Under review, Please Don't Distribute

Off-chip Decoupling Capacitor Allocation for Chip Package Co-Design

Hao Yu, Chunta Chu, and Lei He

ABSTRACT

Off-chip decoupling capacitor (decap) allocation is a demanding task during package and chip codesign. Existing approaches can not handle large numbers of I/O counts and large numbers of legal decap positions. In this paper, we propose a fast decoupling capacitor allocation method. By applying a spectral clustering, a small amount of principal I/Os can be found. Accordingly, the large power supply network is partitioned into several blocks each with only one principal I/O. This enables a localized macromodeling for each block by a triangular-structured reduction. In addition, to systemically consider a large legal position map in a manageable fashion, the map of legal positions is decomposed into multiple rings, which are further partitioned and parameterized in each block. The decaps are then allocated according to the sensitivity incrementally calculated, generated from the parameterized macromodel for each block. Compared to the previous approach, experiments show that our macromodel is 25X faster and has 3.04X smaller error. Moreover, our decap allocation reduces the optimization time by 97X, and reduces decap cost by up to 16% to meet the same power-integrty target.

1. INTRODUCTION

The demand of high-performance system on chip (SoC) or system in package (SiP) integration leads to chip-package interface (I/Os) operating in the Giga-bit range. Because the power supply planes in the package show strong electromagnetic resonance [1–3] under the injection of simultaneously switching I/O currents, they act as a significant source of noise in supply voltage and may create non-negligible jitters that limit the performance of I/Os. Therefore, it is necessary to obtain a clean power deliver system. Decoupling capacitors can be used to short power and ground planes at high frequencies to control power fluctuations. Different from the on-chip decap, off-chip decaps are discrete passive components with given capacitance, equivalent-series resistance (ESR) and equivalent-series inductance (ESL). ESL and ESR are among decisive factors for the cost (dollar-amount) of one decap. Considering congestion from signal and power routing, off-chip decaps can be inserted only at selected slots, called *legal positions* in this paper, and legal positions are used to connect terminals of decaps inside or outside the package. The off-chip decap optimization often minimizes the total decap cost subject to power integrity constraints and congestion from package routing.

There are two types of decap optimization flows. The first one assumes a pre-designed package with a limited number of legal positions and the decap optimization can be called *decap insertion* to decide what type of decap are used at which (usually not all) legal positions. The second flow assumes chip and package co-design and is increasingly more popular, because more designs become package limited but not chip limited as we move to more advanced technologies such as 65nm [4]. In this case, the IO cell number is often big due to high integration level, increased current demand, and needs to support multiple IO standards (for example in FPGA). The legal positions for decaps need to be decided or we can view this as a case with a large number of legal positions. Because of the much bigger solution space, this type of decap optimization called as *decap allocation* in this paper may lead to better designs compared to decap insertion for pre-design packages.

The following decap insertion algorithms have been developed recently. [5] calculates a multi-input multi-output (MIMO) impedance by model order reduction, and uses the inverse inductance [6] for a stable sparsification of massive magnetic coupling. It optimizes the cost of decaps and reduces the resonance impedance in the frequency-domain. The upper bound of impedance is implicitly determined by the rising-time of the worst-case I/O current profile. To explicitly consider the rising-time of the I/O current, [7] allocates decap to directly reduce the noise in the time-domain and avoids over-design compared to [5]. While [7] is faster than [5] due to an incremental impedance calculation, the simulated annealing (SA) based algorithms in [5, 7] is capable of dealing with a pre-designed package with only a limited legal positions as the algorithm virtually tries on each legal position. Therefore, they are not efficient for the chip-package co-design with a large number of legal positions.

In addition, the models used in [5, 7] have much room to improve. The MIMO reduction in [5] needs to match block moments. Its accuracy decays when the input port number increases. The legal positions are ports as well and they further increase the port number and size of impedance matrix. Moreover, the reduction in [5] ignores the structure information. The reduced model is dense and non-localized, and is inefficient to handle large-scale packages. On the other hand, [7] starts with a given macromodel and calculates the noise incrementally. However, it considers only the noise amplitude but not the noise pulse width. Because a very narrow noise with a large amplitude may not necessarily lead to noise violation, the noise model in [7] may lead to over design.

Considering chip-package co-design, this paper formulates a decap allocation problem to minimize the decap cost subject to noise violation constraint with consideration of noise pulse width. We develop a scalable algorithm using ring-based decap allocation followed by the legalization to complete detailed placement of decaps. The primary contributions of our paper are two folds. First, to generate a effective macromodel considering large numbers of input ports, we propose a spectral clustering to find a small amount of principal I/Os based on the I/O correlation. This enables an effective model order reduction. In addition, the information of clustered I/Os can be further used to partition the large RLC-network for power supply. By further incorporating the structure macromodeling [8, 9], we can perform a localized reduction and analysis for each partitioned block. Compared to the macromodel used in [5], our method is 3.04X more accurate and 25X more efficient.

