Chapter 9
Exploratory methodology for power delivery

Power delivery is pivotal to the performance of modern integrated systems [381]. Violating limitations in power delivery such as load voltage drop, thermal characteristics, and power dissipation, may cause a variety of issues, such as circuit malfunction or overheating. Due to the high level of complexity in modern systems, it is difficult to monitor power delivery characteristics throughout the system development process [65]. This approach adds risks to the entire development flow. Unsatisfied power quality constraints at later stages of the design process may require unacceptable time and resources.

One strategy for reducing the burden of modifying the power network is overdesign, such as using additional interconnections and pins for power or larger and more numerous decoupling capacitors. This strategy increases cost and allocates less metal and pin resources for signaling, and less area for the functional circuitry [528]. In addition, external factors, such as cooling power or cost, shift the resulting system even farther from the optimal objective.

Numerous works on power delivery optimization at varying levels of abstraction exist in the literature. On-chip voltage regulation is discussed in [359, 495, 511, 529]. In [511], a framework for combining switching and linear regulators within a single system is presented that combines high efficiency linear regulators with superior regulation characteristics in switching converters. Power management has been deeply investigated from an architectural perspective. The work of [530] presents a framework for system-wide dynamic voltage scaling with thermal considerations that improves overconstrained circuits based on worst case scenarios. In [531], the GradualSleep strategy has been proposed to minimize on-chip static energy dissipation. More recent works describe paradigms suitable for modern circuit-level power management solutions. A system-level framework for optimizing decoupling capacitor and parasitic inductance is proposed in [145, 532]. A system-level power management system is described in [533], where the electrical and thermal characteristics are monitored to make appropriate adaptations, such as

R. Bairamkulov, E. G. Friedman, *Graphs in VLSI*,
https://doi.org/10.1007/978-3-031-11047-4_9

dynamic voltage and frequency scaling (DVFS) based on system temperature and workload.

Despite the maturity of the field, power delivery in VLSI systems is rarely approached from a constrained optimization perspective. In [534], quadratic programming methods are exploited to reduce the impedance profile of the power delivery network at frequencies of interest by sacrificing the impedance at less relevant frequencies. More recent work [535] utilizes differential evolutionary optimization to suggest the impedance profile of a physical structure. A significant omission in the literature is the almost exclusive focus on optimizing the electrical parameters, only indirectly addressing external metrics such as power and cost. Constrained global optimization provides a natural framework for design exploration of power delivery systems. The primary strength of the proposed technique is flexibility, allowing different design objectives and constraints to be considered including thermal and cost parameters. The subsequent sections provide a deeper insight into this proposed methodology. In Section 9.1, the necessary components of the proposed framework are described. Two case studies are presented in Section 9.2 to demonstrate the validity and discuss the strengths and limitations of the proposed approach The chapter is concluded by a summary in Section 9.3.

9.1 Optimization framework

The standard design process in the absence of power network design exploration is shown in Fig. 9.1 [536]. Due to the lack of preliminary information, power delivery network analysis is performed during the placement and routing stage [536]. If the circuit does not comply with power quality and voltage drop objectives, the power network is changed or resynthesized. The verification and redesign processes repeat until the resulting power network satisfies the required specifications. Due to the significant time required to evaluate and refine the power delivery network at the system level, multiple design iterations at later stages of the development process are highly undesirable, as these changes may cause delays on the order of days.

To mitigate potential losses, the number of power network redesigns needs to be minimized, preferably to zero. Power delivery exploration can provide valuable guidelines for power network synthesis, bringing the resulting system close to the optimal state. Two important characteristics of the early design stages are worth noting. First, the lack of accurate electrical data creates a high degree of uncertainty in the power network development process. The assumptions made at this stage are crucial. Second, before the primary design parameters are fixed, a high degree of flexibility exists. For example, the number of voltage domains may significantly affect the efficiency of the system at the expense of additional metal resources or increased power noise. Exploiting these tradeoffs is crucial to unlocking the full potential of the overall power delivery system.

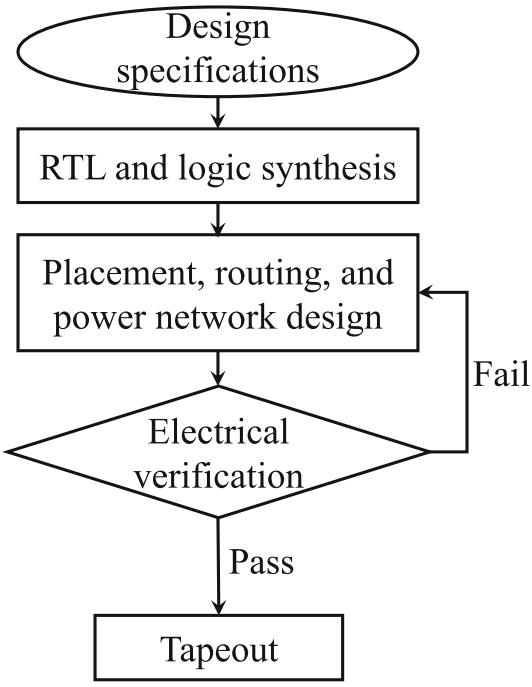

Fig. 9.1 Conventional IC development process [536]

The proposed power delivery exploration process is illustrated in Fig. 9.2. The process is general and varies greatly with different inputs. The process starts with the analysis of the design specifications. A model of the power network is used to estimate the electrical metrics. Non-electrical metrics of interest are also identified and certain design flexibilities are identified. After the required components are characterized, the functions are passed to optimization algorithms. The result of the optimization process is a set of design guidelines that ensure proper operation without excessive overdesign. A more detailed explanation of the proposed exploration process is provided in the following subsections.

