized using $\lfloor N_{\text{max}}/2 \rfloor$ regulators. Accurate circuit analysis is necessary to precisely determine the voltage drop. If the final placement satisfies the target voltage drop, the number of regulators is reduced; otherwise, the number of regulators is increased. The algorithm terminates after finding the minimum number of regulators $N^+ \leq N_{\max}$ satisfying the voltage drop constraint. Since a binary search procedure is adopted, the expected number of iterations is $\log_2 N_{\max}$.

B. Regulation-Aware Power Network Design

Modern high-performance integrated systems typically utilize multiple voltage domains to separate the modules requiring a high voltage (e.g., RF and analog circuits) and low voltage (e.g., memory and digital circuitry) [16]. A lower supply voltage significantly reduces the dynamic power consumption. Each additional voltage domain however utilizes a separate power network, requiring additional on-chip area and metal. Complex tradeoffs therefore exist among the power quality, power efficiency, and cost.

Separate voltage domains produce multiple power networks and an associated set of loads. Each power network can be efficiently analyzed using the proposed framework to determine the power quality and minimum number of regulators. By considering the placement of the regulators during the power network design process, a holistic co-design procedure is enabled, potentially yielding a superior power management solution.

C. Dynamic Power Management

Modern VLSI systems utilize dynamic power management techniques, such as DVFS, and power and clock gating. DVFS is a power management technique where the supply voltage and clock frequency are adjusted based on workload demands. By reducing the voltage and frequency during low activity, power consumption is reduced, while scaling up during high demand periods maximizes performance. Power and clock gating can be considered as extreme cases of DVFS. The idle circuit blocks are disconnected from the power and clock distribution networks, reducing the leakage power and capacitive load.

The proposed placement methodology minimizes the voltage drop in response to a particular number, location, and magnitude of the current loads. With dynamic power management, each of k operating scenarios produces a distinct power network $\mathcal{L}_i$, $0 < i \leq k$, with a different location and magnitude of the load currents. The objective function can therefore be transformed to

$$v_{\text{drop}}(\mathcal{S}) = -\min_{i=1}^{k} \left(\mathbf{v}^{\mathbf{g}}(\mathcal{L}_i)\right)|_{\mathcal{S}} \tag{41}$$

where the position of the regulators is optimized for the worst-case voltage drop across all scenarios.

VIII. Conclusion

To tackle stringent power quality and efficiency requirements in modern VLSI complexity systems, heterogeneous power regulation is necessary, incorporating both off-chip as well as on-chip point-of-load voltage regulators. In practical systems, the voltage regulators are however limited in number and current capacity. The whitespace available for regulator placement is also limited. A voltage regulator allocation algorithm considering these constrains is presented in this article. With the IMT-based grid analysis method, the on-chip power distribution system is efficiently analyzed, enabling a large number of placement options to be evaluated during the optimization process. The proposed algorithm is independent of the size of the grid, enabling the efficient analysis of large scale power networks. The technique is validated using the IBM power grid benchmark suite. With the proposed algorithm, the parasitic voltage drop is reduced by up to 88%. The computational runtime is reduced by several orders of magnitude as compared to placement tools based on MNA circuit analysis.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments.

References

[1] I. Partin-Vaisband, R. Jakushokas, M. Popovich, A. V. Mezhiba, S. Köse, and E. G. Friedman, *On-Chip Power Delivery and Management*, 4th ed. Cham, Switzerland: Springer. 2016.

[2] C. Wang et al., "An efficient approach for power delivery network design with closed-form expressions for parasitic interconnect inductances," *IEEE Trans. Adv. Packag.*, vol. 29, no. 2, pp. 320–334, May 2006.

[3] E. A. Burton et al., "FIVR–Fully integrated voltage regulators on 4th generation Intel Core SoCs," in *Proc. IEEE Appl. Power Electron. Conf. Expos.*, Mar. 2014, pp. 432–439.

[4] D. Hackenberg, R. Schöne, T. Ilsche, D. Molka, J. Schuchart, and R. Geyer, "An energy efficiency feature survey of the Intel Haswell processor," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshop*, May 2015, pp. 896–904.

[5] N. Butzen and M. S. J. Steyaert, "Scalable parasitic charge redistribution: Design of high-efficiency fully integrated switched-capacitor DC–DC converters," *IEEE J. Solid-State Circuits*, vol. 51, no. 12, pp. 2843–2853, Dec. 2016.

[6] C. Schaef et al., "A IMax fully integrated multi-phase voltage regulator with 91% peak efficiency at 1.8 to 1V, operating at 50MHz and featuring a digitally assisted controller with automatic phase shedding and soft switching in 4nm class FinFET CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2022, pp. 1–3.

[7] T. Yu and M. D. F. Wong, "Efficient simulation-based optimization of power grid with on-chip voltage regulator," in *Proc. ACM/IEEE Asia South Pacif. Design Autom. Conf.*, Jan. 2014, pp. 531–536.

[8] S. X.-D. Tan, C.-J. R. Shi, and J.-C. Lee, "Reliability-constrained area optimization of VLSI power/ground networks via sequence of linear programmings," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 22, no. 12, pp. 1678–1684, Dec. 2003.