Secondly, given a large number of legal positions, we introduce

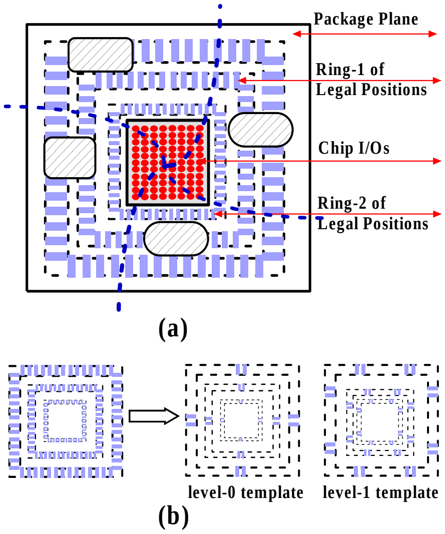


Figure 1: (a) The typical digram of package plane, chip I/Os and legal positions for decaps. The legal positions are represented by multiple rings. (b) Rings are decomposed into levelized templates, and are further partitioned and analyzed independently in each block.

a ring-based decap allocation to avoid trying every legal position as in SA. To systematically allocate decaps, the map of legal positions is first decomposed into multiple rings. By parametrically describing those rings in the state equations, the nominal responses and the sensitivities of I/Os with respect to the ring can be efficiently generated from a structured and parameterized macromodel for each partitioned block. Then, the decaps can be allocated according to the incrementally calculated sensitivity. Compared to the decap allocation in [5,7], experiments show that our allocation is 97X faster, and reduces the decap cost by up to 16%.

The rest of paper is organized as follows. In Section 2, we present the background and problem formulation. In Section 3, we introduce a parameterized description for P/G planes with allocated decaps. In Section 4, we propose a correlation based I/O clustering method. Using the I/O clustering information, in Section 5 we partition the parameterized RLC-plane into several blocks, and apply a triangular block-structured model reduction to locally generate the nominal response and sensitivity for each block. In Section 6, we introduce our decap allocation algorithm using the sensitivity, and present the experiment results. We conclude in Section 7.

2. PROBLEM FORMULATION

Packages often consist of multiple signal planes, power planes and ground planes with dielectric in between. Metal signal traces connecting the chip I/O cells to the PCB traces are routed between planes, and package planes are stapled together with vias, and connected to PCB by balls. We assume that the locations of chip I/O ports are known, and the allowed number of possible locations called legal positions for decaps are predefined for each region in a multi-layer package with consideration of congestion due to packaging routing and ball assignment. The legal positions are slots to connect the terminals of decaps, but not necessarily where decaps are located. As shown in Fig. 1, the I/Os are located in the center of the package. With the consideration of reserved routing area, the legal positions to allocate decaps are surrounded the chip by rings with different distances. After one decap is assigned to one legal location the decap is then called *legally placed.*

Note that a complete RLC model is required for accurate representations of interactions among package layers, C4 bumps, vias, on-chip power grids and all other signal traces. The power/ground plane can be uniformly discretized into N_v tiles, and each tile is modeled by RLC element under the PEEC model [10]. However, a detailed 3D extraction using the PEEC model introduces densely coupled inductances (L) that increase the model complexity. This can solved by stamping a sparsified L^{-1} element as discussed in Section 3.

In our decap allocation problem, the design freedoms are the legal locations and decap types. Brute-forcedly examining every possible combination is computationally expensive if not impossible. To allocate decaps in a manageable way, we propose a ring-based decomposition of all legal positions. This is based on the observation that the impact of decaps to I/O power-integrity can be distinguished by the distance to the center of the chip. As shown in Fig. 1 (a), the legal positions are decomposed into rings. Each ring is composed by a group of legal positions, and has different distance to the center of the chip. The illegal positions due to package routing are not included in each level of rings.

Moreover, because of non-uniformly distributed I/Os in space, the orientation of legal positions can have different impact to I/O power-integrity as well. As a result, the decaps needs to be non-uniformly distributed on one ring. To consider this, all rings are hierarchically divided into M_1 levelized positions, called *templates* in this paper. As shown by Fig. 1 (b), the level-0 template only allocates decaps on the center of rings, and the higher-level template allocates decaps more uniformly on the rings. To further consider M_2 types of decaps, each levelized template is duplicated by M_2 copies, each copy with a different decap type . note that only one copy at one level is selected to allocate decaps.

As a result, there are total $M = M_1 \cdot M_2$ templates, and a vector of templates can be defined by $\mathcal{T}_i$

$$\mathbf{T} = [\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_M], \qquad (1)$$

where $\mathcal{T}_i$ is one template with specified level and decap-type. Usually ,there are less than 5 levels of decomposition and less than 10 types [5, 7] to choose during the realistic design. Therefore, the number of M is still manageable.

Moreover, we need to define an accurate figure of metric to describe the power integrity at each I/O. The power integrity, i.e., the voltage bounce at each I/O is time and space variant during a sufficient long time-period t_p. One obvious metric, called *noise amplitude* could be defined by the maximum voltage bounce during t_p. However, a very narrow noise with a large amplitude may not lead to noise violation.

To avoid over-deign, a noise integral can be defined with the consideration of the noise pulse width. The noise integral above one targeted voltage Vc_i for i-th I/O is

$$f_i = \int_{t_0}^{t_p} max[y_i(\mathbf{T}, t), Vc_i]dt = \int_{t_s}^{t_e} [y(\mathbf{T}, t) - Vc_i]dt, \qquad (2)$$

with a pulse-width (t_s, t_e) and $y_i(\mathbf{T}, t)$ is transient noise waveform at i-th I/O. This applies to all p I/O cells, i.e.,

$$f_j \le Vd_j \qquad (j = 1, .., p). \qquad (3)$$

Recall that our design freedoms are two-folds: one is the level of ring, and another is the decap-type. Accordingly, our problem formulation is