9.1.1 Specification of the electrical design requirements

A model of the power delivery network consists of four components: topology, voltage sources, load currents, and impedances. The topology reflects the relative placement of the elements within the netlist, supporting a comprehensive circuit analysis process. Technology information, such as the number of metal layers or interconnect conductivity, and design specifications, such as the interconnect dimensions, determine the parameters of the power network model [384]. One of the simplest and most widespread power network models is the hierarchical

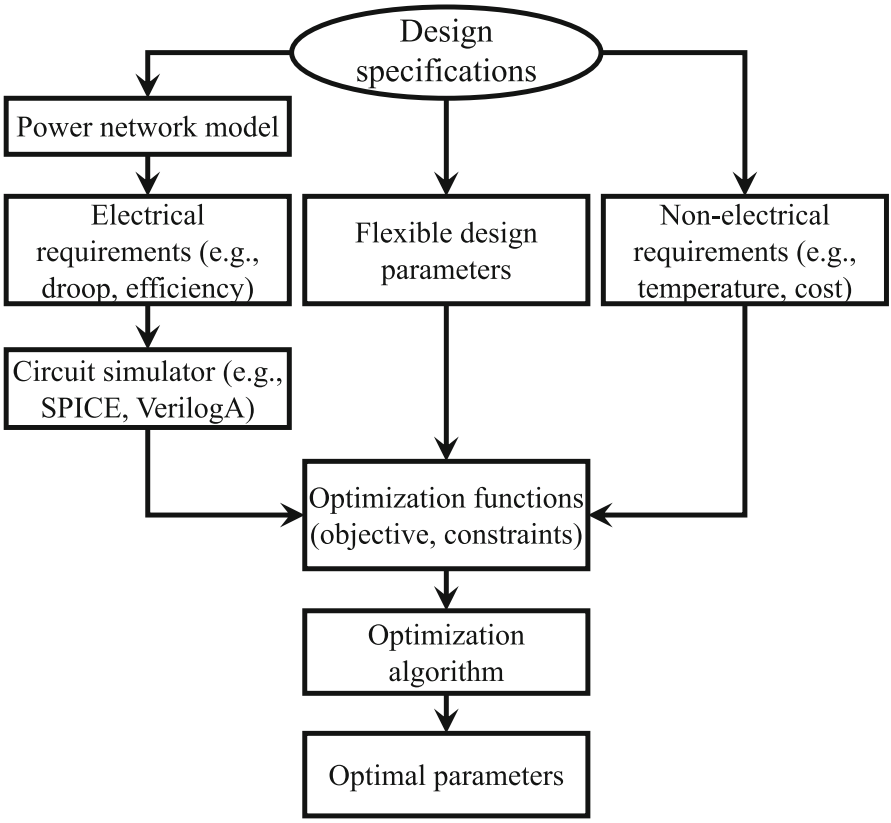

Fig. 9.2 Proposed power network optimization process

model shown in Fig. 9.3 [381], composed of cascaded lumped sections consisting of series RL segments, representing the interconnects and solder bumps, interleaved with parallel RLC segments, representing the decoupling capacitors, with an equivalent series resistance and inductance. More advanced topologies are necessary to evaluate the information from lower abstraction levels, such as the on-chip mesh [71, 72]. However, due to the lack of topology information during the early design phase, the development of a more accurate circuit model of a power network is a complex task.

The voltage source represents an idealized on-board regulator. For simplicity, a constant voltage supply is assumed. The main source of power consumption is modeled as a current source, representing the current delivered to the functional blocks, on-chip regulators, and leakage current. A current profile is necessary to evaluate the reliability of the network. Functional block information is used to model the profile of the load current [537]. Alternatively, the current profile may be modeled as a constant average current with a worst case current pulse [536].

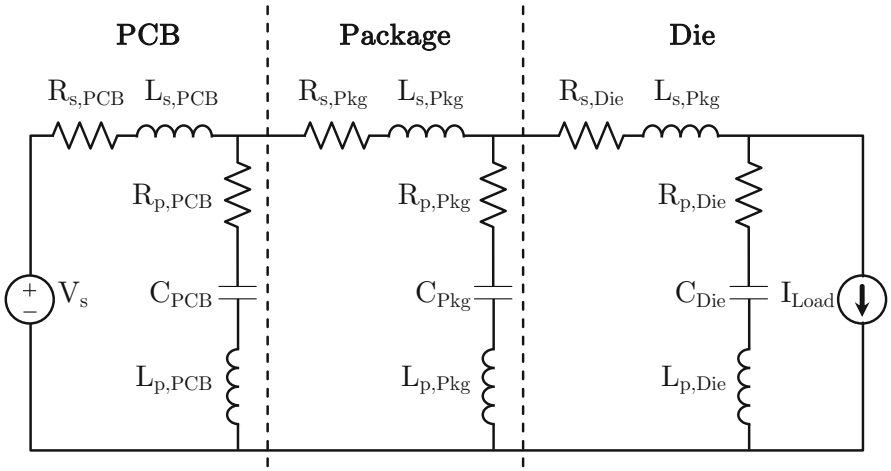

Fig. 9.3 Simplified model of power delivery network for optimization purposes

Once the power network model is determined, the design goals and technology limitations are converted into a functional form. For example, any limitations on voltage drop can be represented as

$$V_{drop} = \frac{\min(V_{Load}(t))}{V_s}, \tag{9.1}$$

where $V_{Load}(t)$ is the load voltage and V_s is the supply voltage. The power distribution efficiency, in turn, is

$$\eta = \frac{P_{Load}}{P_{in}}, \tag{9.2}$$

where P_{Load} and P_{in} are, respectively, the power dissipated by the current source and the total dissipated power. These specifications are necessary to convert the metrics of interest into the optimization functions.

