Congestion-Aware Power Grid Optimization for 3D Circuits Using MIM and CMOS Decoupling Capacitors

Pingqiang Zhou
Department of ECE
University of Minnesota
Minneapolis, MN 55455
pingqiang@umn.edu

Karthikk Sridharan
Department of ECE
University of Minnesota
Minneapolis, MN 55455
sridh019@umn.edu

Sachin S. Sapatnekar
Department of ECE
University of Minnesota
Minneapolis, MN 55455
sachin@umn.edu

Abstract— In three-dimensional (3D) chips, the amount of supply current per package pin is significantly more than in two-dimensional (2D) designs. Therefore, the power supply noise problem, already a major issue in 2D, is even more severe in 3D. CMOS decoupling capacitors (decaps) have been used effectively for controlling power grid noise in the past, but with technology scaling, they have grown increasingly leaky. As an alternative, metal-insulator-metal (MIM) decaps, with high capacitance densities and low leakage current densities, have been proposed. In this paper, we explore the tradeoffs between using MIM decaps and traditional CMOS decaps, and propose a congestion-aware 3D power supply network optimization algorithm to optimize this tradeoff. The algorithm applies a sequence-of-linear-programs based method to find the optimum tradeoff between MIM and CMOS decaps. Experimental results show that power grid noise can be more effectively optimized after the introduction of MIM decaps, with lower leakage power and little increase in the routing congestion, as compared to a solution using CMOS decaps only.

I. Introduction

Three dimensional (3D) circuit technologies, with multiple tiers of active devices stacked above each other, provide the potential to increase transistor packing density and reduce chip area significantly in comparison with today's 2D ICs [1]. In other words, for the same chip footprint, 3D provides a way of continuing along the path of increased integration along the Moore's law curve that is orthogonal to device shrinking and technology scaling. Recent technological advances have permitted 3D tiers to be stacked with very short inter-tier distances. A schematic of a 3D chip is illustrated in Figure 1 showing five tiers stacked over each other. The lowest tier sits over a bulk substrate, while the backs of dies on the other tiers are thinned to remove the substrate. These technologies have been shown to provide intertier distances of the order of a few microns. The tiers may be placed face-to-face, face-to-back, or back-to-back: in this figure, every tier is face-to-back with its neighboring tier.

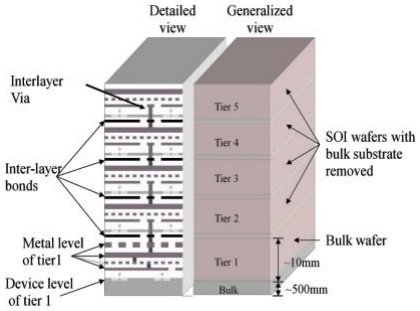

Fig. 1. A schematic view of a 3D integrated circuit.

3D technologies have numerous potential benefits: it can significantly improve circuit performance by reducing interconnect wire lengths and delays, core-to-memory latencies, and compacted wirelength histograms that imply reduced congestion; it permits heterogeneous integration using different materials in each tier; it can allow for improved noise isolation between analog and digital blocks by placing them on different tiers, with noise isolation, etc.

However, there are two significant limitations that 3D technologies must overcome before achieving their full potential, related to on-chip thermal issues and reliable power delivery. Both issues can be illustrated through a simple back-of-the-envelope calculation. A k-tier 3D chip that stacks k similar chips could use k times as much current as a single 2D chip of the same footprint. However, the packaging technology is not appreciably different: with a similar heat sink, the on-chip temperature on such a 3D chip is k times higher than the 2D chip, and with a similar number of pins in the package, the current per pin is k times higher than the 2D case [2].

The above analysis operates under very coarse assumptions (for example, a smart 3D designer may not stack k layers with identical power levels), and a more nuanced approach is necessary for a more accurate analysis – but the eventual conclusions that thermal and power delivery issues are important in 3D – are inescapable. While much research has been conducted on thermal management strategies such as thermal via insertion [3] and spatial distribution of power sources [4]–[7], the power delivery problem has attracted limited attention to date, e.g., in the work in [8].

The power delivery problem can be summarized as follows. The parasitics in the power network, together with temporal variations in the current drawn by a circuit, result in a time-varying voltage drop/surge at nodes in the power grid. These variations can adversely impact the performance and the reliability of a circuit [9]. Such shifts become more acute with technology scaling: on the one hand, noise margins become more stringent with reducing Vdd levels, and on the other hand, with increased switching speeds and larger currents, IR, LdI/dt, and electromigration effects become more prominent [10]. In 3D circuits, robust power supply network design is more challenging, and significant resources have to be invested in building a bulletproof power grid for the 3D chip [2].

Several techniques are available to increase the reliability of power grids and control power grid noise, such as wire widening, grid topology optimization, and decap insertion. Of these techniques, decaps are arguably the most powerful method for reducing transient noise, and are therefore addressed in this paper. Decaps serve as local current reservoirs, and can be used to satisfy sudden surges in current demand by the functional blocks/cells, while keeping supply voltage levels relatively stable.