FORMULATION 1. *Given the allowed noise ($\mathbf{Vc}$), legal positions (M_1) and decap types (M_2), the decap allocation problem is to decide which decap to be placed at which legal position and minimize the total cost of decap under a given bound of decap number ($\mathcal{M}$), such that the voltage violations* $\mathbf{f}$ *at each I/Os are smaller than the allowed noise.*

This problem can be mathematically represented by

$$min \sum_{i=1}^{M} n_i \mathcal{T}_i, \quad (i = 1, ..., M)$$

$$s.t. \ \mathbf{Uf} \le \mathbf{Vc} \ and \ \sum_{j}^{M_1} m_j \le \mathcal{M}. \qquad (4)$$

$x(y)$ ($\in R^{N\times 1}$)	State variable (at output)
$\mathbf{v}_n$ ($\in R^{N_v\times 1}$)	Nodal voltage variables
$\mathbf{a}_l$ ($\in R^{N_l\times 1}$)	Branch vector-potential variables
G ($\in R^{N_v\times N_v}$)	Nominal conductance matrix
C ($\in R^{N_v\times N_v}$)	Nominal capacitance matrix
L^{-1} ($\in R^{N_l\times N_l}$)	Nominal inverse-inductance matrix
$\mathbf{E}_l$ ($\in R^{N_v\times N_l}$)	Inductive incident matrices
$\mathcal{B}$ ($\in R^{N\times P}$)	Input/output port matrix

Table 1: Notations for system equation (5). Note that $N = N_v + N_l$.

where $\mathbf{f} = [f_1, ..., f_N]^T$, $\mathbf{U} = I_{N\times N}$, $\mathbf{Vc} = [Vc_1, ..., Vc_N]^T$. In addition, n_i is the dollar price for ith template ($i = 1, ..., M$), and m_j is the legal position number of jth level ($j = 1, ..., M_1$) As discussed in Section 6, this problem can be efficiently solved by an allocation according to sensitivity. The key is to calculate the parameterized sensitivity from a localized integrity analysis in Section 5.

3. PARAMETERIZED CIRCUIT EQUATION

Because the partial inductance in PEEC introduces massive magnetic couplings, it would slow down the noise analysis. As shown by [9], the inverse of L (L^{-1}) [6] described by VPEC model can be stably sparsified, and stably and passively stamped in the circuit matrix by a vector-potential nodal analysis (VNA). In this paper, the nominal RLC-network for package planes is modeled by VPEC model and is stamped by VNA in frequency (s) domain:

$$(\mathcal{G}_0 + s\mathcal{C}_0)x(s) = \mathcal{B}\mathbf{I}(s),\ y(s) = \mathcal{B}^T x(s) \tag{5}$$

with

$$x(s) = \begin{bmatrix}\mathbf{v}_n\\ \mathbf{a}_l\end{bmatrix}, \mathcal{B} = \begin{bmatrix}\mathbf{E}_i\\ 0\end{bmatrix},$$

and

$$\mathcal{G}_0 = \begin{bmatrix} G & \mathbf{E}_l L^{-1}\\ -\mathbf{E}_l^T L^{-1} & 0\end{bmatrix}, \mathcal{C}_0 = \begin{bmatrix} C & 0\\ 0 & L^{-1}\end{bmatrix} \tag{6}$$

All notations in (6) are summarized in Table 1. Note that $\mathcal{B}$ is the adjacent matrix to describe P identical inputs and outputs, where the inputs $\mathcal{J} = \mathcal{B}\mathbf{I}(s)$ are I/O current sources, and outputs $y(s)$ are the voltage bounce at those I/Os. As discussed in Section 4, studying such a I/O map can guide the network partition.

To obtain the sensitivity, we need to first parameterize the system. Each template $\mathcal{T}_i$ is described by a pair of topology matrices $\mathcal{T}_i^g$ and $\mathcal{T}_i^c$, where $\mathcal{T}_i^g$ describes how to connect the nodal equivalent conductance, and $\mathcal{T}_i^c$ defines how to connect the nodal capacitance and the branch equivalent susceptance (inverse of inductance). For a i-th template, adding decaps between tiles m and n results in:

$$\begin{aligned}&\mathcal{T}_i^g(k,l) = \mathcal{T}_i^g(l,k)\\ &= \begin{cases} -g_i & \text{if } k = m,\ l = n \text{ and } k \neq l\\ \sum_l |\mathcal{T}_i^1(k,l)| & \text{if } k = l\\ 0 & \text{else}\end{cases}\end{aligned} \tag{7}$$

where $k, l \in 1, 2, \cdots, N$, and g_i is the equivalent conductance of one decap. $\mathcal{T}_i^c(k,l)$ can be given similarly to add the equivalent capacitance and susceptance c_i and s_i. This decomposition enables us to apply an efficient decap allocation in Section 6.

Accordingly, the decaps can be parametrically added into the nominal state matrix

$$\begin{aligned}[\mathcal{G}_0 + s\mathcal{C}_0 + \sum_{i=1}^{M}(\mathcal{T}_i^g + s\mathcal{T}_i^c)]x(\mathbf{T}, s) = \mathcal{B}\mathbf{I}(s),&\\ y(\mathbf{T}, s) = \mathcal{B}^T x(\mathbf{T}, s).&\end{aligned} \tag{8}$$