9.1.2 Specification of non-electrical design requirements

In this chapter, non-electrical parameters are described as the system characteristics that are not directly inferred from the circuit model of the power network. These nonelectrical parameters include the on-chip temperature, manufacturing cost of the components, and area of the circuit elements. An externally supplied model is required to link the nonelectrical metrics and electric performance of the system. For example, if the mean time to failure (MTTF) is of concern, optimizing MTTF would place an upper limit on the current density and temperature, as shown in [538],

$$MTTF = \frac{K}{j^n}\exp\left(\frac{E_a}{kT}\right), \tag{9.3}$$

where K and n are material and process constants, E_a is the activation energy, k is the Boltzmann constant, T is the temperature, and j is the current density. Based on the analysis process, such as the individual currents, combined with external data, such as the wire dimensions, the current density in all of the elements is estimated to minimize this metric given the constraints.

9.1.3 Combination of electrical and nonelectrical metrics

The final form of the optimization function is

$$\mathbf{x}_{opt} = \min\left(f\left(\mathbf{x}\right)\right), \quad \text{subject to} \quad c(\mathbf{x}) \leq 0, \tag{9.4}$$

where $\mathbf{x}$ and $\mathbf{x}_{opt}$ are variables and correspond to the optimal parameter vectors, $f(\mathbf{x})$ is the function being optimized, and $c(\mathbf{x})$ is a set of constraint functions. The power delivery exploration process is formulated as in (9.4) to allow the application of constrained optimization algorithms.

The electrical analysis process needs to provide sufficient information to allow the nonelectrical metrics to be evaluated. The comprehensive optimization function requires an expression of the external metrics in terms of the variable parameters, electrical metrics, or both. For example, with adaption of [539], the MTTF of the interconnect segment can be approximated in terms of the interconnect dimensions and current,

$$MTTF = \frac{K_1 W^n H^n}{I_{rms}^n}\exp\left(\frac{K_2 W^2 H^2}{I_{rms}^2}\right), \tag{9.5}$$

where W and H are, respectively, the interconnect width and thickness, I_{rms} is the RMS current through the segment, and K_1, K_2, and n are process and material related constants. Electrical metrics, such as the RMS current through the segment, are evaluated from simulations of the power network. The variable parameters determine the characteristics of the power network model. For example, the dimensional parameters can be used to determine the impedance of the circuit elements. The formulated metrics are combined to create the objective function and set of constraints.

If multiple design objectives exist, a weighted sum of each objective is used to minimize each objective. The resulting formulation is shown in (9.6) to (9.9), where V_s is the supply voltage, W and H are, respectively, the top level interconnect width and thickness, w_1 and w_2 are weight parameters, $A_{int}(\mathbf{x})$ is the total area of the metal expended for the interconnect, and $V_{drop,max}$ and η_{min} are design constraints on, respectively, the voltage droop and efficiency. The objective function is the weighted

sum of the MTTF and cost, minimizing both metrics. To be satisfied, both $c_1(\mathbf{x})$ and $c_2(\mathbf{x})$ need to be greater than or equal to 0, ensuring that the droop is not larger than $V_{drop,max}$ and the efficiency is not less than η_{min}.

$$\mathbf{x} = [V_s, W, H], \tag{9.6}$$

$$f(\mathbf{x}) = \frac{w_1}{MTTF(\mathbf{x})} + w_2 A_{int}(\mathbf{x}), \tag{9.7}$$

$$c_1(\mathbf{x}) = V_{drop}(\mathbf{x}) - V_{drop,max}, \tag{9.8}$$

$$c_2(\mathbf{x}) = \eta_{min} - \eta(\mathbf{x}). \tag{9.9}$$

9.1.4 Circuit simulation procedure

During the optimization process, the circuit parameters are varied and the corresponding electrical parameters are evaluated. An efficient circuit simulator is the cornerstone of this procedure as the quality and timeliness depend upon the speed and accuracy of the simulator. Two simulation methods are utilized. The first method is commercial HSPICE [540] which requires a special interface with the programming language. The primary advantage of this approach is the versatility of the simulator. With the variety of available models, a wide range of circuits can be simulated and, therefore, optimized. The disadvantage of this approach is the communication overhead between the programming language and HSPICE which dramatically increases the simulation time.

Another approach is a custom Laplace transform-based simulator, requiring no interface with the programming language. The Laplace transform is widely used for simulation and optimization of linear circuits and systems [541, 542]. The primary advantage of this approach is the higher speed of the simulation due to the lack of communication with an external language and application-specific code optimization. A significant limitation is the narrow applicability of the method - only linear systems can be simulated using this approach due to the Laplace transform. A variety of methods exist, however, to extend the Laplace transform to nonlinear circuits. In [541], the switching transistors are replaced with lumped RC elements. A piecewise-linear model is another common approach for applying Laplace transforms to nonlinear systems. This method is particularly compatible with sequential switching [543, 544]. A modification of the Laplace Transform applicable to a certain class of nonlinear systems is introduced in [545]. Incorporating this method into the proposed framework may significantly extend the applicability of the proposed tool.