Conventional technologies for implementing decaps are based on SiO_2-based structures that are widely used in robust power delivery network design. In the recent past, the CMOS decap allocation and optimization problem has been investigated by numerous researchers for 2D [9], [11]–[16], [18] and 3D technologies [17]–[19].

Unlike the 2D case, new considerations come into play while optimizing a 3D power grid using CMOS decaps:

- Since CMOS decaps are usually fabricated using white space on the device layer, they must compete for area with through-silicon vias, or with the landing pads of 3D vias, for the limited white space. This leads to a new resource contention problem.
- One way to resolve this contention problem is to increase the chip size in order to make room for CMOS decaps. However, one of the advantages of 3D circuits over 2D implementations

is their reduced chip footprint: increasing the chip size may counteract this benefit.
- Leakage power is an important issue in 3D circuit design. The CMOS decaps added to the 3D circuit will consume extra leakage power, and make things worse. While new high-k dielectrics have been proposed, they will provide temporary relief to the gate leakage problem.

In this work, we address all of these issues. One of the novel features of our work is that it optimizes the power supply network using both conventional CMOS decaps and the newer MIM decap technology. Unlike CMOS capacitors that are built in the device layer, MIM capacitors are fabricated between metal layers. These structures have high capacitance density and low leakage current density [20]–[25]. However, MIM decaps cannot be used unconditionally to replace CMOS decaps, since their use incurs a cost: they present routing blockages to nets that attempt to cross them. The properties of MIM decaps makes them attractive for both 2D and 3D chips, but we pay particular attention to the 3D decap problem in this paper because (i) the power integrity problem is particularly critical in 3D, and requires novel approaches that leverage advances in materials, and (ii) the added complexity of handling routing blockages in a very constrained environment makes the 3D problem especially challenging. To the best of our knowledge, this is the first work to develop CAD solutions for inserting MIM decaps in power grid design and optimization. We formulate the decap budgeting problem as a Linear Programming (LP) problem, and propose an efficient congestion-aware algorithm to optimize the power supply noise, while trying to find a balance between the routing congestion deterioration and leakage power increase. Although we focus on the 3D decap allocation problem in this paper, our algorithm can be extended to solve the 2D power grid optimization problem in cases where MIM decaps are utilized.

II. BACKGROUND

A. MIM Decaps

MIM decaps are typically useful for MPUs, RF capacitors in high frequency circuits, as well as filter and analog capacitors in mixed-signal products [23]. Recently, several successful implementations of high-performance MIM decaps have been reported in [20]–[25]. In [23], a high reliability MIM capacitor is reported to be integrated into a $0.18\mu m$ CMOS foundry technology using copper interconnects. In [20], high performance ALD HfO_2-Al_2O_3 laminate MIM capacitor is fabricated with high capacitance density of $12.8fF/\mu m^2$ (for reference, at the 90nm node the capacitance density of CMOS capacitors can be estimated as ε_{ox}/t_{ox} = $17.3fF/\mu m^2$, where ε_{ox} is the dielectric constant and $t_{ox} \approx 20$Å is the thickness of oxide). A successful implementation of large-area MIM capacitors (exceeding $250nF$) in the power grid of a 90nm SOI microprocessor, with up to $8fF/\mu m^2$ capacitance density, is reported in [24].

A significant advantage of MIM decaps lies in their extremely low leakage: in [24], the leakage current for the $250nF$ MIM decaps is reported to be about $1.0 \times 10^{-8}A$ (with leakage density of $3.2 \times 10^{-8}A/cm^2$), while the leakage current for a $25nF$ CMOS decap in parallel with MIM is approximately $3.2 \times 10^{-6}A$ (with leakage density of $1.45 \times 10^{-4}A/cm^2$).

Figure 2 shows an example with CMOS and MIM decaps in one tier of a 3D circuit. MIM decaps are usually fabricated between the top two metal layers in each 2D tier. In this case, they form a blockage between the top 2 metal layers, and for 3D intertier vias.

It is important to note that MIM decaps also act as a routing blockage, and their use can lead to increased congestion. Although it is possible to slot the decap to allow wires through, this leads to inaccuracy in estimating the decap value, and we regard it as an option of last resort, for ECO fixes or for resolving overconstrained congestion. For the purposes of this paper, we regard a MIM decap as a routing blockage that prevent connections between the top two metal levels, and prevent 3D vias from passing through the region.

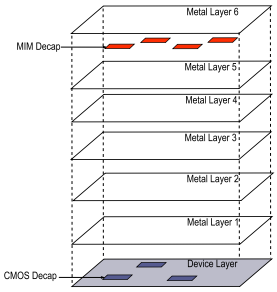

Fig. 2. MIM and CMOS decaps in one 2D tier with 6 metal layers.

B. 3D Power Grid

A 3D circuit consisting of k stacked 2D tiers can be thought to consist of a 3D Power/Ground (P/G) supply network with k stacked 2D P/G networks. Each 2D P/G network may contain several orthogonal metal layers, with increasing P/G pitches from bottom to top. The connections between the P/G subnetworks in adjacent tiers are facilitated by 3D vias, and connections from the topmost tier to the package. In other words, a 3D P/G grid consists of several 2D P/G grids with vertically adjacent P/G nodes connected by 3D vias.