However, the $x(\mathbf{T}_M, s)$ is the total voltage response. For the purpose of design optimization, similar to handle variations in [11], the state variable $x(\mathbf{T}, s)$ is first expanded into Taylor series with respect to $\mathcal{T}_i$, and reconstruct a new state variable x_{ap} using the nominal values and the first-order sensitivities

$$x_{ap} = [x_0^{(0)}, x_1^{(1)}, ..., x_M^{(1)}]^T. \tag{9}$$

A dimension-augmented system can be reorganized according to the expansion order

$$(\mathcal{G}_{ap} + s\mathcal{C}_{ap})x_{ap} = \mathcal{B}_{ap}\mathbf{I}(s), \quad y_{ap} = \mathcal{B}_{ap}^T x_{ap}, \tag{10}$$

where

$$\mathcal{G}_{ap} = \begin{bmatrix} \mathcal{G}_0 & 0 & \dots & 0\\ \mathcal{T}_1^g & \mathcal{G}_0 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ \mathcal{T}_M^g & 0 & \dots & \mathcal{G}_0\end{bmatrix}, \quad \mathcal{C}_{ap} = \begin{bmatrix} \mathcal{C}_0 & 0 & \dots & 0\\ \mathcal{T}_1^c & \mathcal{C}_0 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ \mathcal{T}_M^c & 0 & \dots & \mathcal{C}_0\end{bmatrix}, \tag{11}$$

both have a lower triangular block structure. Although the system size is enlarged by parametrically adding decaps in this fashion, the ports of the augmented system are still the input ports of I/O currents. The size of augmented system can be still reduced by the model order reduction. In contrast, the impedance based approach [5, 7] needs to increase the port number dramatically to accommodate those decaps.

4. I/O CURRENT CORRELATION AND SPECTRAL CLUSTERING

Due to the large number of input ports, the macromodel by model reduction applied by [5] is still ineffective. Because the input current vectors show redundancy the time/space-variant input I/O currents are not mutually independent. If the various inputs are correlated, then they can be represented by a function of a smaller number of independent variables based on the principal component analysis (PCA) using eigen-decomposition (ED).

```
SPECTRAL CLUSTERING ALGORITHM
1 Input: Cluster number $K$, correlation matrix $\mathcal{C} \in R^{N\times N}$, and I/O port matrix $\mathcal{B} \in R^{N\times p}$
2 Compute normalized Laplacian: $\mathcal{L} = \mathcal{D}^{-1/2}\mathcal{C}\mathcal{D}^{1/2}$, where $\mathcal{D} = diag(\mathcal{C})$;
3 Compute the first $K$ eigenvectors $v_1, ..., v_K$ of $L$;
4 Let $V = [v_1, ..., v_K] \in R^{N\times K}$;
5 Let $y_i \in R^K$ $(i = 1, ..., N)$ be the vector of $i$-th row of $V$;
6 Cluster $y_i$ $(i = 1, ..., N)$ by K-means into $C_1, ..., C_K$;
7 Transform $\mathcal{B} \in R^{N\times p}$ by PCA: $\mathcal{B}_x = V\mathcal{B} \in R^{N\times K}$;
8 Output: Clusters $\mathcal{A}_1, ..., \mathcal{A}_K$ with with $\mathcal{A}_i = \{j|y_j \in C_i\}$, and a new I/O port matrix $\mathcal{B}_x$
```

Figure 2: Algorithm 1 for spectral analysis of input current sources with PCA and K-means.

This becomes the motivation to apply the singular value decomposition (SVD) [12–14] based terminal reduction as SVD is equivalent to eigen-decomposition when the matrix to be decomposed is symmetric positive definite. These approaches [12–14] assume that the correlation or similarity of inputs can be inferred from a low rank analysis of system transfer function by SVD, and then compress the system transfer function. Therefore, terminal reduction in fact, studies the similarity of the system since it is based on the singular value (pole) analysis of the system transfer function. However, the real correlation of inputs is dependent on the input signals. As a result, finding the representative ports or ignoring some 'insignificant' ports based on the system similarity may lead to simulation errors, because there could be one significant output response caused by one significant signal that is applied at one port ignored from the system pole analysis. In this paper, we propose to directly study the similarity or correlation of I/O currents. As a result, the large number of I/Os are clustered into K groups, each with one principal I/O current as input.

Given a typical set of P I/O input vectors applied in a sufficient long period, the sampled transient-current $I(t_k, n_i)$ ($k = 1, ..., T$, $i = 1, ..., P$) at time-instant t_k for each I/O n_i can be be described by a random process as follows

$$\mathcal{S}_{n_1} = \{I(t_1, n_1), ..., I(t_T, n_1)\}, \quad \mathcal{S}_{n_2} = \{I(t_1, n_2), ..., I(t_T, n_2)\}$$
$$... \quad \mathcal{S}_{n_P} = \{I(t_1, n_P), ..., I(t_T, n_P)\}.$$

A current spatial-correlation matrix is defined by

$$\mathcal{C}(i, j) = \frac{cov(i, j)}{\sigma_i \cdot \sigma_j}, \tag{12}$$

where $cov(i, j)$ is co-variance between nodes n_i and n_j, and σ_i, σ_j are standard-variations of nodes n_i and n_j. Those correlation coefficients $\mathcal{C}(i, j)$ can be be precomputed and stored in a table.

After extracting the correlation for input currents, we can build a correlation graph by assigning the weight of edge between I/Os n_i and n_j by the correlation value $\mathcal{C}(i, j)$. A fast clustering based on spectral analysis [15] can be applied to efficiently handle a large-scale correlation graph to find K clusters $\mathcal{A}_1, ..., \mathcal{A}_K$ using K-means method, where the I/Os in one cluster all show a similar current waveform. In addition, the number of I/O current sources can be reduced by PCA

$$\mathcal{J}_x = V\mathcal{J} = V\mathcal{B}\mathbf{I}(s) \quad \in R^{1\times K}. \tag{13}$$

It is equivalent to reduce the port matrix

$$\mathcal{B}_x = V\mathcal{B} \qquad \in R^{N\times K}. \tag{14}$$

As such, there is only one principal port selected to represent each cluster.

