Performance Driven Decoupling Capacitor Allocation Considering Data and Clock Interactions

Ajith Chandy and Tom Chen
Department of Electrical and Computer Engineering, Colorado State University, USA

***Abstract*— We propose a sensitivity-based method to allocate decaps incorporating leakage constraints and tighter data and clock interactions. The proposed approach attempts to allocate decaps not only based on the power grid integrity criteria, but also based on the impact of power grid noise on timing criticality and robustness. The resulting algorithm reduces the power grid noise to below a threshold and improves the performance or timing robustness of the circuit at the same time.**

I. Introduction

Traditional methods to reduce the power grid noise to below a threshold involves filling all the white spaces on chip with decoupling capacitors (decaps) and/or manually allocating decaps to the noise prone parts on the chip in an effort to keep the overall power supply noise low [1]–[4]. The conventional device for on-chip decaps is the gate capacitance of transistors. With continued CMOS scaling, the dielectric leakage of thin gate oxide devices will be significant and may become a limiting factor for its usage [5]. An alternative to gate capacitance is the metal-insulator-metal (MIM) capacitors [6], [7]. However, MIN capacitors add more complexity to CMOS process and cost. The premise of this paper is that the gate capacitance, or its variations, will continue to be used as decaps for high performance VLSI chips, and the overall trend of leakage current associated with the decaps is increasing. Therefore, we should use limited amount of decap to achieve the highest quality and performance possible under a leakage power constraint. In addition to the simple method of filling all white space on a chip with decaps, sensitivity based methods [2], [8], [9] have been proposed to minimize the power grid noise. In a sensitivity based approach, the amount of decaps is allocated based on the sensitivity of power grid noise in the form of voltage droop as a function of decaps at each location of the chip. This iterative process stops when all the available decaps are allocated, resulting in an allocation with the maximum possible improvement of power grid noise. However, after reducing the maximum power grid noise to below a given threshold, allocating more decaps to an area where no critical timing paths exist will add power consumption with little impact on performance. A better use of the remaining decaps is to allocate them based on the sensitivity of performance improvement and timing robustness.

II. Proposed Approach and Algorithm

The proposed approach consists of two phases. In the first phase, decaps are allocated to reduce the power grid noise to a specified threshold. A sensitivity based approach, similar to that in [2], is used in this phase. Once the maximum V_{DD} droop is reduced to below the threshold, the remaining decaps are allocated in the second phase. The goal of the second phase is to allocate the remaining decaps to improve the overall system performance. It uses the sensitivity of gate delay with respect to decap, $\frac{\partial(gate\ delay)}{\partial(decap)}$, to gauge the worthiness of adding decap at a given location.

The value $\frac{\partial(gate\ delay)}{\partial(decap)}$ cannot be found directly. We therefore break it down according to the chain rule as illustrated in Equation 1.

$$\frac{\partial(gate\ delay)}{\partial(decap)} = \frac{\partial(gate\ delay)}{\partial(V_{DD}\ droop)} \times \frac{\partial(V_{DD}\ droop)}{\partial(decap)} \qquad (1)$$

The first term on the R.H.S. of Equation 1 can be calculated using direct sensitivity analysis with a polynomial equation. The second term on the R.H.S. of Equation 1 is calculated using the adjoint sensitivity method [2], [8], [10], [11].

Based on the required sensitivity values described previously in this section, the proposed algorithm iteratively calculates the amount of decaps at a given location. There are three constraints considered during the decap allocation process:

- the magnitude of power grid droop at any location must be less than a pre-defined threshold.
- the total amount of decaps to be placed on a chip determined by the leakage power budget of the chip.
- the total amount of decap associated with a given grid point limited by the available space around the grid point.

Table I shows the steps in the proposed algorithm. The sensitivity values used in the algorithm is calculated at each grid point location representing it's surrounding area. Once the sensitivity at each grid point is known, the amount of decaps for each location is determined using a Linear Programming (LP) formulation.

The LP formulation considers the interaction between the clock and data signals under the influence of added decaps. The setup time and hold time constraints are strictly enforced when considering adding decaps at a given location. This is extremely important for critical max. timing paths and min. timing paths as speeding up data signals relative to clock signals and vice versa by added decaps can potentially create timing violations. Table II illustrates the LP formulation. The objective parameter, T, is the clock cycle time that is expected after the addition of decaps in the system. The first constraint of the LP is to maximize the change in clock cycle time of the most critical path. The second constraint ensures that the hold time constraint is satisfied with the added decaps.

III. Simulation Results

A power grid with 18x18 C4 bumps in a commercial 0.18μm CMOS process is used in our experiments. The die size is around 5x5mm^2. A total of four critical paths in terms of both maximum delay time and minimum delay time are

1530-1591/05 $20.00 © 2005 IEEE

TABLE I

STEPS IN THE DECAP ALLOCATION ALGORITHM

```
Do until No Decap budget remaining{
1    Simulate Original Network;
2    Extract data from output file;
3    Use data to reconstruct Adjoint Network;
4    Simulate Adjoint Network;
5    Extract data from output file;
6    Calculate sensitivity $\frac{\partial(V_{DD}\ droop)}{\partial(decap)}$
     using adjoint sensitivity method;
7    Calculate over all sensitivity $\frac{\partial(gate\ delay)}{\partial(decap)}$;
8    Input required data values;
9    Formulate the LP problem;
10   Solve the LP problem;
11   Recalculate the remaining Decap Budget;}
```