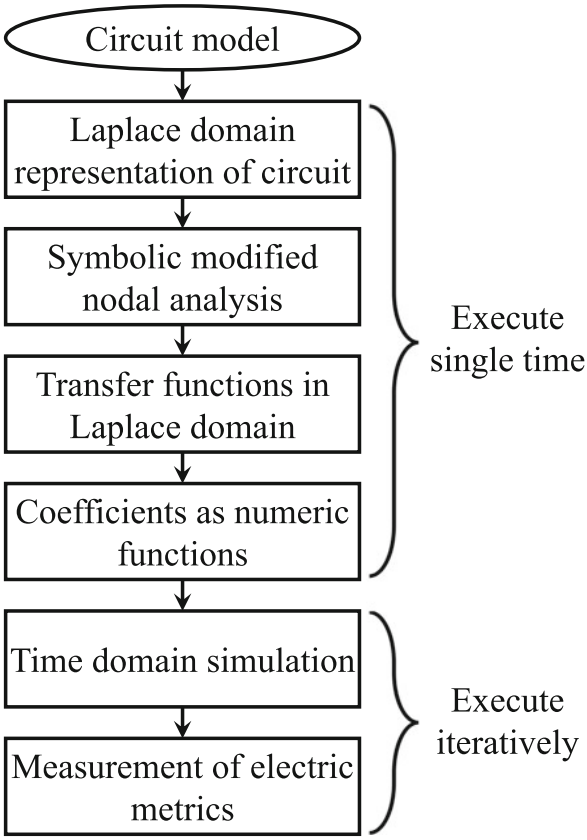

Fig. 9.4 Proposed Laplace transform-based optimization process

The proposed optimizer is applied to a model of a power network, which typically consists of passive RL-RLC branches [381]. The active devices, such as a voltage regulator or load transistors, are replaced with equivalent linear models to offset the error due to the assumption of linearity, which enables the use of a Laplace transform-based optimizer. In cases where the power network model is nonlinear (e.g., a power gated network), typically slower, numerical simulation tools can be utilized, such as HSPICE [540] or Verilog-AMS [546]. The choice between an active and passive power network model, therefore, becomes a tradeoff between accuracy and computational speed.

The Laplace transform-based process is shown in Fig. 9.4. The circuit elements are represented in the s domain. The fixed parameters are expressed numerically, while the variables are represented as symbolic variables. For instance, the impedance of a capacitor with a variable capacitance, fixed equivalent series resistance of 1 mΩ, and fixed equivalent series inductance of 10 pH can be presented as

$$Z_c = 1\text{m}\Omega + 10\text{pH} \times s + \frac{1}{Cs}, \tag{9.10}$$

where the capacitance C is shown as a symbolic variable, Z_c is the equivalent impedance of the capacitor, and s is the Laplace domain parameter.

After the circuit elements are expressed in the Laplace domain, a modified nodal analysis is applied. The circuit is modeled in terms of six input matrices, representing connections and parameter values, as shown in [64]

$$\begin{bmatrix} \mathbf{Y} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{V} \\ \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{J} \\ \mathbf{F} \end{bmatrix}, \tag{9.11}$$

where $\mathbf{V}$ and $\mathbf{I}$ are, respectively, the node voltages and currents through the voltage sources, $\mathbf{Y}$ is the matrix of nodal admittances, while $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{J}$, and $\mathbf{F}$ encode current and voltage sources, including controlled sources. The constructed matrix equation is solved for $[\mathbf{V}, \mathbf{I}]^T$.

The resulting vector represents the node voltages and source currents in terms of symbolic parameters in the Laplace domain. Dividing the resulting vectors by the source produces the transfer function, as shown in

$$H(s) = \frac{b_n s^n + \ldots + b_0}{a_m s^m + \ldots + a_0}. \tag{9.12}$$

The coefficients of the transfer function are expressed as a function of the variable parameters,

$$b_i = f_{i,num}(\mathbf{x}), \tag{9.13}$$

$$a_i = f_{i,den}(\mathbf{x}). \tag{9.14}$$

While the aforementioned procedure is computationally expensive, requiring a solution of the symbolic matrix system, the process only needs to be performed once for a particular circuit topology. Modifications of the variable parameters only change the value of the coefficients, $b_n \ldots b_0$ $a_n \ldots a_0$, while the symbolic representation remains intact. The speedup due to the proposed simulator is, therefore, largely dependent upon the number of iterations N during the optimization process. The speedup is estimated as

$$Speedup = \frac{t_n}{\frac{t_{setup}}{N} + t_{\mathcal{L}}}, \tag{9.15}$$

where t_n and $t_{\mathcal{L}}$ are the time per iteration using, respectively, numerical analysis and the Laplace transform-based simulator, and t_{setup} is the time required to determine the transfer function (9.12). Note that typically $t_{setup} > t_n > t_{\mathcal{L}}$, thus

the speedup converges to a positive value with large N, while approaching zero with small N. Since most optimization procedures require a large number of iterations to determine the global minimum, the creation of a symbolic transfer function represents a negligible fraction of the total computational time.

To simulate the transfer functions and extract the numeric data, the coefficients of the transfer functions of interest are calculated and converted into a state space model. A variety of efficient state space model simulation packages are available, such as LAPACK [547] and LTITR [548]. The input waveform and state space model are passed to the simulators to calculate the output waveform. This approach achieves significant speedup as compared to conventional, purely numerical algorithms. Applying a state-space model, the output waveform can be determined without solving the matrix equation during each time step. Conversion of a circuit into a matrix form is performed only once, greatly reducing the computational overhead. With the large number of circuit simulations during the optimization process, significant optimization speedup is achieved, as described in Section 9.2.

9.2 Case studies

Two practical case studies are presented in this section. Allocation of area for decoupling capacitors within a single rail system is analyzed in Subsection 9.2.1. The cost of decoupling capacitor placement is minimized while satisfying power consumption and the voltage droop constraints. The framework is then applied to a multi-rail system to determine the optimal number of voltage domains as described in Subsection 9.2.2.