The overall clustering is outlined in Algorithm 1. Usually, 1000 sources can be approximated by around 10 sources if the inputs are strongly correlated. In addition, note that with the use of spectral analysis, the result by PCA or K-means is equivalent [15]. Therefore, there is only one principal port for each cluster.

5. LOCALIZED INTEGRITY ANALYSIS

5.1 Network Decomposition

Because the I/O currents are distributed non-uniformly in space, it has different impact to voltage bounces along different orientations. Therefore, it is possible that the one level of ring can be non-uniformly allocated with different typed decaps. To this end, it better to decompose the I/O cells, the RLC-network for power supply, and the M templates into K blocks (See Fig. 1). A corresponding localized analysis can be then preformed to decide how many decaps for one block of I/Os.

The decomposition needs to partition the network based on physical properties such as couplings and latency. The TBS method in [8] leverages the property of latency, which is more suitable for timing simulator. But for the verification of power integrity, it is more meaningful to study the partition based on I/O inputs. Moreover, the partition in TBS [8] is to tear nodal voltage variables v_n for conductance and capacitance matrices, which is not suitable for inductance/susceptance partition because inductance/susceptance is described by the branch current/vector-potential. This can be solved as follows.

The flat VNA network $(\mathcal{G}_0, \mathcal{C}_0, \mathcal{B}_x)$ in (5)is first mapped into a circuit graph, where three different weights (2,1,0) are assigned for the resistor, capacitor and self-susceptor (branch L^{-1}). A fast multi-level *min-cut* partition *hmetis* in [16] is applied to tear those interconnection branches with specified ports $A_1, ..., A_K$ obtained from the spectral clustering. As a result, the network is decomposed into two-levels with the torn resistors, capacitors and self-susceptors in an interconnection block, and all remaining blocks are connected with the interconnection block by incident matrices as shown below

$$\begin{aligned}
\mathcal{G}_{ap} \quad \rightarrow \quad \mathbf{G}_{ap} &= \begin{bmatrix} \mathbf{G}_1 & \cdots & 0 & X_{1,0} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \mathbf{G}_K & X_{K,0} \\ -X_{1,0}^T & \cdots & -X_{K,0}^T & Z_r \end{bmatrix} \\
\mathcal{C}_{ap} \quad \rightarrow \quad \mathbf{C}_{ap} &= \begin{bmatrix} \mathbf{C}_1 & \cdots & 0 & X_{1,0} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \mathbf{C}_K & X_{K,0} \\ -X_{1,0}^T & \cdots & -X_{K,0}^T & Z_i \end{bmatrix} \\
\mathcal{B}_x \quad \rightarrow \quad \mathbf{B} &= \begin{bmatrix} \mathbf{B}_1 & & & \\ \vdots & \ddots & & \\ & & \mathbf{B}_K & \\ & & & \mathbf{0} \end{bmatrix}
\end{aligned} \tag{15}$$

with

$$\mathbf{G}_i = \begin{bmatrix} \mathcal{G}_i & 0 & \cdots & 0 \\ \mathcal{T}_{1,i}^g & \mathcal{G}_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{T}_{M,i}^g & 0 & \cdots & \mathcal{G}_i \end{bmatrix}, \ \mathbf{C}_i = \begin{bmatrix} \mathcal{C}_i & 0 & \cdots & 0 \\ \mathcal{T}_{1,i}^c & \mathcal{C}_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{T}_{M,i}^c & 0 & \cdots & \mathcal{C}_i \end{bmatrix} \tag{16}$$

where $\mathcal{G}_0$ and $\mathcal{C}_0$ are partitioned into K blocks $\mathcal{G}_j$ and $\mathcal{C}_j$ ($j = 1, ..., K$). Accordingly, those parameterized templates $\mathcal{T}_i$ are also partitioned into $\mathcal{T}_{ij}$ ($i = 1, ..., M$ $j = 1, ..., K$). Note that a block matrix structure is implemented to avoid building the large sized matrix.

Because the couplings are relocated into one interconnection block $Z_{r,i}$, each partitioned block in diagonal can be analyzed or reduced individually but with the same accuracy. However, the system poles are not determined only by those blocks in diagonal. To achieve a high-order accuracy but with only a low-order reduction, the TBS reduction in [8] is extended to consider inductance and is presented below.

5.2 Triangular Block-Structured Reduction

After tearing the VNA network into a two-level form, we further transform it into a localized triangular block form with the use of replication [8]. Basically, as shown by (17), a replica block of $\mathbf{G}_{ap}$ is first stacked diagonally to construct a size-doubled $\mathbf{G}_{tb}$, and then those lower triangular blocks are moved to the upper triangular parts of $\mathbf{G}_{tb}$. The resulting triangularized system is

$$\mathbf{G}_{tb} = \left[\begin{array}{cccc|cccc} \mathbf{G}_1^x & \ldots & 0 & X_{1,0} & \mathbf{G}_1^y & \ldots & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & \mathbf{G}_K^x & X_{K,0} & 0 & 0 & \mathbf{G}_K^y & 0 \\ 0 & \ldots & 0 & Z_r & -X_{1,0}^T & 0 & -X_{K,0}^T & 0 \\ \hline & \mathbf{0} & & & & & \mathbf{G}_{ap} & \end{array}\right], \tag{17}$$

where

$$\mathbf{G}_i^x = diag[\underbrace{\mathcal{G}_i, ..., \mathcal{G}_i}_{M}], \ \mathbf{G}_i^y = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \mathcal{T}_{1,i}^g & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{T}_{M,i}^g & 0 & \cdots & 0 \end{bmatrix}. \tag{18}$$

