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

Hao Yu
Berkeley-DA Inc.
Santa Clara, CA 95054

Chunta Chu, Lei He
EE Department UCLA
Los Angeles, CA 90095

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 parameterized in each block. The decaps are then allocated according to the sensitivity obtained from the parameterized macromodel for each block. Compared to the PRIMA-based macromodeling, experiments show that our method (TBS2) 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

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. However, different from the on-chip decap, off-chip decaps are discrete passive components. Moreover, 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 following decap insertion algorithms have been developed recently. [4] calculates a multi-input multi-output (MIMO) impedance by model order reduction. It reduces the cost of decaps and resonance impedance in the frequency-domain. To explicitly consider the rising-time of the I/O current, [5] allocates decap to directly reduce the noise in the time-domain and avoids over-design compared to [4]. However, the simulated annealing (SA) based algorithms in [4, 5] is capable of dealing with a pre-designed package with only a limited legal positions as the algorithm virtually tries on each legal position. Moreover, the models used in [4, 5] need to be improved. The accuracy of PRIMA-based reduction in [4] decays when the I/O port number increases. In addition, the reduction in [4] ignores the structure information. The reduced model is dense and non-localized, and is inefficient to handle large-scale packages. On the other hand, [5] 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 [5] can lead to over-design.

Considering chip-package co-design, this paper formulates a decap allocation problem to minimize the decap cost subject to a dynamic noise violation constraint with consideration of noise pulse width. We develop a scalable algorithm using a localized and parameterized macromodel. The primary contributions of our paper are two-fold. First, to generate an effective macromodel considering large numbers of I/O ports, we propose a spectral clustering of I/Os based on correlation. The information of clustered I/Os is further used to partition the large RLC-network and leads to a structured macromodel with localized integrity analysis. Compared to the macromodel used in [4], our method is 3.04X more accurate and 25X more efficient. Secondly, given a large number of legal positions, we introduce a systematical decap allocation based on sensitivity of the transient response with respect to the decap location and type. Then, the decaps can be allocated according to sensitivity from a structured and parameterized macromodel that is only calculated once. Compared to the SA-based allocation in [4, 5], 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

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. 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

*This work was partially supported by NSF-CAREER 0401682 and UC-MICRO sponsored by Altera, Intel and Rio Design Automation. Address comments to hao@berkeley-da.com and lhe@ee.ucla.edu.

DAC 2007, June 4–8, 2007, San Diego, California, USA.

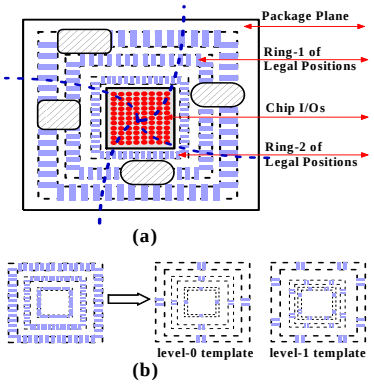

Figure 1: (a) The typical digram of package plane, chip I/Os and legal positions for decaps. The legal positions are represented by multiple rings. (b) Rings are decomposed into levelized templates, and are further partitioned and analyzed independently in each block.

rings with different distances. After one decap is assigned to one legal location the decap is then called *legally placed*.

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 uniformly distributed on one edge. The illegal positions are not included in each ring.

Moreover, because of non-uniformly distributed I/Os in space, the orientation of legal positions can have different impact on 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 edge centers, 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 $\mathbf{T} = [\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_M]$, 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 [4, 5] to choose during the realistic design. Therefore, the number of M is still manageable.

In addition, to consider the dynamic noise taking in to account the noise pulse width, the noise integral instead of noise amplitude is used

$$f_i = \int_{t_0}^{t_p} max[y_i(\mathbf{T}, t), Vc_i]dt = \int_{t_s}^{t_e} [y(\mathbf{T}, t) - Vc_i]dt, \qquad (1)$$

with a pulse-width (t_s, t_e) in a sufficient long time-period t_p. $y_i(\mathbf{T}, t)$ is transient noise waveform at i-th I/O. This applies to all p I/O cells, i.e., $f_j \leq Vd_j$ $(j = 1, .., p)$.

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

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

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

$$s.t.\ \mathbf{Uf} \leq \mathbf{Vc}\ and \sum_{j}^{M_1} m_j \leq \mathcal{M}. \qquad (2)$$

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 $(j = 1, ..., M_1)$. This problem can be efficiently solved by calculating 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 integrity analysis. As shown by [6], the inverse of L (L^{-1}) 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:

$$(\mathcal{G}_0 + s\mathcal{C}_0)x(s) = \mathcal{B}\mathbf{I}(s),\ y(s) = \mathcal{B}^T x(s) \qquad (3)$$

with

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

and

$$\mathcal{G}_0 = \begin{bmatrix} G & \mathbf{E}_l L^{-1} \\ -\mathbf{E}_l^T L^{-1} & 0 \end{bmatrix}, \mathcal{C}_0 = \begin{bmatrix} C & 0 \\ 0 & L^{-1} \end{bmatrix} \qquad (4)$$