9.2.1 Single rail system

A typical power network represented by serially cascaded RL branches and parallel RLC branches is shown in Fig. 9.5. A three-level power network including the PCB, package, and die levels is considered here. The series resistance and inductance of the power network are assumed fixed. The on-die parallel inductance is neglected assuming point-of-load on-die decoupling capacitors with small inductance [511]. The profile of the load current has been adapted from [536] and shown in Fig. 9.6(a). The load current profile models the fluctuations of the workload during system operation. The supply voltage is used as a design variable to explore the effects of supply voltage on system performance. Other controllable parameters are the number and magnitude of the decoupling capacitors within the PCB, package, and die levels. Minimization of the decoupling capacitor placement cost is the primary objective of this case study, subject to power consumption, power quality, and frequency requirements.

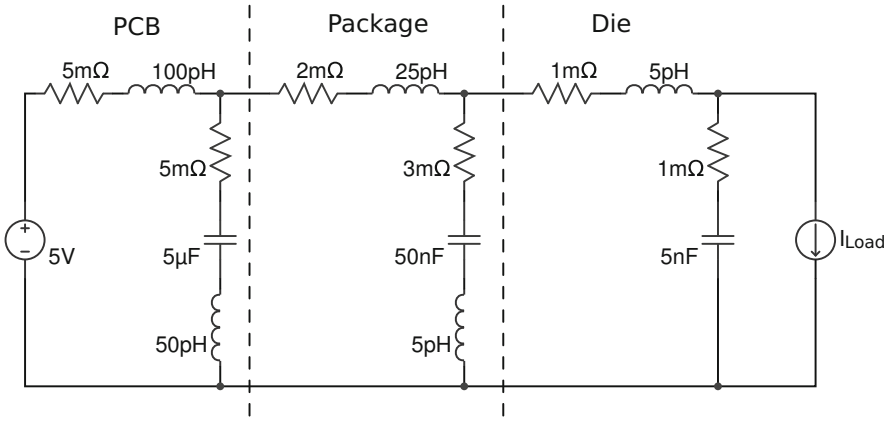

Fig. 9.5 Model of 1-D power delivery network with initial parameters

9.2.1.1 Optimization setup

The cost of each system level (PCB, package, die) is assumed to be a function of the physical area which is affected by the area of the decoupling capacitors. The decoupling capacitor placement cost Q_{die} is

$$Q_{die} = w_{die} A_{die}, \tag{9.16}$$

where A_{die} is the area of the on-chip decoupling capacitor and w_{die} is the cost of the unit on-die area. The total cost of the decoupling capacitors is therefore

$$Q = \frac{1}{\varepsilon_0} \sum_{i \in S} \frac{w_i C_i d_i}{\varepsilon_i}, \tag{9.17}$$

where S is the set of levels in the system (e.g., PCB, package, and die), ε_0 is the permittivity of free space, C_i is the parallel plate capacitance at level i, and d_i and ε_i are, respectively, the insulator thickness and relative permittivity at level i.

The oxide thickness and dielectric constant are described in [549–551]; however, the cost per area is not as clear. Based on the review of publicly available cost information [552–556], the cost per unit area of a package is approximately 3 to 6 times greater than the cost of unit PCB area, and approximately 3 to 10 times lower than the cost of unit die area. To simplify the cost estimate, the cost per unit area of a PCB is normalized to 1, the package area cost is assumed to be 4.5, and the cost per unit on-die area is assumed to be 20.25, 4.5 times greater than the cost per unit area of the package. The normalized cost estimates used in this case study are listed in Table 9.1.

Note the important tradeoffs that affect the optimization process [381]. A higher supply voltage enhances the speed but significantly increases the power

