Timing-Aware Decoupling Capacitance Allocation in Power Distribution Networks

Sanjay Pant, David Blaauw

University of Michigan, Ann Arbor

Abstract

Power supply noise increases the circuit delay, which may lead to performance failure of a design. Decoupling capacitance (decap) addition is effective in reducing the power supply noise, thus making the supply network more robust in presence of large switching currents. Traditionally, decaps have been allocated in order to minimize the worst-case voltage drop occurring in the power grid. In this paper, we propose an approach for timing-aware decap allocation which uses global time slacks to drive the decap optimization. Non-critical gates with larger timing slacks can tolerate a relatively higher supply voltage drop as compared to the gates on the critical paths. The decap allocation is formulated as a non-linear optimization problem using Lagrangian relaxation, and modified adjoint method is used to efficiently obtain the sensitivities of objective function to decap sizes. A fast path-based heuristic is also implemented and compared with the global optimization formulation. The two approaches have been implemented and tested on ISCAS85 benchmark circuits and with grids of different sizes. Compared to uniformly allocated decaps, the proposed approach utilizes 35.5% less total decap to meet the same delay target. For the same total decap budget, the proposed approach is shown to improve the circuit delay by 10.1% on an average.

1 Introduction

Power supply networks are essential in providing the devices on a die with a reliable and constant operating voltage. Due to on-chip and package interconnect parasitics, the supply voltage delivered to devices on a die exhibits both spatial and temporal fluctuations. With technology scaling and decreasing nominal supply voltage, gate delay is becoming increasingly sensitive to supply variations as the headroom between the supply voltage V_{dd} and device threshold voltage V_{th} is consistently getting reduced [1]. Therefore, it is extremely important to model the impact of power supply noise on circuit performance and improve the robustness of the power delivery network from the aspect of circuit timing.

Capacitance between the power and ground distribution networks, commonly referred to as decap, provides local charge storage and is helpful in mitigating the voltage drop in the presence of rapidly switching current transients. Parasitic capacitance between metal lines of the power distribution grid, device capacitance of the non-switching transistors and N-Well substrate capacitance occur naturally in a power distribution network and act as implicit decoupling capacitance. Unfortunately, the amount of the naturally occurring decoupling capacitance is not sufficient to meet stringent power supply integrity constraints and designers have to often add substantial amount of explicit decoupling capacitance on the die at various strategic locations.

Gate capacitance of *n* or *p* type devices is normally used as the explicit decap. These explicitly added decaps not only result in area overhead, but also increase the leakage power consumption of the chip due to their gate leakage current. With technology scaling, gate leakage has become a significant percentage of the overall leakage power and has been cited as a significant limitation on the maximum amount of decap that can be introduced [2]. Hence, the goal of the designers is to meet the desired performance and signal integrity constraints with the least possible total amount of explicitly added decaps.

A number of methods have been proposed to allocate explicit decap in order to confine the voltage drops in the power grid within a pre-specified bound. The decap allocation problem was formulated as a non-linear optimization problem in [3][4][5] with constraints on the worst-case voltage drops. An adjoint sensitivity based method [9] was used in [4][5] to obtain the sensitivities of decaps to the power supply noise metric. Decap allocation methods tend to be computationally intensive because of expensive transient power grid and adjoint grid simulations. The method in [5] proposes a partitioning-based approach to reduce the power grid simulation runtime during decap allocation.

All the above mentioned approaches aim at restricting the voltage drop at all the supply nodes within a pre-specified margin for a given total decap budget. However, in high performance designs, circuit performance is a more pressing concern and the above approaches, although optimal for supply noise reduction, may not be optimal for optimizing circuit performance. For instance, in a logic block, only the delay of gates on the critical and near-critical paths are of concern and the gates having larger timing slacks can afford relatively higher voltage drops. To address this issue, two recent approaches [6][7] have been proposed for timing aware reduction of power supply noise. The approach in [6] requires enumeration of all critical paths and then formulates the problem as a non-linear optimization problem. However, this approach is computationally intensive, requiring many adjoint circuit simulations (equal to the number of gates in the enumerated paths) during each optimization iteration. The approach in [7] uses a *prediction* and *correction* based algorithm for power noise reduction. In the *prediction* step, the amount of decap at various locations is predicted based on switching frequency and placement of standard cells. The *correction* step involves gate sizing to improve timing after placement. However, this approach is heuristic-based and may lead to over-design in certain scenarios.