$\mathbf{C}_{tb}$ can be transformed in a similar fashion. The triangularized system has a localized pole distribution, where poles are determined only by those blocks in the diagonal. In addition, the factorization cost only comes from those block in diagonal. However, due to the replica block the overall factorization cost of the triangulated system is still the same as the original. To reduce the overall computational cost, we further apply a block-structured projection to reduce the system size.

As the network is decomposed and further triangularized, each block $(\mathbf{G}_i, \mathbf{C}_i, \mathbf{B}_i)$ can be reduced independently [8,9] by finding a q-th projection matrix Q_i ($R^{n_{b_i}\times q}$) ($1 \leq i \leq K$) to contain the moment space of the diagonal block

$$\{\mathbf{R}_i, \mathbf{A}_i\mathbf{R}_i, ..., \mathbf{A}_i^{q-1}\mathbf{R}_i\},$$

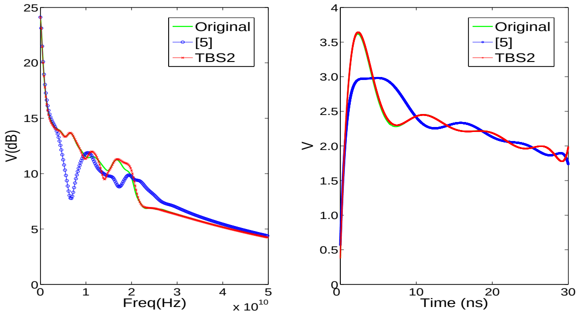


Figure 3: Waveform accuracy comparisons between the original, the method in [5] and TBS2 in (a) time-domain and (b) frequency-domain for 4th principal port. The original and TBS2 are visually identical.

where $\mathbf{A}_i = \mathbf{G}_i^{-1}\mathbf{C}_i$ and $\mathbf{R}_i = \mathbf{G}_i^{-1}\mathbf{B}_i$, and $(n_b)_i$ is the size of original block. Accordingly, a block-diagonal projection matrix

$$\mathcal{Q} = diag[\underbrace{\mathcal{Q}_1, ...\mathcal{Q}_1}_{M}, ..., \underbrace{\mathcal{Q}_K, ...\mathcal{Q}_K}_{M}, \mathcal{Q}_0, \mathcal{Q}_{ap}] \qquad (19)$$

is constructed to reduce the original matrix $\mathbf{G}_{tb}$, $\mathbf{C}_{tb}$ and $\mathbf{B}_{tb}$, respectively.

$$\widetilde{\mathbf{G}}_{tb} = \mathcal{Q}^T\mathbf{G}_{tb}\mathcal{Q},\ \widetilde{\mathbf{C}}_{tb} = \mathcal{Q}^T C_{tb}\mathcal{Q},\ \widetilde{\mathbf{B}}_{tb} = \mathcal{Q}^T B_{tb}. \qquad (20)$$

In addition, note that $\mathcal{Q}_0$ is an identity matrix to project those interconnection branches, and $\mathcal{Q}_{ap}$ is either obtained by directly applying a lower-order PRIMA to $(\mathcal{G}_{ap}, \mathcal{C}_{ap}, \mathcal{B}_x)$, or it can be accurately approximated by $[\mathcal{Q}_1, \mathcal{Q}_2, ..., \mathcal{Q}_K, \mathcal{Q}_0]^T$ [8].

Moreover, one important observation is that, since only one principal port at each block is selected, a SIMO-reduction can be easily applied to achieve q-th order moment matching for each block, and the reduced macromodel for each block can be repeatedly used for any input signals.

As a result, a localized integrity analysis can be efficiently performed for each block to generate both nominal responses and sensitivities in time-domain

$$(\widetilde{\mathbf{G}}_{tb} + \frac{1}{h}\widetilde{\mathbf{C}}_{tb})\widetilde{x}_{tb}(t) = \frac{1}{h}\widetilde{\mathbf{C}}_{tb}\widetilde{x}_{tb}(t-h) + \widetilde{\mathbf{B}}_{tb}\mathbf{I}(t)$$
$$\widetilde{y}_{tb}(t) = \widetilde{\mathbf{B}}_{tb}^T\widetilde{x}_{tb}(t). \qquad (21)$$

The k-th block power integrity at one principle I/O perturbed by i-th template is

$$\widetilde{y}_{tb}(t) = \widetilde{y}_{tb}^{(0)}(t) + \widetilde{y}_{tb}^{(1)}(t). \qquad (22)$$

Note that although it is a localized solution, the couplings between different blocks are still taken into account due to the two-level network decomposition and the triangularization. Below, we present the decap allocation algorithm using the block integrity including nominal responses and sensitivities.