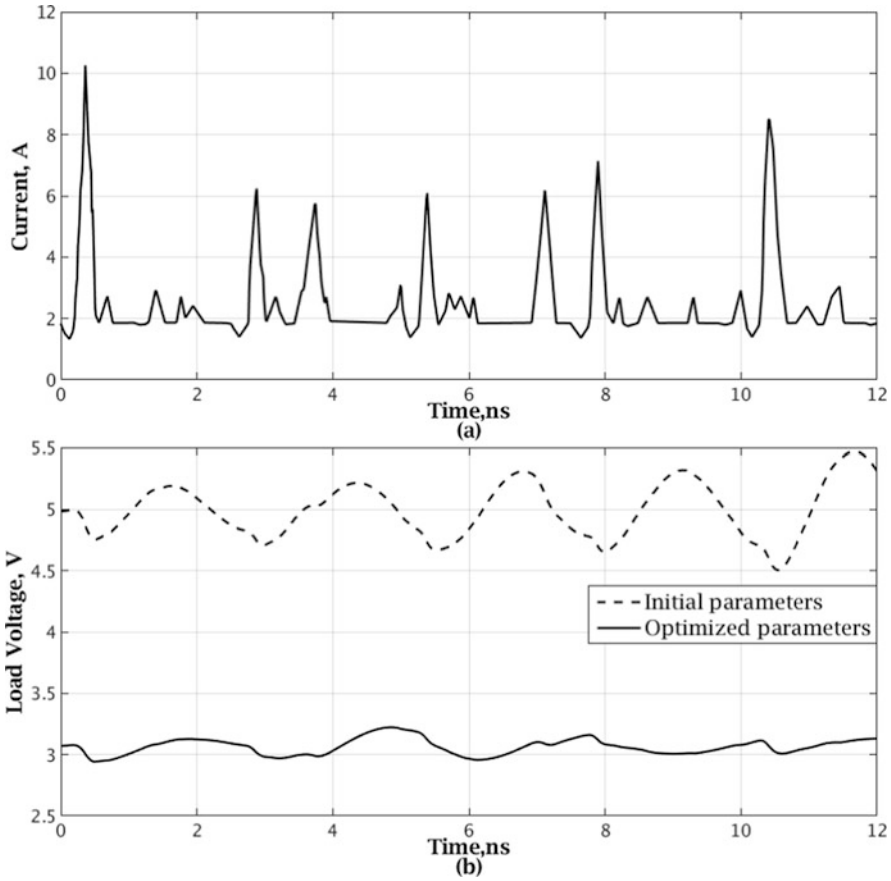

Fig. 9.6 Waveform of power network, a) load current adapted from [536], and b) load voltage with initial and optimized parameters

Table 9.1 Parameters of decoupling capacitor cost

Parameter	Die	Package	PCB
Cost per unit area, normalized	20.25	4.5	1
Insulator thickness	0.9 nm [549]	12 μm [550]	250 μm [551]
Insulator permittivity	3.9 [549]	4.6 [550]	4.5 [551]

consumption. Insertion of parallel decoupling capacitances is a powerful technique for reducing ripple currents since the high frequency components of the current bypass the load. Larger decoupling capacitors, however, require significant on-chip area, leading to greater system cost.

The target constraint metrics are power consumption, power quality, and speed. The power consumption is directly measured through simulation, and the corresponding constraint function is

$$c_1(\mathbf{x}) = P - P_{max}, \tag{9.18}$$

where $c_1(\mathbf{x})$ is the initial constraint function, P is the measured power, and P_{max} is the upper bound on the power consumption. Since the constraint function is negative, (9.18) ensures that the power dissipation does not exceed the maximum allowable power level.

For frequency, the constraint is

$$t_{p,CP} \leq T_{min}, \tag{9.19}$$

where $t_{p,CP}$ is the propagation delay of the critical path and T_{min} is the lower bound on the clock period. Evaluation of this metric, however, is computationally expensive and requires identification of the critical paths and extensive parameter extraction. This level of precision is typically not available during the early stages of the design process. In this case, accuracy is sacrificed for higher computational efficiency. The load voltage is, therefore, used as the speed metric,

$$c_2(\mathbf{x}) = V_{min} - \min(V_{load}(t)), \tag{9.20}$$

where $V_L(t)$ is the instantaneous voltage at the load, and V_{min} is the minimum voltage to maintain reliable high speed operation.

The third design constraint is power quality, described as voltage fluctuations, and is formulated as

$$c_3(\mathbf{x}) = \frac{\max(V_{load}(t)) - \min(V_{load}(t))}{V_{rail}} - \Delta V_{max}, \tag{9.21}$$

where V_{rail} is the supply voltage, and ΔV_{max} is the maximum allowed fluctuation. The optimization constraints are listed in columns two and three of Table 9.2.

9.2.1.2 Optimization results

The Interior Point Algorithm, part of MATLAB Optimization Toolbox [557] and HSPICE [540], is used in this case study. The optimization functions, circuit parameters, and external parameters are inputs to the optimization algorithm. The optimization procedure has been run on an Intel Core i7-6700 3.40 GHz 8-core computer using different initial conditions to avoid any local minima. The initial parameters that produce the lowest cost under specified constraints are listed in column four of Table 9.2.

Table 9.2 Optimization constraints, with initial and optimal parameters

	Lower bound	**Upper bound**	**Initial value**	**Optimized value**
Supply voltage	1.4 volts	10.0 volts	5.0 volts	3.09 volts
PCB decap	25.0 nF	10.0 μF	5.00 μF	2.71 μF
Package decap	50.0 pF	100 nF	50.0 nF	9.77 nF
Die decap	2.00 pF	10.0 nF	5.00 nF	9.32 nF
Minimum load voltage	1.40 volts	—	2.96 volts	2.94 volts
Power dissipation	—	10.0 watts	10.6 watts	6.51 watts
Load voltage	—	10.0%	19.3%	9.07%
Normalized cost	—	—	0.317	0.270

The optimization process is completed in 28 seconds, requiring 66 function evaluations to converge. The load voltage waveforms are shown in Fig. 9.6(b). The power network initially exhibits an underdamped response, resulting in relatively large droops and overshoots. After optimization, the voltage fluctuations are reduced in the optimized power network by choosing an appropriate decoupling capacitor. The reduction in the load voltage fluctuations allows the supply voltage to be scaled since fluctuations are less likely to drop below the minimum allowed level. Reducing the supply voltage, in turn, leads to lower power dissipation.

The optimization results are listed in column five of Table 9.2. As compared to the initial suboptimal parameters, the cost has decreased by almost 15% from 0.317 to 0.270. The initial parameters do not satisfy the power dissipation and load voltage constraints. A 38.6% reduction in power consumption is achieved, from 10.6 watts to 6.51 watts. Most of the reduction in power originates from the reduced supply voltage, from 5 volts to 3.09 volts. In addition, a 53% decrease in fluctuations is achieved, from 19.3% to 9.07%. As a result, the optimized parameters satisfy the target requirements, including the power and voltage constraints.