[9] M. Zhao, Y. Fu, V. Zolotov, S. Sundareswaran, and R. Panda, "Optimal placement of power-supply pads and pins," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 1, pp. 144–154, Jan. 2006.

[10] K. Wang, B. H. Meyer, R. Zhang, M. R. Stan, and K. Skadron, "Walking pads: Managing C4 placement for transient voltage noise minimization," in *Proc. ACM/IEEE Design Autom. Conf.*, Jun. 2014, pp. 1–6.

[11] S. Köse and E. G. Friedman, "Distributed on-chip power delivery," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 4, pp. 704–713, Dec. 2012.

[12] M. Chang, "Power distribution network optimization for on-die regulator with Laplace transform technique," in *Proc. IEEE Electr. Design Adv. Packag. Syst.*, Dec. 2020, pp. 1–3.

[13] Z. Zeng, X. Ye, Z. Feng, and P. Li, "Tradeoff analysis and optimization of power delivery networks with on-chip voltage regulation," in *Proc. AMC/IEEE Design Autom. Conf.*, Jun. 2010, pp. 831–836.

[14] S. A. Sadat, M. Canbolat, and S. Köse, "Optimal allocation of LDOs and decoupling capacitors within a distributed on-chip power grid," *ACM Trans. Design Autom. Electron. Syst.*, vol. 23, no. 4, pp. 1–15, Jul. 2018.

[15] D. J. Wales and J. P. K. Doye, "Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms," *J. Phys. Chem. A*, vol. 101, no. 28, pp. 5111–5116, Jul. 1997.

[16] R. Bairamkulov, A. Roy, M. Nagarajan, V. Srinivas, and E. G. Friedman, "SPROUT–Smart power routing tool for board-level exploration and prototyping," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 41, no. 7, pp. 2263–2275, Jul. 2022.

[17] A. Mezhiba and E. Friedman, "Scaling trends of on-chip power distribution noise," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 4, pp. 386–394, Apr. 2004.

[18] R. Bairamkulov and E. G. Friedman, *Graphs in VLSI*. Cham, Switzerland: Springer. 2022.

[19] C. Ho, A. Ruehli, and P. Brennan, "The modified nodal approach to network analysis," *IEEE Trans. Circuits Syst.*, vol. CS-22, no. 6, pp. 504–509, Jun. 1975.

[20] S. Köse and E. G. Friedman, "Effective resistance of a two layer mesh," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 58, no. 11, pp. 739–743, Nov. 2011.

[21] R. Bairamkulov and E. G. Friedman, "Effective resistance of finite two-dimensional grids based on infinity mirror technique," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 9, pp. 3224–3233, Sep. 2020.

[22] L. A. Wolsey and G. L. Nemhauser, *Integer and Combinatorial Optimization*. New York, NY, USA: Wiley, 1999.

[23] S. R. Nassif, "Power grid analysis benchmarks," in *Proc. ACM/IEEE Asia South Pacif. Design Autom. Conf.*, Mar. 2008, pp. 376–381.

[24] *HSPICE Quick Reference*, Mountain View, CA, USA: Synopsys, Mar. 2017.

[25] I. Vaisband and E. G. Friedman, "Heterogeneous methodology for energy efficient distribution of on-chip power supplies," *IEEE Trans. Power Electron.*, vol. 28, no. 9, pp. 4267–4280, Sep. 2013.

[26] K. Wang and M. Marek-Sadowska, "On-chip power-supply network optimization using multigrid-based technique," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 3, pp. 407–417, Mar. 2005.

[27] A. V. Mezhiba and E. G. Friedman, "Inductive properties of high-performance power distribution grids," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 10, no. 6, pp. 762–776, Dec. 2002.

[28] Z. Feng, Z. Zeng, and P. Li, "Parallel on-chip power distribution network analysis on multi-core-multi-GPU platforms," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 10, pp. 1823–1836, Oct. 2011.

[29] J. N. Kozhaya, S. R. Nassif, and F. N. Najm, "A multigrid-like technique for power grid analysis," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 21, no. 10, pp. 1148–1160, Oct. 2002.

[30] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion?" *J. Classif.*, vol. 31, no. 3, pp. 274–295, Oct. 2014.

[31] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[32] D. Sculley, "Web-scale K-means clustering," in *Proc. Int. Conf. World Wide Web*, Apr. 2010, pp. 1177–1178.

[33] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM Sigmod Rec.*, vol. 25, no. 2, pp. 103–114, Jun. 1996.

[34] C. J. Burnham and N. J. English, "Crystal structure prediction via basin-hopping global optimization employing tiny periodic simulation cells, with application to water–ice," *J. Chem. Theory Comput.*, vol. 15, no. 6, pp. 3889–3900, May 2019.

[35] S. Yang and G. M. Day, "Exploration and optimization in crystal structure prediction: Combining basin hopping with quasi-random sampling," *J. Chem. Theory Comput.*, vol. 17, no. 3, pp. 1988–1999, Feb. 2021.