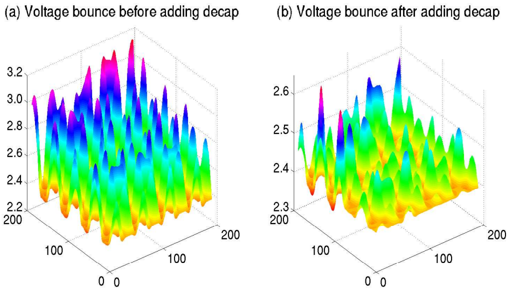
6. ALGORITHM AND EXPERIMENTAL RESULTS

6.1 Sensitivity based Optimization

The problem in Section 2 can be efficiently solved by the sensitivity based optimization. The key is to calculate sensitivities from the structured and parametrized macromodel in Section 5. Then, the decap is allocated for each block according to the sensitivity of I/O power integrity with respect to templates. The partitioned template $\mathcal{T}_{i,j}$ is recursively added according to the order of the gain. As a result, a minimum number of decaps are added to reduce the voltage violations in problem formulation (4). Such a greedy flow is able to solve large-scale designs efficiently and effectively.

```
MULTIPLE RING-BASED ALLOCATION ALGORITHM
1 Input: Integrity vector $\mathbf{Vc}$
2 Compute initial $y^{(0)}$ and $y^{(1)}$ using (21);
3 Reorder $\mathcal{T}_k = \{\mathcal{T}_{i_1,k}, \mathcal{T}_{i_2,k}, ..., \mathcal{T}_{i_M,k}\}$ $(k = 1, ..., K)$;
4 Do allocation with max $\mathcal{T}_k$ for block $k$
5     Delete max $\mathcal{T}_k$ from $\mathcal{T}_k$ and $M = M - 1$;
6     Compute $y_k = y_k^{(0)} + y_k^{(1)}$;
7 Until $y_k$ satisfies the block integrities $\mathbf{Vc}_k$
8 Output: Allocated template-vector $\mathbf{T}$ for detailed decap placement
```

Figure 4: Algorithm 2 for sensitivity based decap allocation.


(a) Voltage bounce before adding decap

(b) Voltage bounce after adding decap

Figure 5: Voltage bounce at P/G plane (a) before decap allocation and (b) after decap allocation.

The overall optimization is outlined in Algorithm 2. The nominal value and sensitivity are computed one-time from the structured and parameterized macromodel from (21). Afterwards, the decap is added into each block independently. In k-th block, the template-vector $\mathbf{T}$ is ordered according to the magnitude of sensitivities

$$\{\delta y_{i_1,k}, \delta y_{i_2,k}, ..., \delta y_{i_M,k}\}$$

and is added according to this order until the integrity constraint of k-th block is satisfied. The algorithm then iterates to the next block until all the power integrities of all blocks are satisfied. Because each input-template is legalized initially to exclude those illegal positions, the output template vector $\mathbf{T}$ can be directly used for the detailed placement of decaps.

6.2 Results

The proposed macromodeling and allocation algorithm has been implemented in C and Matlab. We call our macromodeling method as TBS2, and our optimization as multi-ring based allocation (MRA). Experiments are run on a Linux workstation with 2G RAM. A typical FPGA package model is assumed with the a specific application inputs . Four packages P/G planes are assumed with same the size of $1cm \times 1cm$. The Vdd is assumed to be 2.5V, and the targeted noise is 10% of Vdd, i.e., 0.25V. The worst-case I/O current sources are modeled as triangle-waveform with rising time 0.1ns, width 1ns and period 150ns, which are randomly distributed in a square of $0.2cm \times 0.2cm$ located in the center of a $1cm \times 1cm$ package plane. The 30% of remaining area are reserved for legal postions. The 4 decap types in [7] are used and summarized in Table 2. The total number of decaps is bounded by 80, and the total number of rings is 5, each ring decomposed

Type	1	2	3	4
ESC(nF)	50	100	50	100
ESR(Ohm)	0.06	0.06	0.03	0.03
ESL(pH)	100	100	40	40
Normalized Price	1	2	2	4

Table 2: Settings of decaps.

Table 3: Results of decap allocations by SA and our MRA method. The cost of decaps is normalized.

ckt	#level	#legal-pos	#partition	SA-NA		MRA-NA		MRA-NI	
(#node+#I/O)				opt	norm-cost	opt	norm-cost	opt	norm-cost
280+40	0,1	20	4	192.2s	16	5.2s	10	5.4s	10
1160+160	0,1	80	4	2hrs	55	62.3s	50	64.2s	40
4720+640	0,1	320	4	7hrs	102	277.1s	96	280.2s	80
10680+1440	0,1,2	720	8	1day	233	783.7s	216	773.5s	200
19521+3645	0,1,2	1701	8	NA	NA	932.4s	277	972.2s	265
55216+10880	0,1,2,3	5440	16	NA	NA	51mins	340	54mins	312

into four levels (0-3). We increase the circuit complexity by increasing the number of discretized tiles, and need more levels for legal positions when the tile number becomes larger. We allocate decaps by MRA and SA methods to satisfy the power integrity at I/Os under constraints of either the noise amplitude (NA) or the noise integral (NI).

6.2.1 Comparison of Macromodels

We first compare our method with macromodeling in [5] as follows. The packages planes are discretized into 4720 tiles, described by a RLC-mesh with 12,810 resistors, 11,800 capacitors and 64,000 susceptors. There are 420 I/O current sources as inputs. As discussed in Section 4, the sequences of I/O currents are generated by simulating the specified application of input vectors for millions of cycles. One spatial correlation matrix $\mathcal{C}$ is extracted from the sequences. Then, the spectral clustering finds 8 principal ports by PCA and clusters the ports into 8 groups. Accordingly, the network is partitioned into 8 blocks by *hmetis*. Fig. 3 compares the frequency and time domain responses at 4th principal port. Due to the I/O port reduction and a localized reduction and analysis, our method is 21X faster (765s vs. 35.2s) to build and 25X faster (51mins vs.2mins) to simulate compared to [5]. Moreover, because the TBS reduction can achieve a higher accuracy with use of triangularization, the waveform by TBS2 is visually identical to the original. But the reduced waveform by [5] has about 3.04X larger waveform error in the time-domain.