9.2.2 Multiple rail system

The problem of choosing the optimal number of rails is an important power delivery exploration issue. Utilizing several voltage domains may bring considerable savings in terms of power, while achieving performance goals [361]. At early stages of the design process, planning the circuit topology is problematic since the resulting power delivery characteristics are difficult to estimate in advance. In particular, it is unclear whether the power network is sufficiently conductive to satisfy voltage droop requirements. Separation of the low voltage circuitry from the rest of the IC is an attractive option to reduce power consumption due to the quadratic

Table 9.3 Voltage domain specifications of power delivery network adapted from [558]

Power network	**Rail #**	**Voltage**, V		**Current**, mA		**Peak slew rate**, A/μs	**Function**
		max	min	max	min		
A	A1-4	1.42	0.97	5,830	416	1,000	CPU core
	A5	1.20	0.99	3,150	225	500	GPU
	A6	1.33	1.00	10	1	500	USB
	A7	1.93	1.67	10	1	500	GPS
	A8	1.93	1.72	30	1	500	DSP
	A9	1.93	1.67	10	1	500	Camera
	A10	1.93	1.67	10	1	500	Audio
	A11	1.93	1.67	1,500	58	500	LTE+WiFi
	A12	1.55	1.00	3,150	225	500	Memory
B	B1-4	1.42	0.97	5,830	416	1,000	CPU core
	B5	1.20	1.00	3,160	226	*	GPU+USB
	B6	1.93	1.67	1,500	58	500	LTE+WiFi
	B7	1.93	1.72	60	4	*	GPS+DSP+ Camera+Audio
	B8	1.55	1.00	3,150	225	500	Memory
C	C1	1.42	1.00	26,470	1,889	*	CPU+Memory
	C2	1.20	1.00	3,160	226	*	GPU+USB
	C3	1.93	1.72	1,560	62	*	GPS+DSP+Camera +Audio+LTE+WiFi

relationship between power consumption and operating voltage. The scaled voltage is, however, less robust to sudden load current fluctuations, possibly violating droop requirements, allowing the device to malfunction. Moreover, utilizing separate power networks requires less metal resources for each rail, resulting in a power delivery network exhibiting higher impedance.

To investigate this problem, three power networks are considered, twelve rail (A), eight rail (B), and three rail (C) systems. The impedance characteristics of these networks are based on [558] and assume the power network topology shown in Fig. 9.3. The rail specifications are listed in Table 9.3. The maximum and minimum voltages represent the range of allowed values of the voltage. The model of the load current is a worst case triangular current waveform [532].

In system B, the rails with the closest voltage levels are merged to minimize energy losses due to the voltage conversion process. Rail A5 is merged with rail A6 to produce rail B5, and rails A7 through A10 are merged into rail B7, resulting in the eight rail system B. Further, rails B1 to B4 and B8 are merged, while rail B6 is merged with rail B7 to produce the three rail system C. The variables are the voltage supply of each rail, as well as the decoupling capacitance at each level of each rail.

For simplicity, the power rails are assumed to be mutually isolated, allowing each rail to be evaluated separately.

The objective of the design exploration process is to determine the set of rails that delivers the lowest possible cost of decoupling capacitance area. The objective function of the multiple rail system is adapted from (9.17),

$$Q = \frac{1}{\varepsilon_0} \sum_{j \in D} \sum_{i \in S_j} \frac{w_i C_i d_i}{\varepsilon_i}, \tag{9.22}$$

where D is the set of rails (voltage domains), and S_j is the set of layers of the power network (printed circuit board (PCB), package, or die) within the rail j.

Moving the decoupling capacitance farther from the load makes the system more vulnerable to inductive noise [493], limiting the cost benefits of a small on-chip capacitance. The greater fluctuations in the load voltage result in a need for a higher voltage supply to offset the potential voltage droops, resulting in higher power consumption. In addition, the inductive system response may result in significant overshoots [354] that may damage the transistors. For each rail in D, the aforementioned tradeoffs are expressed as constraint functions, as shown in (9.23) to (9.25),

$$c_1(V_s, C_{PCB}, C_{Pkg}, C_{Die}) = V_{load,min} - \min(V_{load}(t)), \tag{9.23}$$

$$c_2(V_s, C_{PCB}, C_{Pkg}, C_{Die}) = \max(V_{load}(t)) - V_{load,max}, \tag{9.24}$$

$$c_3(V_s, C_{PCB}, C_{Pkg}, C_{Die}) = P_{total} - P_{max}, \tag{9.25}$$

where $V_{load}(t)$ is the waveform of the load voltage, $V_{load,min}$ and $V_{load,max}$ are, respectively, the minimum and maximum bounds on the load voltage, and $Power_{total}$ and $Power_{max}$ are, respectively, the total power consumption and upper limit on the consumed power. The constraint functions place strict requirements on the quality of the power rails. If the voltage waveform violates the constraint functions, the objective function (or cost) is severely penalized, invalidating the result.

The power network model used in this case study does not include any nonlinear elements. A Laplace transform-based simulator has therefore been chosen. Particle swarm optimization is chosen as the optimization algorithm due to the robustness and efficiency characteristics of this algorithm. The optimization procedure is run on an eight core 3.40 GHz Intel Core i7-6700 machine. The results for 23 separate rail configurations are obtained in 26 minutes, with an average time of 67 seconds per rail. The results of the optimization are shown in Fig. 9.7. Note that the lowest value of the objective function is achieved with eight rails. In the eight rail and twelve rail scenarios, certain rails (e.g., rails seven to eleven in the twelve rail scenario) do

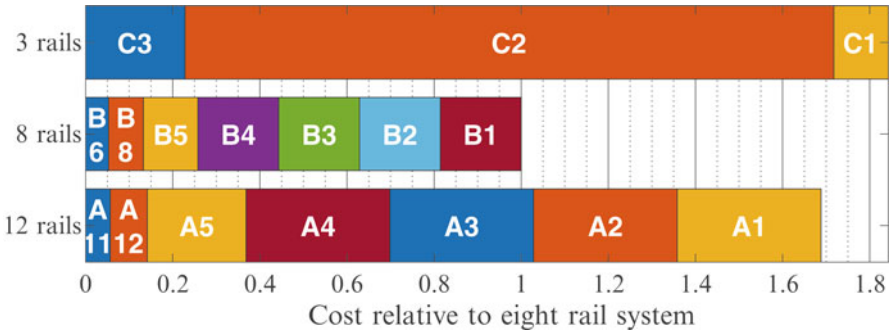

Fig. 9.7 Decoupling capacitor placement for three power delivery networks

not require decoupling capacitors due to the low load currents and high tolerance to variations.