In this paper, we propose an approach that is based on global static timing analysis of a circuit and does not require the enumeration of circuit paths. The approach naturally incorporates timing slacks at the gates in the decap allocation algorithm. The objective of the proposed approach is to improve the overall circuit performance, given a total decap budget. Rather than confining the voltage drop at *all* the nodes in a power grid within a pre-specified bound, the proposed approach automatically reduces the voltage drop near the timing critical regions while non-critical gates may have relatively higher voltage drops. Arrival time constraints are handled using Lagrangian relaxation and the relaxed subproblem is solved efficiently using the modified adjoint sensitivity method. We show how the sensitivities of the node voltages with respect to the decap sizes can be computed with a single adjoint simulation, significantly improving the runtime of the algorithm. We also propose

1-4244-0630-7/07/$20.00 ©2007 IEEE.

a fast and accurate heuristic to allocate decaps for improving circuit performance. We implemented the proposed approaches and tested them on benchmark circuits and power grids of different sizes. We demonstrate that, on average, the proposed approach improves the timing delay by 10.1% for a given decap budget or the amount of required decap by 35.5% for a given delay target, compared to uniform decap allocation. The results also show that the heuristic approach results in slightly lesser delay/ decap reduction with significantly smaller runtimes.

The remainder of the paper is organized as follows. Section 2 presents the global decap minimization formulation using Lagrangian relaxation. Section 3 presents the proposed path-based heuristic approach. Experimental results are presented in Section 4 and conclusions are drawn in Section 5.

2 Proposed Global Optimization Approach

A typical power supply network model of a chip consists of ideal supply voltage sources, power and ground wires modeled as a linear RLC network, time varying current sources representing switching transistors, and decoupling capacitances [8]. Each logic block in the design is simulated at nominal supply voltage to obtain the current drawn by the blocks over time. Each block of non-linear devices is then replaced by its time-varying current distributed among its power grid supply points. Figure 1 shows a combinational logic circuit in a power grid consisting of two metal layers. For clarity, only the *Vdd* distribution grid is shown. The ideal voltage sources, time-varying currents and decaps are not shown in the figure.

The objective of the proposed optimization problem is to allocate decaps such that the circuit delay in each clock cycle is less than a pre-specified clock period, T. Conversely, the optimization can be formulated to minimize the decap allocation while meeting a specified delay constraint. The optimization variables are the decap sizes C_n attached to power grid nodes n. The proposed approach is a post-placement method and we assume that the white space available for decap insertion is known. Furthermore, the placement information of gates is assumed to be available. Since the voltage variations in a power grid are typically very slow compared to the transition time of a switching gate [10], we make the simplifying assumption that the supply voltages are constant during the switching transition of a gate.

The next subsection describes the problem formulation for the global timing aware decap allocation.

2.1 Gate Delay Model

Consider a combinational circuit comprising of g gates with the power and ground supplies at a gate i denoted as Vdd_i and Vss_i respectively. Two fictitious nodes *src* and *sink* are added to the circuit. All the primary inputs are connected to the *src* node and all the primary outputs are joined together to form the *sink* node. The *sink* node connecting all the POs is labeled as node 0 and all other gates are numbered in the reverse topological order. Let the arrival time at the output of a gate i be denoted by a_i. Let *input*(i) be the set of indices of gates driving the inputs of gate i and let *output*(i) be the set of indices of gates in the fanout of gate i. For example in Figure 2, g=4, *input*(0)={1,2}, *output*(3)={1,2}.

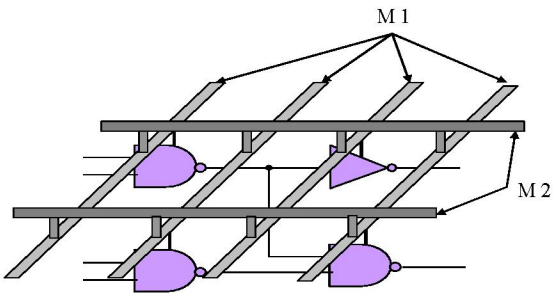

Figure 1. A combinational circuit in a power distribution network

The delay of a gate i from one of its inputs j, D_{ji}, and output transition time, tr_{ji}, are represented as a linear function of the voltage drops in supply voltages at the gate i and the driver gate driving the node j. This linear approximation have been shown to be accurate for power supply variations within a range of ±10% in [11].

$$D_{ji} = D_{ji}^0 + k_{ji}\Delta Vdd_i + l_{ji}\Delta Vss_i + m_{ji}\Delta Vdd_j + n_{ji}\Delta Vss_j \quad (1)$$

$$tr_{ji} = tr_{ji}^0 + p_{ji}\Delta Vdd_i + q_{ji}\Delta Vss_i + r_{ji}\Delta Vdd_j + s_{ji}\Delta Vss_j \quad (2)$$
$$\forall j \in input(i)$$

where, D_{ji}^0 is the delay of gate i from input j under ideal supply voltages; tr_{ji}^0 is the transition time at the output of gate i under ideal supply voltages; ΔVdd_i and ΔVss_i are power voltage drop and ground bounce respectively at gate i; k, l, m, n, p, q, r and s are constants obtained by simulating the gate delay over a range of supply voltages and performing multi-variable linear regression.