Given the similarity between power and ground grids, we only describe the algorithm for power grid in this paper. As in other work on this topic, e.g., [12], we use a linear circuit model to analyze the voltage noise of the power supply grid. The power grid is modeled as a resistive mesh, the cells/blocks as time-varying current sources, decoupling capacitors as lumped capacitors connected to the power and ground planes, and 3D vias as RC elements. The top-level metal of the top die is connected to a package pin, which is modeled as an inductance connected to an ideal off-chip constant voltage source. The behavior of such a 3D circuit is described by a first order differential equation formulated using the modified nodal analysis (MNA) method [26].

C. Noise Violation Metric

The voltage waveform at every node can be computed through a transient analysis of the circuit. To efficiently measure the dynamic voltage drop, we follow definition of the noise metric in [12] and formulate the *violation area*, S_i, at each power node i as:

$$\begin{aligned} S_i &= \int_0^T max\{V_{limit} - v_i(t), 0\}dt \\ &= \sum_j \int_{t_{s,j}}^{t_{e,j}} max\{V_{limit} - v_i(t)\}dt \end{aligned} \quad (1)$$

Here, V_{limit} is the voltage threshold, usually set to be 10% of V_{dd}. The symbol $v_i(t)$ is the transient voltage at node i, and $[t_{s,j},\ t_{e,j}]$ is the j^{th} interval during which the constraint is violated.

A key advantage of this metric is therefore that it permits the noise violation, which is usually expressed as a constraint for power grid optimization, to be incorporated into an objective function to be minimized. The goal of our optimization is to reduce the violation area to zero at all nodes, with optimal resource usage.

III. PROBLEM FORMULATION

We tile the layout using an uniform grid G' that is coarser than the original power grid, G, so that each tile of G' contains less than 20 power nodes in G. Our algorithm proceeds iteratively, adding a small amount of decap to the circuit in each iteration. An observation node is dynamically chosen from G for each tile in G', and all newly added decaps in this tile are connected to this observation node in each iteration. This helps in reducing the number of possible decap insertion spots, thus controlling the size of the problem that we solve.

A. Objective function

We denote the newly added CMOS and MIM decaps in tile k by Δx_k and Δy_k, respectively in each iteration. Let $S = \sum_{i=1}^{n} S_i$ be the total violation area over all the n nodes in the supply grid. The objective function in each iteration is to minimize the total increase in the leakage power, ΔP, while maximizing the reduction in the violation area S, i.e.,

$$\text{minimize} \qquad \alpha \cdot \Delta S + (1-\alpha) \cdot \Delta P \tag{2}$$

Here,

1) α is a weighting parameter that sets the objective to be a convex combination of the noise violation and the leakage power.
2) ΔS is the change in the violation area S when a small amount of CMOS decap and/or MIM decap is added to each tile k. Since the amount of decap inserted in each iteration is small, this change may be computed as

$$\Delta S = \sum_{k=1}^{m'} \{(\partial S/\partial x_k) \cdot \Delta x_k + (\partial S/\partial y_k) \cdot \Delta y_k\}$$

where m' is the number of tiles in G', $\partial S/\partial C$ is the sensitivity of S with respect to the decap $C \in \{x_k, y_k\}$, and Δx_k and Δy_k are as defined above. We note that $\partial S/\partial x_k$ and $\partial S/\partial y_k$ are nonpositive, since the violation area must decrease when decaps are added to the circuit. Therefore, minimizing ΔS, which is nonpositive, implies that we maximize the absolute reduction in S.

3) ΔP corresponds to a leakage term, and is calculated as $\sum_{k=1}^{m'} (a_k \cdot \Delta x_k + b_k \cdot \Delta y_k)$. In other words, it is the weighted sum of the increase in leakage due to the newly added decaps Δx_k and Δy_k. The weights a_k and b_k are given by

$$a_k = \frac{LD_{CMOS}}{CD_{CMOS}} \cdot \phi(T_k) \tag{3}$$

$$b_k = \frac{LD_{MIM}}{CD_{MIM}} \cdot \phi(T_k) \tag{4}$$

Here, LD_{CMOS}, LD_{MIM}, CD_{CMOS}, and CD_{MIM} are, respectively, the leakage densities of CMOS and MIM decaps, and the capacitance densities of CMOS and MIM decaps, and T_k is the average temperature in the tile k. The ratio $\frac{\Delta x_k}{CD_{CMOS}}$ provides the area of the added decap, which when multiplied by LD_{CMOS} determines the corresponding leakage. The penalty term $\phi(T_k) = T_k^2 \cdot exp(\mu/T_k^2)$ captures the effect of temperature on each leakage term, where μ is a constant negative number [7]. A higher temperature T_k corresponds to a larger $\phi(T_k)$, which means that the increase in leakage in tile k is controlled more strictly.

Considering that ΔS and ΔP may have different orders of magnitude, to better control the coefficients of the objective function, we first normalize $\partial S/\partial x_k$, $\partial S/\partial y_k$, a_k and b_k to the interval [0,1] and then add them up using the weighting parameter α.