Note that $\mathcal{G}_0$ and $\mathcal{C}_0$ are nominal state matrices composed by conductance G, capacitance C, inverse-inductance L and incident-matrix $\mathbf{E}_l$. x is the state variable composed by the nodal voltage $\mathbf{v}_n$ and the branch vector-potential $\mathbf{a}_l$. In addition, $\mathcal{B}$ is the adjacent matrix to describe P identical input and output ports, 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 when we insert decaps with associated equivalent conductance g_i, capacitance c_i and susceptance s_i (inverse of inductance). We describe each template $\mathcal{T}_i$ 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 susceptance. For an i-th template, adding decaps between tiles m and n results in:

$$\mathcal{T}_i^g(k,l) = \mathcal{T}_i^g(l,k)$$
$$= \begin{cases} -g_i & \text{if } k = m,\ l = n \text{ and } k \neq l \\ \sum_l |\mathcal{T}_i^1(k,l)| & \text{if } k = l \\ 0 & \text{else} \end{cases} \qquad (5)$$

where $k, l \in 1, 2, \cdots, N$. $\mathcal{T}_i^c(k,l)$ can be given similarly to add the equivalent capacitance and susceptance.

To separate the sensitivity from the nominal response, the state variable $x(\mathbf{T}, s)$ is first expanded into Taylor series with respect to $\mathcal{T}_i$, and a new state variable x_{ap} is reconstructed by the nominal values and the first-order sensitivities $x_{ap} = [x_0^{(0)}, x_1^{(1)}, ..., x_M^{(1)}]^T$. Then, a dimension-augmented system can be reorganized according to the expansion order

$$(\mathcal{G}_{ap} + s\mathcal{C}_{ap})x_{ap} = \mathcal{B}_{ap}\mathbf{I}(s), \quad y_{ap} = \mathcal{B}_{ap}^T x_{ap}, \qquad (6)$$

where

$$\mathcal{G}_{ap} = \begin{bmatrix} \mathcal{G}_0 & 0 & \dots & 0 \\ \mathcal{T}_1^g & \mathcal{G}_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{T}_M^g & 0 & \dots & \mathcal{G}_0 \end{bmatrix}, \quad \mathcal{C}_{ap} = \begin{bmatrix} \mathcal{C}_0 & 0 & \dots & 0 \\ \mathcal{T}_1^c & \mathcal{C}_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{T}_M^c & 0 & \dots & \mathcal{C}_0 \end{bmatrix}, \qquad (7)$$

both have a lower triangular block structure. Due to the increased dimension, the augmented system needs be reduced and represented by a macromodel.

4. SPECTRAL CLUSTERING OF I/O

Due to the large number of I/O ports, the macromodel by model reduction applied by [4] is still ineffective. Because the time/space-variant input I/O currents are not independent, 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). This becomes the motivation to apply the singular value decomposition (SVD) [7, 8] based port reduction as SVD is equivalent to eigen-decomposition when the matrix to be decomposed is symmetric positive definite. These approaches [7, 8] assume that the correlation or similarity of inputs can be inferred from a SVD analysis of the system transfer function. However, the correlation of I/O ports is actually 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.

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 I/O current sources with PCA and K-means.

In this paper, we propose to directly study the similarity or correlation of I/O currents. The large number of I/Os are clustered into K groups, each with one principal I/O current as port. Given a typical set of P 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

$$\mathcal{C}(i, j) = \frac{cov(i, j)}{\sigma_i \cdot \sigma_j}, \tag{8}$$

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 precomputed and stored in a table.

After extracting the correlation for I/O 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 [9] 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

$$\mathcal{J}_x = V\mathcal{J} = V\mathcal{B}\mathbf{I}(s) \quad \in R^{1\times K}. \tag{9}$$

It is equivalent to reduce the port matrix

$$\mathcal{B}_x = V\mathcal{B} \qquad \in R^{N\times K}. \tag{10}$$

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 I/Os are strongly correlated.

5. LOCALIZED INTEGRITY ANALYSIS

Because the I/O currents are distributed non-uniformly in space, they have different impact on voltage bounces at different locations. 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 [10] leverages the property of latency, which is more suitable for timing simulators. But for the verification of power integrity, it is more meaningful to study a partition based on I/O currents. Moreover, the partition in TBS [10] 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. As shown in [6], this can be solved by Bordered-Block-Diagonal (BBD) decomposition of the VNA network $(\mathcal{G}_0, \mathcal{C}_0, \mathcal{B}_x)$. Note that each block has the specified ports A_i obtained from the spectral clustering.