Traditional standard cell libraries are composed of two dimensional tables with table entries representing delay and transition times for different load-transition time combinations. For the proposed algorithm, the cell library is re-characterized for different input slope-output load combinations to also incorporate the delay coefficients k through s, along with the nominal delay and transition time entries, D_{ji}^0 and tr_{ji}^0. For a given set of transition time, load capacitance and supply voltages at a gate and its driver, the delay is computed appropriately by table look-up and interpolation for delay coefficients and using equations (1) and (2).

In the next section, we state the primal optimization problem (*PP*) for decap allocation under timing constraints, followed by the Lagrangian relaxation formulation. The discussions are focussed on the decap minimization problem with constraints on circuit delay. The problem of circuit timing improvement with constraints on total decap can be solved in a similar manner.

2.2 Primal Problem

The variables in the primal problem are the decap sizes C_ns and arrival times at the output of gates, a_is. The objective of the primal problem (*PP*) is to minimize the total decap area:

$$\text{minimize} \quad \sum_{n=0}^{N-1} C_n \quad (3)$$

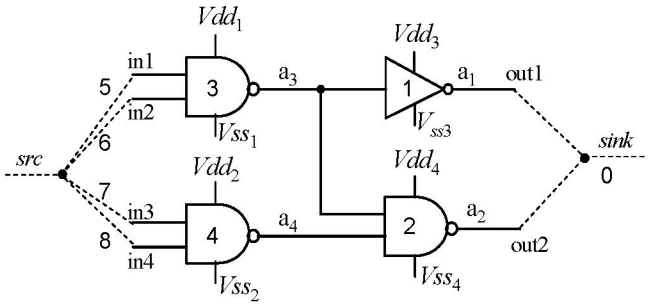

Figure 2. A combinational circuit under non-ideal supply voltages

where, N is the total number of decap candidate locations. The maximum decap sizes are bounded based on the white space available at each candidate decap location, n:

$$0 \le C_n \le Cmax_n \tag{4}$$

The constraint on arrival times, a_is, are stated as follows (For simplicity in explanation, but without loss of generality, we do not differentiate between rise and fall transitions):

$$a_j \le T \qquad \forall\, j \in input(0)$$

$$a_j + D_{ji} \le a_i \qquad \forall\, j \in input(i) \wedge (1 \le i \le g) \tag{5}$$

where, T is the pre-specified delay requirement of the given circuit.

The delay of gates, D_{ji}, are expressed as a linear function of supply voltages as shown in (1). The supply voltages, on the other hand, are a function of decap sizes, and are given by the following modified nodal analysis (MNA) relation [12]:

$$\left[G + \frac{C}{h}\right] x[k] \;=\; I[k] + \frac{C}{h} x[k-1] \tag{6}$$

where, x is the vector of unknowns: node voltages, inductor currents and currents from voltage sources; I is the vector of current and voltage sources; G and C are the conductance and capacitance matrices of the power grid; k is the simulation time; and h is the time step for simulation.

The primal problem is difficult to solve in the current form because of the large number of unknowns and constraints in the problem. Also, it requires a prohibtively large number of circuit simulations to compute the sensitivity of the objective function (3) to decap sizes. In the next subsection, we describe the use of Lagrangian relaxation to remove the constraints and present a formulation that reduces the number of required simulations.

2.3 Lagrangian Relaxation Problem

Lagrangian relaxation is a standard technique to eliminate difficult constraints in an optimization problem [13]. For our purpose, a non-negative Lagrange multiplier, λ_{ji}, is associated with each input-output pair (j,i) for gate i, and the corresponding arrival time constraint is incorporated into the objective function. For a given set of Lagrange multipliers, the problem in Section 2.1 can be expressed as the following Lagrangian relaxation problem:

minimize

$$\sum_{n=0}^{N-1} C_n + \sum_{j \in input(0)} \lambda_{j0}(a_j - T) + \sum_{i=1}^{g} \sum_{j \in input(i)} \lambda_{ji}(a_j + D_{ji} - a_i)$$

subject to:

$$0 \le C_n \le Cmax_n$$

$$D_{ji} = D_{ji}^0 + k_{ji}\Delta Vdd_i + l_{ji}\Delta Vss_i + m_{ji}\Delta Vdd_j + n_{ji}\Delta Vss_j \qquad \forall j \in input(i)$$

$$\left[G + \frac{C}{h}\right] x[k] \;=\; i[k] + \frac{C}{h} x[k-1]$$

For a given set of Lagrange multipliers λ, the above problem has two sets of variables: arrival times, a and decap sizes C. Kuhn-Tucker conditions [13] state that if (a^*, C^*) is at the optimal solution to the above problem, then the derivative of the objective function with respect to all the variables must be zero:

$$\frac{\partial\, objective}{\partial a_i}\bigg|_{a = a^*,\, C = C^*} = 0$$

and hence,

$$\sum_{j \in input(i)} \lambda_{ji} \;=\; \sum_{k \in output(i)} \lambda_{ik} \qquad \forall i \tag{7}$$

This condition states that at the optimal solution, the sum of Lagrange multipliers from the inputs of a node i is equal to the sum of Lagrange multipliers emanating from node i to all its fanout gates. Let Ω_λ be the set of Lagrange multipliers satisfying the above condition in (7). Thus, if we search the Lagrange multipliers only in the set Ω_λ, we can eliminate arrival time variables a_i from the objective function. The simplified objective function for such choices of Lagrange multipliers is expressed below:

$$\text{obj:} \quad \sum_{n=0}^{N-1} C_n - \sum_{j \in input(0)} \lambda_{j0} T \;+\; \sum_{i=1}^{g} \sum_{j \in input(i)} \lambda_{ji} D_{ji} \tag{8}$$

Substituting the expression for D_{ji} from (1) into (8), we get the objective function as follows:

$$\sum_{n=0}^{N-1} C_n - \sum_{j \in input(0)} \lambda_{j0} T \;+\; \sum_{i=1}^{g} \sum_{j \in input(i)} \lambda_{ji} D_{ji}^0 + \tag{9}$$

$$\sum_{i=1}^{g} \alpha_i \cdot \Delta Vdd_i + \sum_{i=1}^{g} \beta_i \cdot \Delta Vss_i$$

where,

$$\alpha_i \;=\; \sum_{j \in input(i)} \lambda_{ji} \cdot k_{ji} + \sum_{k \in output(i)} \lambda_{ik} \cdot m_{ik} \tag{10}$$

and $$\beta_i \;=\; \sum_{j \in input(i)} \lambda_{ji} \cdot l_{ji} + \sum_{k \in output(i)} \lambda_{ik} \cdot n_{ik} \tag{11}$$

Thus, given the optimal value of λ, we can solve the above simplified Lagrangian relaxation problem to arrive at the optimal solution. With the simplified objective function in (9), removal of arrival time variables and arrival time constraints, the Lagrangian problem is much easier to solve as compared to the Primal Problem. The next subsection describes the solution of the Lagrangian problem for a given set of Lagrange multipliers.

2.4 Solving the Lagrangian Relaxation

To solve the Lagrangian subproblem described in the previous subsection, we need to compute the sensitivities of the objective function expressed in (9) to decaps C_n. Adjoint sensitivity [9] is the preferred method when sensitivity of one measurement response to multiple parameters is required. On the other hand, direct sensitivity method is efficient in computing the sensitivity of multiple measurement responses to a single parameter. In our case, as shown in objective function in (9), we require the sensitivity of worst supply voltage values at each gate, to changes in all the decap sizes. Thus, using either the adjoint or the direct method in its standard form will require multiple power grid simulations. For instance, using the direct sensitivity method would require N transient power simulations, where N is the number of decaps. On the other hand, the adjoint sensitivity method would require one original power grid simulation and g adjoint network simulations, where g is the number of gates in the combinational circuit. We therefore use a modified adjoint sensitivity method [14] which uses the principle of

superposition and computes the sensitivities of multiple measurement variables to multiple circuit parameters in only one adjoint simulation. Since, the objective function in (9) is a linear function of voltage drops at all the gate locations, the principle of superposition holds and the modified adjoint sensitivity method can be used to solve the Lagrangian relaxation problem, as explained in the next subsection.