TABLE II

LP FORMULATION

Minimize : **T**

Subject to :

$\sum(S_{gi} * D_i) + \sum(S_{c1i} * D_i) - \sum(S_{c2i} * D_i) + PD_j \leq T$

$\sum(S_{c2i} * D_i) - \sum(S_{c1i} * D_i) - \sum(S_{gi} * D_i) \leq PD_j$

$j = 1 \ldots$ total number of paths

$\sum d_i \leq$ D

$d_i \leq max(d_i)$

where,

$D_i = d_i + (X_1 * \sum d_{i1}) + (X_2 * \sum d_{i2}) + ...$

observed during the process of decap allocation using the proposed algorithm. The gate switching current of all gates (on or off the 4 signal paths) are modeled using triangular current sources. The current sources are attached to their corresponding grid points. SPICE simulations are used to calculate the adjoint sensitivities.

Two different power cases resulting in two different switching current distributions are studied, where Power Case 1 has more switching activities than Power Case 2. The pre-defined threshold for power grid droop is set to 10% and 8.33% of the nominal supply voltage for Power Case 1 and 2, respectively. The first iteration of the algorithm aims at eliminating all the power grid droops greater than the pre-defined threshold. Initially there were 168 and 10 violations of the pre-defined thresholds in Power Case 1 and 2, respectively. 55nf and 4.01nf were allocated during the first iteration for Power Case 1 and 2, respectively, and all the threshold violations were eliminated after the first iteration. The remaining decaps were allocated in the subsequent iterations for performance improvement. The total decap budget for both power cases is 70nf. After 4 iterations, the potential for performance improvement measured by the sensitivity has been reduced to below a threshold. The allocated decaps are 55nf, 5nf, 5.24nf and 3nf for 4 iterations for Power Case 1; and 4.01nf, 5.25nf, 3nf, and 3nf for 4 ietrations for Power Case 2.

To determine the effectiveness of the proposed algorithm, we compare our results with the results obtained using the existing sensitivity based approach in [2]. We also examine the effectiveness of taking the clock and data interaction into account when allocating decaps for performance improvement. These simulation results are compared with the results where only data signal delays are considered during decap allocation. All simulations were performed using a test chip structure of 5x$5mm^2$ with the two power cases. Table III shows the performance improvement and the final power grid droop in term of percentage of nominal V_{DD} after the decap allocation process has terminated.

TABLE III

NORMALIZED CHANGE IN OP. FREQ. AND WORST CASE DROOP AS % OF V_{DD} FOR POWER CASES 1 AND 2

Cases	Power case 1		Power case 2	
Condition	Freq.	max droop	Freq.	max droop
Before decap addition	1	12	1	8.67
After decap addition (existing method [2])	1.0377	7	1.00257	8.14
After decap addition (data signals)	1.0477	7.8	1.00313	8.3
After decap addition (data & clock signals)	1.0505	8	1.01182	8.31

IV. CONCLUSIONS

The proposed algorithm demonstrated its ability to eliminate power grid noise violations and to achieve the maximum performance improvement possible. One may argue that the amount of performance improvement of 1.233% is small compared to delay improvements using conventional timing optimization methods. First, the proposed method is not intended to replace the conventional timing optimization methods. Rather, it is used after the timing optimization is done to further improve performance or timing robustness. Second, the amount of improvement depends very much on the power grid design, circuit's switching activities, and circuit's sensitivity to power grid droop.

REFERENCES

[1] S. Z. et. al., "Power Supply Noise Aware Floorplanning and Decoupling Capacitance Placement," *DAC*, 2002.

[2] H. S. et. al., "Optimal Decoupling Capacitor Sizing and Placement for Standard Cell Layout Designs," *IEEE TCAD*, vol. 22, 2003.

[3] H. Chen and D. Ling, "Power Supply Noise Analysis Methodology For Deep-submicron VLSI Chip," *DAC97*, 1997.

[4] H. Chen and S. Schuster, "On-chip Decoupling Capacitor Optimization for High-performance VLSI Design," *Proc. Tech. Papers and Int. Symp. on VLSI Tech., Sys., and Appl.*, pp. 99–103, 1995.

[5] S. B. et. al., "IC Power Distribution Challenges," *ICCAD01*, 2001.

[6] S.-J. Ding and H. Hu, "High-performance mim capacitor using ald high-k $HfO_2 - Al_2O_3$ laminate dielectrics," *IEEE Electron Device Letters*, vol. 24, 2003.

[7] T. Ishikawa and D. Kodama, "High-capacitance $Cu/Ta_2/O_5/Cu$ MIM structure for SoC applications featuring a single-mask add-on process," *Int. Electron Devices Meeting Tech. Digest*, pp. 940–942, 2002.

[8] H. Su, K. Gala, and S. Sapatnekar, "Fast Analysis and Optimization of Power/Ground Networks," *ICCAD00*, 2000.

[9] G. B. et. al., "Simulation and Optimization of the Power Distribution Network in VLSI circuits," *ICCAD00*, 2000.

[10] L. T. Pillage, R. A. Rohrer, and C. Visweswariah, *Electronic Circuit and System Simulation Methods*. McGraw-Hill, 1995.

[11] S. W. Director and R. A. Roher, "The Generalized Adjoint Network and Network Sensitivities," *IEEE Trans. Circuit Theory*, vol. 16, pp. 318–323, 1969.