As a result, the parameterized system can be described in a BBD matrix by

$$\mathcal{G}_{ap} \rightarrow \mathbf{G}_{ap} = \begin{bmatrix} \mathbf{G}_1 & \cdots & 0 & X_{1,0} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \mathbf{G}_K & X_{K,0} \\ -X_{1,0}^T & \cdots & -X_{K,0}^T & Z_r \end{bmatrix}$$

where $\mathcal{G}_0$ is partitioned into K blocks $\mathcal{G}_j$ $(j = 1, ..., K)$, and $\mathbf{C}_{ap}$ is built similarly from $\mathcal{C}_{ap}$. Note that 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 for the BBD formulation. As such, the reduction can be performed locally [6]. 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 [10] is extended to consider the parameterized BBD state matrices with the inverse-inductance. 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

$$(\widetilde{\mathbf{G}}_{tb} + \frac{1}{h}\widetilde{\mathbf{C}}_{tb})\widetilde{x}_{tb}(t) = \frac{1}{h}\widetilde{\mathbf{C}}_{tb}\widetilde{x}_{tb}(t-h) + \widetilde{\mathbf{B}}_{tb}\mathbf{I}(t)$$
$$\widetilde{y}_{tb}(t) = \widetilde{\mathbf{B}}_{tb}^T\widetilde{x}_{tb}(t). \tag{11}$$

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

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

.

6. ALGORITHM AND RESULTS

With use of the structured and parametrized macromodel in Section 5, the problem formulation (2) in Section 2 can be efficiently solved by the sensitivity based optimization. The overall optimization is outlined below. The nominal value and sensitivity are computed one-time from the structured and parameterized macromodel from (11). 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: $\{\delta y_{i_1,k}, ..., \delta y_{i_M,k}\}$, 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.

The proposed macromodeling and allocation algorithm are implemented in C/Matlab. We call our macromodeling method as TBS2, and our optimization as multi-ring based allocation

Table 1: Results of decap allocations by SA and our MRA method. The cost of decaps is normalized.

ckt (#node+#I/O)	#level	#legal-pos	#partition	SA-NA		MRA-NA		MRA-NI	
				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

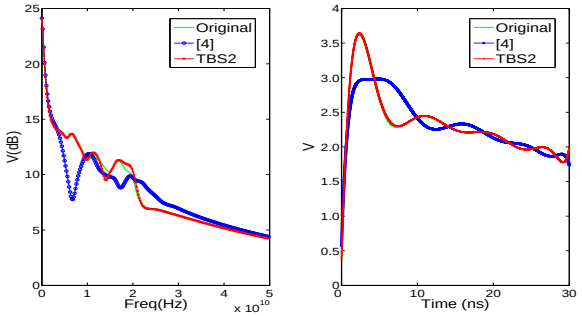

Figure 3: Waveform accuracy comparisons between the original, the method in [4] and TBS2 in (a) time-domain and (b) frequency-domain for 4th principal port. The original and TBS2 are visually identical.

(MRA). Experiments are run on a Linux workstation with 2G RAM. A typical FPGA package model is assumed with the specific application inputs . Four packages P/G planes are assumed with the size of $1cm \times 1cm$. The Vdd is assumed to be 2.5V, and the targeted noise is 0.25V. The worst-case I/O currents are modeled as triangle-waveforms with rising time 0.2ns, width 1ns and period 150ns, which are randomly distributed in a square of $0.2cm \times 0.2cm$ located in the center of package planes. The 30% of remaining area are reserved for legal postions. The 4 decap-types in [5] are used. The total number of decaps and rings are 80 and 5. Each ring decomposed 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 the noise amplitude (NA) or the noise integral (NI).

6.1 Comparison of Macromodels

We first compare our method with macromodeling in [4] 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 [4]. 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 [4] has about 3.04X larger waveform error in the time-domain.

6.2 Comparison of Allocations

We also compare the runtime and the cost of allocated decaps between SA and MRA, where 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.

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 NI can consider the noise pulse width, and accurately predict the decap allocation. As a result, MRI reduces the cost by up to 16% compared to the SA using NA [5].

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. This clustering enables an I/O-based network partition with a localized integrity analysis. In addition, to systemically allocate decaps, we have also proposed a ring-based decap allocation based on the sensitivity, which is generated from a structured and parameterized macromodel. Experiments show that compared to the existing methods, 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 et.al., "Power supply noise analysis methodology for deep-submicron VLSI chip design," in *Proc. DAC*, 1997.

[2] K. Sheth and et.al., "The importance of adopting a package-aware chip design flow," in *Proc. DAC*, 2006.

[3] S. Pant and et.al., "Power grid physics and implications for CAD," in *Proc. DAC*, 2006.

[4] H. Zheng and et. al., "On-package decoupling optimization with package macromodels," in *Proc. CICC*, 2003.

[5] J. Chen and et.al., "Noise-driven in-package decoupling capacitance insertion," in *Proc. ISPD*, 2006.

[6] H. Yu and et.al., "A fast block structure preserving model order reduction for inverse inductance circuits," in *Proc. ICCAD*, 2006.

[7] P. Feldmann and et.al., "Sparse and efficient reduced order modeling of linear sub-circuits with large number of terminals," in *Proc. ICCAD*, 2004.

[8] P. Liu and et.al., "Efficient method for terminal reduction of interconnect circuits considering delay variations," in *Proc. ICCAD*, 2005.

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

[10] H. Yu and et.al., "Fast analysis of structured power grid by triangularization based structure preserving model order reduction," in *Proc. DAC*, 2006.