2.4.1 Sensitivity Computation

We first describe the sensitivity computation of worst voltage drop, ΔVdd_i, at a single gate location i. The power grid is simulated with a given current profile and the derivative of the voltage waveform across all the capacitors, $\dot{V}_n(t)$ is stored. The time point of the occurrence of the worst voltage drop at the supply node of gate i is observed. Let τ_i denote the time of occurrence of the worst drop at gate i. Then, the adjoint power grid network is constructed with all the voltage sources shorted to ground and all the current sources removed from the network. The adjoint network is excited backwards in time with unit delta, $\delta(t-\tau_i)$ current waveform applied at supply node of gate i. The voltage waveform at all the supply nodes of all the decaps, denoted by $\Psi_n(t)$, is then observed and stored. Lastly, the convolutions between $\dot{V}_n(t)$ and $\Psi_n(t)$ provide the sensitivity of worst case voltage drop at the gate i to all the decap sizes. Since there are g gates in the design, this method will require g different adjoint grid simulations, one for each gate, to compute the sensitivity of the voltage drops at all the gates.

However, the objective function of Lagrangian relaxation subproblem in (9) consists of a linear combination of voltage drops ΔVdd_i and ΔVss_i, weighted by constants α_i and β_i. As mentioned in the last subsection, the principle of superposition is therefore used and all the current excitations are applied simultaneously to the adjoint circuit. Let $[\tau]$ be a gx1 vector (g is the total number of gates), representing the time of occurrence of worst drop at all the gates. In the adjoint network, for every gate i, a current source, represented by a scaled delta function, $\alpha_i\delta(t-\tau_i)$ for power grid (and $\beta_i\delta(t-\tau_i)$ for ground grid) is applied and time-varying voltage responses at the decaps $\Psi_n(t)$ are observed. Then, the convolutions between the derivative of voltages across the decaps in the original grid and voltages across the decaps in the adjoint grid provide the required sensitivities. Thus, the sensitivity of (9) to C_i is given as follows:

$$\frac{\partial\, objective}{\partial C_n} = 1 + \int_0^T \psi_n(T-t)\cdot \dot{V}_n(t)\, dt \qquad (12)$$

where, T is the duration of original grid simulation; n is the power grid node where decap C_n is connected; $\Psi_n(t)$ is the voltage waveform at node n in the adjoint network and $V_n(t)$ is the voltage waveform at node n in the original network.

The gradients of the objective function to decap values obtained in this manner can be used to solve the Lagrangian relaxation problem using the conjugate gradient method [13].

2.5 Finding the Optimal λ

The previous subsection described the modified adjoint method to find the optimal solution given a set of Lagrange multipliers λ. In this subsection, we present a method for obtaining the optimal set of λs based on timing slacks available in the circuit. The Lagrange multiplier λ_{ji} intuitively represents the criticality of the delay of gate

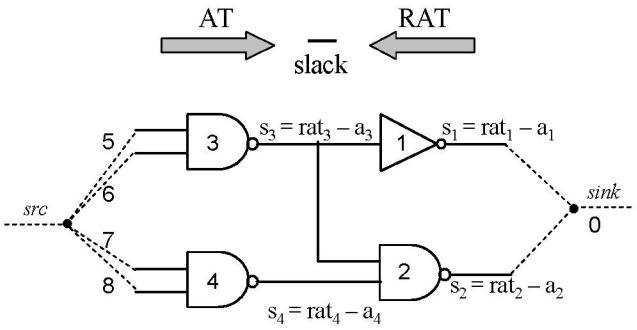

Figure 3. Slack computation using required arrival times

i from its input j. If the circuit consists of only one prominent critical path under supply variations, then at the optimal solution point, only the Lagrange multipliers along the critical path will remain non-zero, while all other λs associated with the non-critical paths will be zero. We propose to update the set of Lagrange multipliers based on the gate criticality or in other words, the timing slack available at each node.

At the start of the algorithm, we assume that every gate delay is critical and as an initial guess, all the Lagrange multipliers are non-zero and satisfy condition (7). Then the Lagrangian relaxation subproblem is solved optimally as described in Section 2.4 based on the given set of Lagrange multipliers. In the next iteration, timing slack is computed at each node in the circuit based on new arrival times (AT) and new required arrival times (RAT) as shown in Figure 3. Each Lagrange multiplier λ_{ji} is decreased in proportion to the slack available at the output of gate i as follows:

$$\lambda_{ji}^{k+1} = \lambda_{ji}^{k} - \rho_k \cdot s_i^k \qquad \forall 1 \le i \le g, j \in input(i) \qquad (13)$$

where, λ_{ji}^{k} and s_i^k is the value of Lagrange multiplier in iteration k; s_i^k is the slack at the output of gate j in iteration k; ρ_k is the step-size in iteration k.