B. Constraints

1) *Congestion constraints*. As mentioned in Section II.A, the MIM decaps inserted between metal layers may become potential routing blockages. Therefore, it is necessary to impose a constraint that restricts the deterioration of congestion with MIM decap insertion. This constraint is written as:

$$\Delta Cong_k \leq \gamma \cdot Cong_k \tag{5}$$

where $Cong_k$ is the current congestion value in tile k, $\Delta Cong_k$ is the change of the congestion in tile k in the current iteration, and γ is a bounding parameter, which is empirically set to be 0.03 to 0.05 in our experiments.
Since each iteration imposes only a small change in the inserted decaps, it is reasonable to formulate $\Delta Cong_k$ as a linear function of the inserted MIM decaps $\Delta Cong_k = \sum_{i \in R_k} (c_i \cdot \Delta y_i)$, where the set R_k and the justification for this term are described in detail in Section IV.

2) *Decap resource constraints*. For a tile k, the amount of CMOS decap that can used is limited by its available white space, and the amount of MIM decap is restricted by its capacity. If C^k_{CMOS} and C^k_{MIM} are the current maximum allocatable amount of CMOS and MIM decaps in tile k, then the decap resource constraints for tile k can be formulated as:

$$0 \leq \Delta x_k \leq min\{\Delta_{CMOS}, C^k_{CMOS}\} \tag{6}$$

$$0 \leq \Delta y_k \leq min\{\Delta_{MIM}, C^k_{MIM}\} \tag{7}$$

where Δ_{CMOS} and Δ_{MIM} are upper bounds that are chosen to control the amount of CMOS and MIM decaps inserted in each iteration.

Equations (2)-(7) together formulate a linear programming problem, which can be solved by any standard linear programming solver.

IV. Congestion Analysis And Linear Congestion Model

A. 3D Congestion Analysis

We estimate the routing congestion for decap optimization in 3D circuits using a probabilistic method. Given a placed 3D netlist, the core area is discretized using a 3D mesh, using the grid G' described in Section III, and the congestion in each tile of this mesh is estimated. The congestion cost of a tile cell in the X, Y and Z directions is defined as the ratio of the usage to the capacity in that direction. The capacities of the tile cells depend on the sizes of the tile cells and process technology parameters. For the purposes of our algorithm, the congestion in the Z direction is the most important: since the uppermost two layers primarily consist of supply/clock wires rather than signal wires within a single tier, MIM decaps primarily affect signal routes in the Z direction. However, other terms in the objective function can act to provide disincentives to large area capacitors which would create significant bottlenecks to power/clock wires in the X, Y, and Z directions as well.

The routing usage values are calculated using a probabilistic congestion model similar to the one proposed in [27], extended to the three dimensional case. A minimum spanning tree (MST) is constructed for each multipin net, and this is used to decompose the multipin net into 2-pin net pairs. For each 2-pin net, a bounding box is constructed. Assuming monotonic routes, the probabilistic usage within a tile cell is calculated as the ratio of the number of tracks in that direction used in the tile cell to the total number of possible routes in the bounding box.

Consider a box of dimension $p \times q \times r$, where p, q, and r are the number of tile cells in the X, Y and Z direction respectively. Let $F(p, q, r)$ be the total number of possible routes in this bounding box, starting from the origin, at the bottom left lowermost corner, going to the top right uppermost corner. This must satisfy the following recurrence relation:

$$F(p,q,r) = F(p-1,q,r) + F(p,q-1,r) + F(p,q,r-1) \tag{8}$$

Using an argument parallel to [27], Equation (8) is valid because the number of routes in the bounding box of size $p \times q \times r$ is the sum of the mutually exclusive set of routes from the origin to its extreme distal corner in a set of bounding boxes of dimension $(p-1) \times q \times r$, $p \times (q-1) \times r$ and $p \times q \times (r-1)$. The basis case of this recurrence relation can be stated as:

$$F(p,1,1) = F(1,q,1) = F(1,1,r) = 1 \tag{9}$$

In other words, when the bounding box has many tile cells in only one direction and a single tile cell each in the other two directions, there is only one possible route.

After computing the probabilities corresponding to each 2-pin decompositions using the method stated above, the usage of a tile cell is computed by adding up all such probabilities.

B. Approximate Linear Congestion Model

The method described in Section IV.A is used to calculate the initial congestion map of the circuit, and is predicated on the assumption that there are no blockages in the region. When a MIM decap is inserted, it results in a blockage and causes a perturbation in the congestion values. As mentioned in Section III.B, we model this change in the congestion in a tile cell, assuming a small perturbation as a linear function. We now describe the procedure used to calculate this linear function using the initial congestion map.

Let R_k be a set of tile cells (including k) within a specified Manhattan radius, $maxDist$, of a tile cell k. We assume that the size of R_k is bounded by a small number, reflecting the fact that we operate under small perturbations that do not cause widespread congestion changes far away from k. For each tile cell $i \in R_k$, let W_i be the current number of routes in tile cell i, and let $CurCap_i$ and $NewCap_i$ be the current and new routing capacities in tile cell i after the insertion of a MIM decap.