6.2.2 Comparison between SA and MRA

We also compare the runtime and the cost of allocated decaps between SA and MRA. During this comparison, both methods use the noise amplitude as the constraint. As shown in Table 3, due to the systematical allocation with use of sensitivity, MRA reduces the allocation time by 97X on average compared to SA. In addition, SA can only handle circuits up to $\sim 10,000$ nodes. To obtain a result in a reasonable time, SA usually can not find the optimal solution. For a circuit with 10,680 nodes, MRA finds a solution with dollar cost about 216 in 13mins, but SA finds a solution with dollar cost about 233 (+9%) in 1day.

In addition, Fig. 5 shows the voltage-bounce map (at 80ns) across the top plane. The initial noise amplitude is around 1.0V, and its voltage bounce profile is shown in Fig. 5 (a). In contrast, the decap-allocation by MRA results in a smaller voltage bounce that closely approaches the targeted bounce (0.25V) as shown in Fig. 5 (b).

6.2.3 Comparison between NA and NI

We further compare the runtime and the cost of allocated decaps by noise amplitude (NA) and noise integral (NI), both using MRA for allocation. As shown in Table 3, compared to the optimization with NA, the optimization with NI reduces the cost of allocated decaps by up to 7% within a similar allocation time. This is because the constraint by the noise amplitude ignores the accumulated effect of the transient noise waveform. In contrast, the constraint by noise integral can consider the noise pulse width, and can accurately predict the decap allocation using the transient noise waveform. As a result, NI reduces the dollar cost by up to 16% compared to the SA using NA [7].

7. CONCLUSIONS

To efficiently and accurately allocate the decap, this paper has presented a fast off-chip decoupling capacitor allocation considering I/O Clustering. We have presented a spectral analysis to cluster larger numbers of I/Os and find the principal I/Os with the use of I/O correlation. This clustering enables I/O-based network partition, and also provides an efficient structured macromodel generation by moment matching. In addition, to systemically allocate decaps in a manageable fashion, we have also proposed a ring-based decap allocation based on the sensitivity, which is generated from a localized integrity analysis using a structured and parameterized macromodel. Experiments with four layers of power/ground planes show that compared to the existing SA based allocation, our method is up to 97X faster, and also reduces decap cost by up to 16% to meet the same noise bound in time-domain.

8. REFERENCES

[1] H. Chen and J. Neely, “Interconnect and circuit modeling techniques for full-chip power supply noise analysis,” *IEEE Trans. Compon., Packag. and Manuf. Tech.*, pp. 209–215, 1998.

[2] M. Swaminathan, J. Kim, I. Novak, and J. Libous, “Power distribution networks for System-on-Package: Status and challenges,” *IEEE Trans. on Adv. Packag.*, pp. 286–300, 2004.

[3] S. Pant and E. Chiprout, “Power grid physics and implications for CAD,” in *Proc. DAC*, 2006.

[4] K. Sheth, E. Sarto, and J. McGrath, “The importance of adopting a package-aware chip design flow,” in *Proc. DAC*, 2006.

[5] H. Zheng, B. Krauter, and L. Pileggi, “On-package decoupling optimization with package macromodels,” in *Proc. CICC*, 2003.

[6] A. Devgan, H. Ji, and W. Dai, “How to efficiently capture on-chip inductance effects: introducing a new circuit element K,” in *Proc. ICCAD*, 2000.

[7] J. Chen and L. He, “Noise-driven in-package decoupling capacitance insertion,” in *Proc. DAC*, 2005.

[8] H. Yu, Y. Shi, and L. He, “Fast analysis of structured power grid by triangularization based structure preserving model order reduction,” in *Proc. DAC*, 2006.

[9] H. Yu, Y. Shi, L. He, and D. Smart, “A fast block structure preserving model order reduction for inverse inductance circuits,” in *Proc. ICCAD*, 2006.

[10] A. E. Ruehli, “Equivalent circuits models for three dimensional multiconductor systems,” *IEEE Trans. MTT*, pp. 216–220, 1974.

[11] X. Li, P. Li, and L. Pileggi, “Parameterized interconnect order reduction with explicit-and-implicit multi-parameter moment matching for inter/intra-die variations,” in *Proc. ICCAD*, 2005.

[12] P. Feldmann and F. Liu, “Sparse and efficient reduced order modeling of linear sub-circuits with large number of terminals,” in *Proc. ICCAD*, 2004.

[13] P. Liu, X. Tan, and et. al., “Efficient method for terminal reduction of interconnect circuits considering delay variations,” in *Proc. ICCAD*, 2005.

[14] P. Li and W. Shi, “Model order reduction of linear networks with massive ports via frequency-dependent port packing,” in *Proc. DAC*, 2006.

[15] C. Ding, “Spectral clustering, principal component analysis and matrix factorizations for learning,” in *Int’l Conf. on Machine Learning (Tutorial)*, 2005.

[16] G. Karypis, R. Aggarwal, and V. K. S. Shekhar, “Multilevel hypergraph partitioning: application in VLSI domain,” *IEEE Trans. VLSI*, pp. 69–79, 1999.