2.6 Overall Optimization Flow

The overall optimization flow for the algorithm is presented in Figure 4. We start with all Lagrange multipliers as non-zero such that the condition in (7) is satisfied. Then, the Lagrangian relaxation

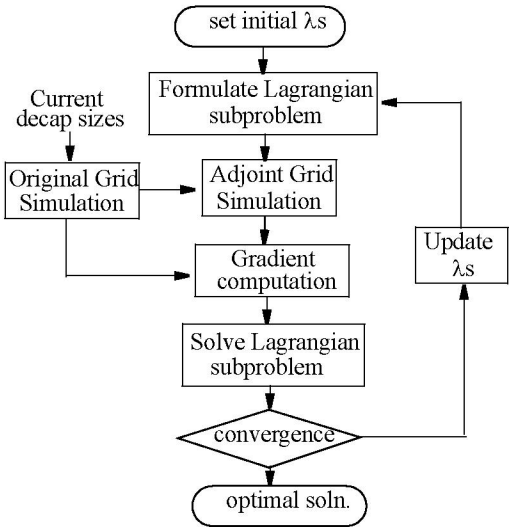

Figure 4. Overall optimization flow

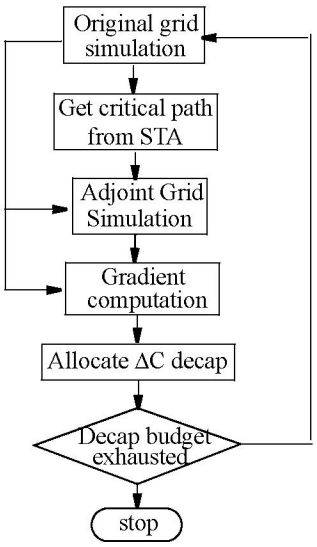

Figure 5. Optimization flow of the path based greedy algorithm

problem is formulated and values of voltage drop coefficients α_i and β_i are obtained. The voltage drop coefficients are used to provide current excitations to the adjoint network. Convolution is performed between the original and adjoint network waveforms to obtain the sensitivities of the objective in the Lagrangian subproblem. Using these sensitivities, the Lagrangian subproblem is solved optimally for the given set of Lagrange multipliers. Lagrange multipliers are then updated based on the slack available at gates in the combinational circuit and the procedure is repeated until the convergence of Lagrange multipliers.

3 Greedy Path-Based Approach

The Lagrangian subproblem described in the prior subsections involves updating the Lagrange multipliers in each major iteration and solving the Lagrange relaxation subproblem within each such iteration. In this subsection, we present a path based greedy heuristic to reduce the number of iterations. This approach is fast and has been found to give a favorable trade-off in terms of run time and optimization quality compared to the optimal results.

The optimization flow for the greedy approach is illustrated in Figure 5. At the start of the optimization, all the decaps are assigned a small value. Power grid is simulated with these decap values to obtain the worst voltage drops at all the gates in the circuit. Static

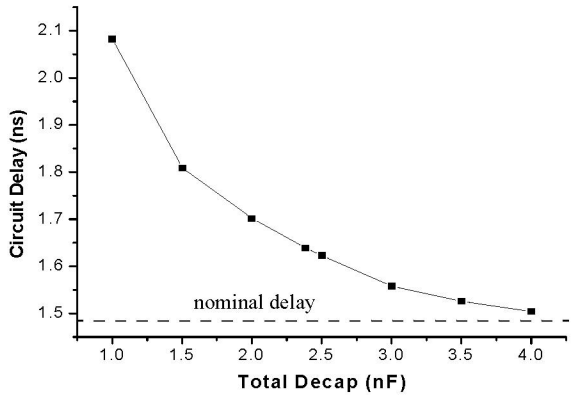

Figure 6. Decap area vs circuit delay tradeoff curve for c432

timing analysis is performed on the circuit with the voltage drops to obtain the path with maximum delay. In the Lagrangian relaxation based algorithm, this amounts to setting all the Lagrange multipliers to 1 in the critical path and 0 in the non-critical paths. Path delay is formulated as a linear function of voltage drops using (1). Gradient of the path delay to all the decaps is then computed using the results from original and adjoint grid simulations. Based on the gradients, a small amount of decap ΔC is allocated in the best search direction. The new allocated capacitance values are stamped in the MNA coefficient matrix (G+C/h) and original grid is re-simulated to obtain the new worst voltage drops at the gates. STA is again performed to obtain the critical path in the circuit which may have changed due to decap insertion in the previous iteration. The process is repeated until the decap budget has been exhausted. In every iteration, only a small amount of decap is available to the optimizer for allocation.