Let ΔW_i be the number of routes in the tile cell i to be redistributed. The redistribution process proceeds as follows after a small additional MIM decap, Δy_i, is inserted in tile cell i. If W_i is smaller than the current capacity, $CurCap_i$, then none of the routes in tile i need to be redistributed but the congestion values are updated to reflect the reduction in the capacity. Otherwise, it is necessary for routes in tile i to be redistributed. The number of routes to be moved out of tile i, to neighboring tile cells, is calculated as:

$$\Delta W_i = W_i \times \frac{CurCap_i - NewCap_i}{CurCap_i} \tag{10}$$

The redistribution depends on the Manhattan distance of a cell from i. For a tile cell k that is at a distance d from cell i ($k \neq i$), the number of routes added is computed as:

$$\Delta W_{k,i} = \frac{1}{4d} \times \frac{\omega}{d} \times \Delta W_i \tag{11}$$

$$\text{where} \quad \omega = \frac{4}{\sum_{j=1}^{maxDist}(1/j^2)} \tag{12}$$

The term $\frac{\omega}{d}$ captures the fact that the number of routes added to a cell varies inversely with its distance d from cell i, and these are equally distributed among the $4d$ cells that lie at a Manhattan distance of d from i. The role of the factor, ω, is to ensure that the total number of routes redistributed equals ΔW_i. In our experiments, the value of $maxDist$ is set to be $1/3$ of the smaller of the number of tile cells in X and Y directions.

We then calculate $\Delta Cong_{k,i}$, the increase in congestion in tile cell k caused by $\Delta W_{k,i}$, as $\Delta Cong_{k,i} = \Delta W_{k,i}/CurCap_k = c_i \cdot \Delta y_i$ ($k \neq i$). This leads to the following linear approximation

$$\begin{aligned} \Delta Cong_k &= \left(\sum_{i \in R_k, i \neq k} \Delta Cong_{k,i} \right) + (c_k \cdot \Delta y_k) \\ &= \sum_{i \in R_k} (c_i \cdot \Delta y_i) \end{aligned} \tag{13}$$

where $c_k \cdot \Delta y_k$ is the congestion increase caused by the MIM decap Δy_k added to tile k.

V. Sequence-of-Linear-Program Based Solution

As stated in Section III, we use an iterative flow to solve the decap allocation problem. In each iteration we allocate a relatively small amount of decap to the current circuit, for two reasons. Firstly, the decap allocation problem is highly nonlinear, and this iterative approach permits us to control the optimization process by solving a sequence of linear programs, one in each iteration. Secondly, it avoids the excessive allocation of decaps that could invalidate the approximate linear model of congestion and violation area used in our algorithm: these models are predicated on the assumption of small perturbations.

The overall optimization flow can be formulated as follows:

1) Initial setup steps: solving the input 3D power grid, determining the set of nodes that violate the voltage specifications and computing the noise violation metric (Section II.C), building the coarser grid G' as described in Section III, generating the temperature map for the circuit using 3D thermal analysis, and evaluating $\phi(T_k)$ in each tile k of G'.
2) If violation node set is empty, then stop. Otherwise, for each tile k that contains at least one node that violates the voltage specification, select one observation node N_k. The node N_k is chosen to be the node i with the maximum violation area, S_i, in tile k.
3) For each tile that contains an observation node N_k, calculate $\partial S/\partial C_{N_k}$, the derivative of the total violation area S with respect to the decap C_{N_k} added at N_k using the adjoint analysis method described in [12].
4) For each tile k, calculate $\Delta Cong_k = \sum_{j \in R_k}(c_j \cdot \Delta y_j)$ using the method described in Section IV.
5) Formulate the linear programming problem described in Section III and solve it.
6) Update the decap budget using the solution from LP solver. For each tile k, if the solution Δx_k or Δy_k of current iteration is not zero, then we insert corresponding Δx_k CMOS decap or Δy_k MIM decap to tile k. Next, we update the current maximum allocatable amount of decap resource C^k_{CMOS} or C^k_{MIM} in tile k correspondingly.
7) Solve the circuit using the updated decap allocation, and update the set of violating nodes.
8) Update the current total violation area S.

VI. Experimental Results

The overall 3D power grid optimization flow has been written using Tcl, and the 3D power grid analyzer and the congestion and leakage aware decap allocation algorithm are implemented in C++. All experiments are performed on an Intel Pentinum 4 CPU 2.8GHz Linux machine with 1G memory running Redhat Linux 2.6.9.

The 3D placement tool in [6] is first applied to generate the 3D layouts from the IBM-PLACE benchmarks [28]. Next, we scale all the layouts to the $90nm$ technology node. Since the time-varying current sources, which model the behavior of each functional unit, are not originally available in these benchmarks, we use a method similar to [12] to generate the waveforms in each circuit. Six layers of regularly distributed power grid are generated for each 2D tier of a 3D circuit when building the 3D power grid. The supply voltage is set to be 1.2V and the voltage drop threshold is chosen to be 0.12V in each of the experiments. The capacitance densities for CMOS and MIM decaps are, respectively, set to be $17.3fF/\mu m^2$ (the oxide thickness is assumed to be $20\mathring{A}$) and $8.0fF/\mu m^2$. The leakage density of a CMOS decap is set to be $1.5 \times 10^{-5} mA/\mu m^2$, which is obtained from the simulation of a CMOS decap using PTM model [29]. For all of our experiments, the leakage density of the MIM decap is sufficiently small that it can be neglected.