[36] J. Kwon and K. M. Lee, "Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping Monte Carlo sampling," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, Jun. 2009, pp. 1208–1215.

[37] J. A. Englander and A. C. Englander, "Tuning monotonic basin hopping: Improving the efficiency of stochastic search as applied to low-thrust trajectory optimization," in *Proc. Int. Symp. Space Flight Dyn.*, May 2014, pp. 1–33.

[38] M. Guo, Y. Liu, and J. Malec, "A new Q-learning algorithm based on the metropolis criterion," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 34, no. 5, pp. 2140–2143, Oct. 2004.

[39] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1, pp. 503–528, Aug. 1989.

[40] P. T. Boggs and J. W. Tolle, "Sequential quadratic programming," *Acta Numer.*, vol. 4, pp. 1–51, Jan. 1995.

[41] D. Kraft, *TOMP: FORTRAN Modules for Optimal Control Calculations*. Düsseldorf, Germany: VDI-Verlag, 1991.

[42] R. Bairamkulov, K. Xu, M. Popovich, J. S. Ochoa, V. Srinivas, and E. G. Friedman, "Power delivery exploration methodology based on constrained optimization," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 9, pp. 1916–1924, Sep. 2020.

[43] R. Kirby, S. Godil, R. Roy, and B. Catanzaro, "CongestionNet: Routing congestion prediction using deep graph neural networks," in *Proc. IFIP/IEEE Int. Conf. Very Large Scale Integr.*, Dec. 2019, pp. 217–222.

[44] H. Q. Tay, V. T. Nam, N. H. Duc, and B. N. Chau, "A current sensing circuit using current-voltage conversion for PMOS-based LDO regulators," in *Proc. IEEE Int. Symp. Comput. Appl. Ind. Electron.*, Dec. 2012, pp. 1–4.

[45] J. A. De Lima and W. A. Pimenta, "A current limiter for LDO regulators with internal compensation for process and temperature variations," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2008, pp. 2238–2241.

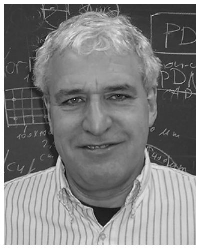
[46] J. Li et al., "An adaptively biased LDO regulator with 11nA quiescent current and 50mA available load," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2021, pp. 1–5.

[47] H. Niederreiter, "Low-discrepancy and low-dispersion sequences," *J. Number Theory*, vol. 30, no. 1, pp. 51–70, Sep. 1988.

[48] M. Konakovic Lukovic, Y. Tian, and W. Matusik, "Diversity-guided multi-objective Bayesian optimization with batch evaluations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 17708–17720.

Rassul Bairamkulov received the B.Eng. degree in electrical and electronic engineering from Nazarbayev University, Astana, Kazakhstan, in 2016, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Rochester, Rochester, NY, USA, in 2018 and 2022, respectively.

In Summer 2018 and 2020, he interned with Power Design Team, Qualcomm Inc., San Diego, CA, USA. He is currently a Postdoctoral Scholar with the Integrated Systems Laboratory, École polytechnique fédérale de Lausanne, Lausanne, Switzerland. His current research interests include power integrity, logic synthesis, and electronic design automation of conventional and emerging VLSI technologies.



Eby G. Friedman (Life Fellow, IEEE) received the B.S. degree in electrical engineering from Lafayette College, Easton, PA, USA, in 1979, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Irvine, Irvine, CA, USA, in 1981 and 1989, respectively.

He was with Hughes Aircraft Company, Glendale, CA, USA, from 1979 to 1991, rising to Manager of the Signal Processing Design and Test Department, where he was responsible for the design and test of high-performance digital and analog ICs. He has been with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA, since 1991, where he is a Distinguished Professor and the Director of the High Performance VLSI/IC Design and Analysis Laboratory. He is also a Visiting Professor with the Technion–Israel Institute of Technology, Haifa, Israel. He has authored almost 600 articles and book chapters and authored or edited 21 books in the fields of high-speed and low-power CMOS design techniques, 3-D design methodologies, high-speed interconnect, superconductive circuits, and the theory and application of synchronous clock and power distribution networks, and he holds 29 patents. His current research and teaching interests include high-performance synchronous digital and mixed-signal circuit design and analysis with application to high-speed portable processors, low-power wireless communications, and server farms.

Dr. Friedman is a recipient of the IEEE Circuits and Systems Mac Van Valkenburg Award, the IEEE Circuits and Systems Charles A. Desoer Technical Achievement Award, the University of Rochester Graduate Teaching Award, and the College of Engineering Teaching Excellence Award. He was the Editor-in-Chief (EIC) and the Chair of the Steering Committee of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS and the EIC of the *Microelectronics Journal*, a Regional Editor of the JOURNAL OF CIRCUITS, SYSTEMS AND COMPUTERS, an editorial board member of numerous journals, and a program and technical chair of several IEEE conferences. He is a Senior Fulbright Fellow, a National Sun Yat-sen University Honorary Chair Professor, and an Inaugural Member of the UC Irvine Engineering Hall of Fame.