4 Experimental Results

The proposed Lagrangian relaxation-based and heuristic decap allocation algorithms were implemented in C++ and tested on ISCAS85 benchmark circuits with power grids of different sizes. For the experiments, two power grids, Grid1 and Grid2, consisting of 4 metal layers were constructed using pitches and widths of an industrial microprocessor in 0.13µ technology. Grid1 is a M2-M5 power grid of 700µx700µ die-area consisting of 10,804 nodes, 17,468 elements, 12 *Vdd* C4s and 12 *Vss* C4s. Grid2 is a M2-M5 1.2mmx1.2mm grid with 17,530 nodes, 29,746 elements, 28 *Vdd* C4s and 28 *Vss* C4s. To model package impedance, an inductance

grid	ckt	# gates	# decaps	decap budget	Circuit Delay				% delay reduction	runtimes	
					nom.	uniform decaps	global optim.	greedy optim.		global optim.	greedy optim.
Grid1	c432	212	476	2.38nF	1.498ns	1.798ns	1.621ns	1.640ns	9.84%	11m15s	1m15s
Grid1	c499	553	595	2.98nF	1.233ns	1.480ns	1.308ns	1.394ns	11.62%	9m41s	1m57s
Grid1	c1355	654	793	3.97nF	1.839ns	2.207ns	1.878ns	1.913ns	14.90%	11m43s	2m58s
Grid1	c1908	543	579	2.89nF	2.088ns	2.506ns	2.251ns	2.256ns	10.17%	20m24s	25.83s
Grid1	c2670	1043	1190	3.57nF	1.622ns	1.946ns	1.754ns	1.764ns	9.86%	52m33s	8m41s
Grid2	c3540	1492	1559	7.79nF	2.301ns	2.761ns	2.498ns	2.564ns	9.52%	109m59s	23m49s
Grid2	c5315	2002	2217	6.65nF	2.080ns	2.769ns	2.409ns	2.416ns	9.97%	221m24s	61m18s
Grid2	c6288	3595	3712	8.15nF	5.186ns	6.223ns	-	5.820ns	6.48%	>4hrs	188m36s
Grid2	c7552	2360	2571	7.18nF	2.975ns	3.571ns	-	3.262ns	8.65%	>4hrs	63m03s

Table 1. Experimental results showing delay reduction for a given decap budget

grid	ckt	nom. delay	delay constr.	Decap Allocated		%decap redn.
				uniform	optim.	
Grid1	c432	1.498ns	1.640ns	3.55nF	2.38nF	32.98%
Grid1	c499	1.233ns	1.394ns	3.49nF	2.98nF	17.60%
Grid1	c1355	1.839ns	1.913ns	6.65nF	3.97nF	40.33%
Grid1	c1908	2.088ns	2.256ns	6.15nF	2.89nF	52.92%
Grid2	c2670	1.622ns	1.764ns	6.96nF	3.57nF	95.80%
Grid2	c3540	2.301ns	2.564ns	10.04nF	7.80nF	22.37%
Grid2	c5315	2.080ns	2.416ns	12.20nF	6.65nF	45.56%
Grid2	c6288	5.186ns	5.820ns	9.74nF	8.15nF	16.31%
Grid2	c7552	2.978ns	3.262ns	13.26nF	7.18nF	45.85%

Table 2. Experimental results showing decap reduction for a timing constraint

of 1nH was connected in series with a resistance of 0.1mΩ at each C4 location in the grid. The benchmark circuits were synthesized using 0.13μ standard-cell library and placed using Cadence Silicon Ensemble. Some of the white space available was chosen to be the candidate for decap allocation and the decap in each candidate location was modeled as a capacitor connected to the power grid nodes over the region. For the gates in the modeled power grid region, the peak currents were approximated using a triangular waveform of 1ns duration. The gate peak currents were scaled to cause an appreciable (15%) worst-case voltage drop in the grid. Preconditioned conjugate gradient method was applied for optimization, using the LANCELOT non-linear solver [15].

Figure 6 shows the delay-decap trade-off curve for circuit c432 placed in test-grid Grid1. The dotted line represents the nominal delay of the circuit under ideal supply voltages. The figure shows the effectiveness of the proposed timing-aware approach in improving the circuit delay. Table 1 presents the improvement in circuit delay for a fixed decap budget. Column 1 and 2 show the circuit name and its power grid. Column 3 and 4 show the number of gates in the circuit and number of candidate decap locations in the layout. The total decap budget is shown in column 5. Columns 6, 7, 8 and 9 show the circuit delay under ideal voltage supplies, delay with uniform decap distribution, delay obtained after Lagrangian optimization and delay after application of the heuristic, respectively. It should be noted that the uniform decap allocation, Lagrangian optimization and greedy heuristic utilize the same total decap budget. The greedy heuristic-based approach gave near-optimal results for all the circuits. Column 10 shows the percentage improvement in circuit delay by Lagrangian optimization over uniform decap allocation. On an average, the circuit delay was observed to improve by 10.11% for the given decap budget. The last two columns show the runtimes of the Lagrangian-based optimization and the greedy heuristic. The Lagrangian-based optimization, although optimal, had considerably larger runtimes compared to the greedy algorithm. The Lagrangian-based optimization could not converge within 4 hours of runtime for two largest circuits. Runtimes can be improved by using a better non-linear solver and by accelerating the transient power grid simulations using moment-based solvers [16].