A. Comparison Of Optimization Efficiency

TABLE I
Parameters Of Benchmarks

Circuit	# Nodes	Worst voltage droop (V)	# Violation nodes	Violation Area S (V · ns)
ibm123	18634	0.135	3330	13.739
ibm05	12026	0.122	1359	72.260
ibm08	17030	0.125	3191	41.305
ibm10	29262	0.159	5935	91.286
ibm18	75042	0.163	6392	108.649

Table I lists the parameters of the benchmarks used in our experiments. The circuit ibm123 is the combination of three ibm benchmarks: ibm01, ibm02 and ibm03.

TABLE II
COMPARISON OF OPTIMIZATION EFFICIENCY

Ckt	CMOS only						MIM only					CMOS + MIM					
	VNs	S (V·ns)	Lkg (mA)	Decap (pF)	#Iter	Time (s)	maxC (%)	avgC (%)	Decap (pF)	#Iter	Time (s)	Lkg (mA)	maxC (%)	avgC (%)	Decap (pF)	#Iter	Time (s)
ibm123	368	0.023	2.1	564	25	130	15.8	3.9	607	7	59	1.1	8.4	1.7	628	4	43
ibm05	24	0.049	2.7	480	5	24	19.7	1.7	550	23	111	2.1	0.0	1.2	546	22	109
ibm08	31	0.010	1.2	313	16	82	30.6	1.5	768	24	134	0.6	0.0	0.9	774	20	116
ibm10	351	0.182	1.6	417	12	108	10.6	5.9	511	11	186	0.9	4.5	2.5	520	4	133
ibm18	130	0.071	2.7	698	14	400	39.5	5.3	812	9	339	1.4	7.0	3.6	826	8	307

Table II lists the results of decap optimization in three different cases. First, only CMOS decaps are used: in this case, it is not possible to add enough CMOS decaps to eliminate the the violation area **S** (see column 3) for any of the five circuits. However, we list the results for the best available solution that minimizes this metric, showing the final number of violating nodes (**VNs**) that fail to meet the constraints, the corresponding violation area (**S**), the total leakage current of the CMOS decaps (**Lkg**), the total amount of CMOS decap allocated (**Decap**), the total number of iterations required by the optimizer (**#Iter**), and the total CPU time (**Time**).

Next, only MIM decaps are used: in this case, the violation area is completely eliminated by our procedure. Considering that the allocated MIM decaps will affect the routing congestion, we list the following results: the percentage increase in maximum and average Z-direction routing congestion after optimization (**maxC**, **avgC**), the total amount of MIM decap allocated (**Decap**), as well as the total number of iterations (**#Iter**) and the total CPU time (**Time**) for this case. Since MIM decaps have much smaller leakage density than CMOS decaps, for all practical purposes, their leakage is zero and is not shown in the table.

Finally, when both CMOS and MIM decaps are used, again, the violation area is completely eliminated. We list the total leakage current of the CMOS decaps (**Lkg**), the percentage increase in maximum and average Z-direction routing congestion after optimization (**maxC**, **avgC**), the total amount of Decap allocated (**Decap**), the total number of iterations (**#Iter**) and the total CPU time (**Time**).

From Table II we can see that for each of the five circuits, the violation area cannot be eliminated through the use of CMOS decaps only. This is due to the fact that the amount of CMOS decap that can be added in a circuit is limited by the available area of white space; moreover, for decaps to be effective, it is important for sufficient white space to be available near the area where the voltage constraints are violated. Placing decaps far away from the voltage violation area is of little help in alleviating noise violations. Therefore, unless we disturb the current placement or enlarge the chip size to make more white space available near the violation area, it is not possible to completely eliminate these violations.

The introduction of MIM decaps can effectively eliminate the voltage violations and greatly reduce the decap leakage, at the cost of worsened routing congestion. Table II shows that the use of MIM decaps alone leads to severe congestion problems. Comparing the results of using MIM and CMOS decaps individually with using them together, it can be seen that replacing part of the MIM decaps with CMOS decaps can obtain a better tradeoff between congestion and leakage, while effectively eliminating voltage violations.

Comparing the values of **Decap** for the MIM only and the CMOS+MIM cases, we can see that the decap values are similar (the values for CMOS-only are significantly different, since the constraints are not met in this case). The slight difference is attributable to approximations in linearizing the cost function in each iteration: specifically, in each iteration of our decap budgeting algorithm, an approximate formula, $\Delta S = (\partial S/\partial C) \cdot \Delta C$, is used to estimate the effect of added decap on the violation area, and this holds only when ΔC is small enough. In other words, in order to make the linear model more accurate, a smaller ΔC should be used, implying that the upper bounds for CMOS and MIM decaps in each iteration should be set to be very low (see Section III.B). This may lead to an increase in the number of iterations, impairing the computational efficiency of our approach. In our experiments, we found that a good balance between efficiency and accuracy can be obtained when Δ_{CMOS} and Δ_{MIM} are chosen to be in the region $[0.5pF, 1.0pF]$.