To evaluate the benefits of the Laplace Transform optimization process, a similar optimization is performed using HSPICE [540]. The optimization results are identical to those results obtained from the Laplace transform optimization process due to the absence of nonlinear elements in the model. The total computational time, however, is 265 minutes, ten times greater than the Laplace simulator.

Distribution of the decoupling capacitor costs across the voltage domains normalized to the least expensive system is shown in Fig. 9.7. Certain patterns can be inferred. Comparing the eight and twelve rail systems, allocation of metal resources for separate power rails is unjustified from a cost perspective. The higher contribution of the CPU cores (A1 to A4) in the twelve rail network indicates that voltage fluctuations in this network are greater due to less metal resources allocated to each CPU rail, as compared to the eight rail system. The combination of rails A5 and A6 allocates more metal resources for both networks, resulting in reduced decoupling capacitor cost in combined rail B5.

As compared to the three rail system, where rails B1 to B4 and B8 (CPU cores and memory) are merged into a single voltage domain, the three rail system requires a large decoupling capacitance for the combined rail C2. The reason for the increased decoupling capacitance is the poor compatibility between voltage ranges. While rails B1 to B4 require a range of 0.97 to 1.42 volts, rail B8 has a range of 1.00 to 1.55 volts. The combined rail, therefore, needs to satisfy both ranges and is effectively shrunk to 1.00 to 1.42 volts, placing greater limitations on the voltage fluctuations. The narrow voltage range is compensated by placing a larger on-chip decoupling capacitance, increasing the overall cost of the power network.

A conventional power network design process may require a series of late design backtracking iterations to satisfy target noise performance requirements [559, 560]. Assuming that the post-floorplan power network model requires time t_{sim} for simulation and $t_{correct}$ for hotspot correction, and N iterations are required to reach the acceptable characteristics, the total time for the power integrity analysis process without early exploration is

$$t_{noEE} = (N-1)t_{sim} + Nt_{correct}, \tag{9.26}$$

where, typically, t_{sim} and $t_{correct}$ are on the order of hours and days, and N typically ranges between two and ten iterations. Alternatively, early power delivery exploration requires time t_{exp}, which may require several hours to complete. An expected result of the power delivery exploration process is a significant reduction in the number of iterations. Assuming the updated number of iterations is N_{new}, the total time for the power integrity analysis process is

$$t_{EE} = t_{exp} + (N_{new}-1)t_{sim} + N_{new}t_{correct}. \tag{9.27}$$

The savings in time due to the early power integrity analysis process is

$$t_{noEE} - t_{EE} = (N - N_{new})(t_{sim} + t_{correct}) - t_{exp}, \tag{9.28}$$

therefore, to ensure that the power delivery exploration is justified from the perspective of computational time, the following condition must be satisfied:

$$(N - N_{new})(t_{sim} + t_{correct}) > t_{exp}. \tag{9.29}$$

Note that typically $t_{sim} + t_{correct} > t_{exp}$, therefore, to justify early design exploration, it is sufficient to reduce the number of post-floorplan backtracking iterations, *i.e.*, $N_{new} < N$.

The proposed early power delivery exploration framework may reduce the number of costly iterations by providing an estimate of the optimal parameters at an earlier phase of the development process, shrinking both time and labor. The non-electrical parameters, such as area and cost, are combined with the electrical parameters to produce a system with minimum cost while satisfying target performance metrics. This approach provides useful information for early system exploration, allowing more effective design decisions to be made.

Several limitations of the proposed framework exist. First, the computational time largely depends upon the circuit simulator. Therefore, optimization of more complex circuits with a larger number of nodes may require significant computational time. A Laplace transform-based simulator is proposed for optimization of linear circuits. The speedup due to the Laplace transform-based simulator, however, largely depends upon the number of iterations during the optimization process. Second, a function for the metrics of interest needs to be determined to conduct the power delivery exploration process. Practical assumptions, therefore, need to be made to achieve useful results. An issue of premature convergence exists, resulting in the optimization converging to a local minimum rather than a global minimum [527]. It is, therefore, necessary to ensure that the design space is thoroughly explored, for example, by increasing population sizes (evolutionary algorithms), mutation and migration rates (genetic algorithm), swarm velocities and inertia (particle swarm), and the initial temperature and frequency of reheating (simulated annealing).

9.3 Conclusions

A versatile methodology for power delivery design exploration is described in this chapter. The primary strength of the framework is applicability to a wide range of objectives and constraints, including external, non-electrical parameters. The procedure supports the application of robust, general purpose algorithms to solve power delivery problems. A fast, optimization oriented Laplace transform-based simulator is described. Limitations of the proposed framework include the dependence on the computational time of the circuit simulator, the need for optimization functions during the preliminary design stages, and careful tuning of the optimization algorithms. The effectiveness of the framework is demonstrated by a case study, where the appropriate power delivery network is chosen among existing options.