Table 2 presents the comparison of total decap budget required to meet a pre-specified timing budget using the heuristic method. Column 3 states the circuit delay with ideal supply. The constraint on circuit delay under power supply fluctuations is shown in column 4. Columns 5 and 6 show the amount of total decap required to meet the given timing constraint under uniform distribution of decaps and the decaps allocated by the proposed heuristic approach. Column 6 shows the percentage reduction in total decap area compared to uniform decap distribution, which is 35.51% on an average.

5 Conclusion

In this paper, we proposed an approach for timing aware decoupling capacitance allocation which utilizes the timing slacks available at the gates in a design. The decoupling capacitance allocation is formulated as a non-linear optimization problem and Lagrange relaxation in conjunction with the modified adjoint method is used for optimization. We also presented a fast and near-optimal greedy heuristic for timing-aware decap allocation. The approach has been implemented and tested on ISCAS85 benchmark circuits and power grids of different sizes. Compared to uniformly allocated decaps, the proposed approach utilizes 35.51% less total decap to meet the same delay target. For the same total decap budget, the proposed approach is shown to improve the circuit delay by 10.11%.

References

[1] Semiconductor Industry Association. International Technology Roadmap for Semiconductors, 2004.

[2] S. Bobba, T. Thorp, K. Aingaran and D. Lu, "IC power distribution challenges," in *Proceedings of International Conference on Computer Aided Design*, 2001, pp. 643-650.

[3] S. Zhao, K. Roy and C.K. Koh, "Decoupling capacitance allocation and its application to power-supply noise-aware floorplanning," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, issue 1, Jan. 2003, pp. 81-92.

[4] H. Su, S.S. Sapatnekar and S.R. Nassif, "Optimal decoupling capacitor sizing and placement for standard-cell layout designs", in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, issue 4, April 2003, pp. 428-436.

[5] H. Li, Z. Qi, S.X.D. Tan, L. Wu, Y. Cai and X. Hong, "Partitioning-based approach to fast on-chip decap budgeting and optimization," in *Proceedings of Design Automation Conference,* 2005, pp.170-175.

[6] A. Chandy and T. Chen, "Performance driven decoupling capacitor allocation considering data and clock interactions," in *Proceedings of Design, Automation and Test in Europe,* 2005, pp. 984-985.

[7] C.Y. Yeh and M.M. Sadowska, "Timing-aware power noise reduction in layout," in *Proceedings of International Conference on Computer Aided Design*, 2005, pp. 627-634.

[8] G. Steele, D. Overhauser, S. Rochel and Z. Hussain, "Full-chip verification methods for DSM power distribution systems," in *Proceedings of Design Automation Conference*, 1998, pp. 744-749.

[9] S. Director and R. Rohrer, "The generalized adjoint network and network sensitivities," in *IEEE Transactions on Circuits and Systems*, vol. 16, issue 3, Aug. 1969, pp. 318-323.

[10] A. Chandrakasan, W. J. Bowhill and F. Fox, *Design of high performance microprocessor circuits*. NY: IEEE Press, 2001.

[11] S. Pant, D. Blaauw, V. Zolotov, S. Sundareswaran and R. Panda, "Vectorless analysis of supply noise induced delay variation," in *Proceedings of International Conference on Computer Aided Design*, 2003, pp. 184-191.

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

[13] M.S. Bazaraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, second edition 1993.

[14] A.R. Conn, R.A. Haring, C. Visweswariah and C.W.Wu, "Circuit optimization via adjoint Lagrangians," in *Proceedings of International Conference on Computer Aided Design*, 1997, pp. 281-288.

[15] A.R. Conn, N.I.M. Gould and P.H. Toint, *LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization*, Springer Verlag, 1992.

[16] T.H. Chen, C. Luk, C.C.P. Chen, "INDUCTWISE: inductance-wise interconnect simulator and extractor", in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, issue 7, July 2003, pp. 884-894.