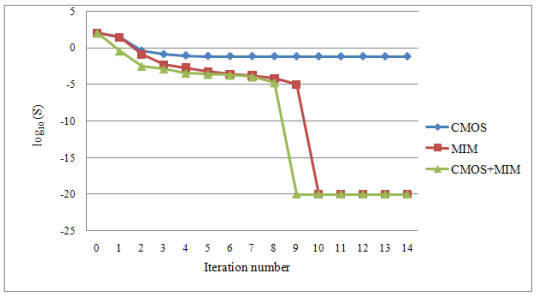

Fig. 3. Change in the total violation area over each iteration.

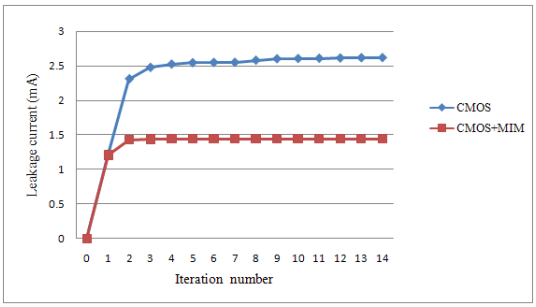

Fig. 4. Change in the total leakage current over each iteration.

Figures 3 and 4 show how the total violation area and total leakage current of circuit ibm18 change as the iterative process progresses. It can be seen that the CMOS-only case cannot bring the violation area down beyond some threshold, while the MIM-only and the CMOS+MIM methods are both successful (note that the extremely low violation value of approximately 10^{-20} is essentially zero).

Figure 3 shows that the total violation area decreases rapidly in the first 5 iterations, and most of the violations are eliminated after this stage (Note that the y-axis in this figure is on a log scale). A similar phenomenon can be observed in Figure 4: the total leakage current goes up quickly in the first 5 iterations and then plateaus out. The reason is as follows: most of the violations of the power nodes are relatively easily resolved by inserting a small amount of decap. Although the violation area of these nodes is individually small, their sum, taken over a large number of nodes, is large. Eliminating these "easy" violation at the beginning of the iterative process cause the violation area to decrease rapidly at first. Beyond this point, a relatively small number of "hard" violation nodes remain, and the change in the violation area is harder to view on the scale of this graph, but is definitely visible at a magnified scale. For the same reason, most of the white space resources that are effective in reducing noise violations are consumed in early iterations, resulting in a fast initial increase in the leakage power. The leakage component of the objective function implies that MIM decaps are preferred over

TABLE III
OPTIMIZATION RESULTS OF DIFFERENT POWER GRID DENSITIES

Cases	Power Grid Density	# Nodes	# Violation Nodes	Worst-case voltage droop (V)	Violation Area (V·ns)	Decap (fF)	Lkg (mA)	maxC (%)	avgC (%)	#Iter	Time (s)
Case1	Nominal	18634	3330 (17.87 %)	0.135	13.739	627564	1.1164	8.35	1.66	4	42.6
Case2	Denser	36433	4210 (11.56 %)	0.126	2.615	488372	0.6296	31.27	4.75	2	45.0
Case3	Densest	72114	4671 (6.48 %)	0.124	1.482	228595	0.2878	58.41	7.62	1	53.1

CMOS decaps when both are available, and when the insertion of MIM decaps does not significantly affect congestion.

B. Effect Of Power Grid Density

In this section, we further investigate how power grid density affects the results of decap budgeting provided by our algorithm. The circuit ibm123 with a size of $2480\mu m \times 2000\mu m$ was selected, and three power grids with different densities were built. In **Case1**, the power pitches in both the x and y directions, for the lowest two metal layers in each 2D tier, are set to be the cell row height. In **Case2**, the power pitch in the y direction in these layers is set to be half of the cell row height, while that in the x direction is set to be the cell row height, and in **Case3** the power pitches in both the x and y directions in these layers are set to be half of the cell row height. In all three cases, the power pitches for the higher metal layers, as well as the number of interlayer vias connecting adjacent 3D tiers, are set proportionately.

Our decap optimization algorithm, using both CMOS and MIM decaps, was then used to individually optimize the power grids in all three cases. The results are shown in Table III. From the table, we can see that:

- First, a denser grid helps to reduce the voltage droop in a circuit. When we increase the power grid density, both the worst-case voltage droop and violation area will be reduced (see column 4 and column 5).
- Second, a denser grid implies a larger number of grid nodes, resulting in larger cost for transient analysis and adjoint sensitivity analysis. Therefore, it takes more time to solve the problem in each iteration. On the other hand, the total number of iterations decreases because the violation area in the circuit is reduced. Therefore, we can see from Table III that the total CPU time for our algorithm increases much more slowly than the power grid size.
- Third, a denser grid implies more power connections in Z direction, and therefore a higher routing congestion. This can be seen in Table III: when the power grid density increases, so do the values of the maximum and average congestion values.

VII. CONCLUSION

We have proposed an efficient decap allocation algorithm to optimize 3D power supply network using both MIM and CMOS decaps. MIM decaps have the desirable properties of high capacitance density and low leakage density, and can be a good complement to the on chip SiO_2-based CMOS decap. Our algorithm uses 3D congestion analysis and a linear congestion model, as well as linearized noise models based on adjoint sensitivity analysis, to guide the decap allocation among CMOS and MIM decaps. Experimental results show that power grid noise can be more effectively optimized using both MIM and CMOS decaps, with lower leakage power and low routing congestion costs.

REFERENCES

[1] B. Black *et al.*, "3D processing technology and its impact on iA32 microprocessors," *Proc. ICCD*, pp. 316–318, 2004.

[2] S. S. Sapatnekar, "CAD for 3D circuits: Solutions and challenges," *Proc. VMIC*, pp. 245–251, 2007.

[3] B. Goplen and S. S. Sapatnekar, "Thermal via placement in 3D ICs," *Proc. ISPD*, pp. 167–174, 2005.

[4] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs," *Proc. ICCAD*, pp. 306–313, 2004.

[5] B. Goplen and S. S. Sapatnekar, "Efficient thermal placement of standard cells in 3D ICs using a force directed approach," *Proc. ICCAD*, pp. 86–89, 2003.

[6] B. Goplen and S. S. Sapatnekar, "Placement of 3D ICs with thermal and interlayer via considerations," *Proc. DAC*, pp. 626–631, 2007.

[7] P. Zhou *et al.*, "3D-STAF: scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits," *Proc. ICCAD*, pp. 590–597, 2007.

[8] Y. Zhan, T. Zhang, and S. S. Sapatnekar, "Module assignment for pin-limited designs under the stacked-Vdd paradigm," *Proc. ICCAD*, pp. 656–659, 2007.

[9] G. Bai, S. Bobba, and I. N. Hajj, "Simulation and optimization of the power distribution network in VLSI circuits," *Proc. ICCAD*, pp. 481–486, 2000.

[10] S. S. Sapatnekar and H. Su, "Analysis and optimization of power grids," *IEEE Design & Test*, vol. 20, no. 3, pp. 7–15, 2003.

[11] L. Smith, "Decoupling capacitor calculations for CMOS circuits," *Proc. EPEP*, pp. 101–105, 1994.

[12] H. Su, S. S. Sapatnekar, and S. R. Nassif, "An algorithm for optimal decoupling capacitor sizing and placement for standard cell layouts," *Proc. ISPD*, pp. 68–73, 2002.

[13] J. Fu *et al.*, "Decoupling capacitor allocation for power delivery network noise reduction based on standard cell layouts," *Proc. ASIC*, pp. 101–104, 2003.

[14] J. Fu *et al.*, "VLSI on-chip power/ground network optimization considering decap leakage currents," *Proc. ASPDAC*, pp. 735–738, 2005.

[15] H. Chen and D. Ling, "Power supply noise analysis methodology for deep-submicron VLSI chip design," *Proc. DAC*, pp. 638–643, 1997.

[16] S. Zhao, K. Roy, and C.-K. Koh, "Decoupling capacitance allocation for power supply noise suppression," *Proc. ISPD*, pp. 66–71, 2001.

[17] J. R. Minz, S. K. Lim, and C.-K. Koh, "3D module placement for congestion and power noise reduction," *Proc. GLSVLSI*, pp. 458–461, 2005.

[18] E. Wong, J. Minz, and S. K. Lim, "Decoupling capacitor planning and sizing for noise and leakage reduction," *Proc. ICCAD*, pp. 395–400, 2006.

[19] G. Huang *et al.*, "Power delivery for 3D chip stacks: Physical modeling and design implication," *Proc. EPEP*, pp. 205–208, 2007.

[20] H. Hu *et al.*, "High performance ALD HfO_2-Al_2O_3 laminate MIM capacitors for RF and mixed signal IC applications," *Proc. IEDM*, pp. 15.6.1–15.6.4, 2003.

[21] Y. Tu *et al.*, "Characterization and comparison of high-k metal-insulator-metal (MiM) capacitors in $0.13\mu m$ Cu BEOL for mixed-mode and RF applications," *VLSI Symp.*, pp. 79–80, 2003.

[22] P. Zurcher *et al.*, "Integration of thin film MIM capacitors and resistors into copper metallization based RF-CMOS and Bi-CMOS technologies," *Proc. IEDM*, pp. 153–156, 2000.

[23] M. Armacost *et al.*, "A high reliability metal insulator metal capacitor for $0.18\mu m$ copper technology," *Proc. IEDM*, pp. 157–160, 2000.

[24] D. Roberts *et al.*, "Application of on-chip MIM decoupling capacitor for $90nm$ SOI microprocessor," *Proc. IEDM*, pp. 72–75, 2005.

[25] H. Sanchez *et al.*, "Increasing microprocessor speed by massive application of on-die high-k MIM decoupling capacitors," *Proc. ISSCC*, pp. 2190–2199, 2006.

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

[27] J. Lou *et al.*, "Estimating routing congestion using probabilistic analysis," *IEEE Trans. on Comput.-Aided Des.*, vol. 21, no. 1, pp. 32–41, 2002.

[28] "IBM-PLACE Benchmarks (version 1.0)," Available at http://er.cs.ucla.edu/benchmarks/ibm-place/.

[29] "Predictive Technology Model," Device Group at Arizona State University, Available at http://www.eas.asu.edu/~ptm.