Power/Ground Supply Network Optimization for Power-Gating

Hailin Jiang
ECE Department, UCSB
hailin@ece.ucsb.edu

Malgorzata Marek-Sadowska
ECE Department, UCSB
mms@ece.ucsb.edu

***Abstract* -- Power-gating is a technique for efficiently reducing leakage power by shutting off the idle blocks. However, the presence of power-gating may also introduce negative effects on power supply network, which have not been considered in the earlier design stages. Ignoring those effects may result in suboptimal power supply network designs and could potentially even nullify the intended power savings. In this paper, we analyze mutual dependencies between the sleep transistors and the P/G network, and we present a general flow to optimize the P/G supply network for power-gating. Experimental results show that sizing sleep transistor and power network separately cannot achieve optimal solution in terms of power. By compromising only 1% of the total area, our optimization method allows us to save 10% of power dissipated on decaps and sleep transistors, which is a practical solution for a power-gated system. We also report results of a study on optimal solutions for various gated areas and current densities.**

***Index Terms* -- Power gating, power/ground network, decap, sleep transistor, optimization.**

I. INTRODUCTION

Due to continuous shrinking of the minimal feature size, lowering of supply voltages, and lowering of threshold voltages, leakage power is emerging as a major challenge for current and future CMOS designs. Power-gating is an efficient technique for reducing leakage power by shutting off the idle blocks. Implementation of power-gating requires a multi-threshold CMOS process. Logic blocks are implemented using low-V_t, high-performance transistors, whereas high-V_t transistors (also called sleep transistors) connect the gated blocks to the power supply [6].

However, the power-gating is not without cost. In [4], the authors discuss and quantify the negative effects of power gating, which include noise, performance penalty, area and power overhead, etc. Recently, research efforts have been made in the physical design stage to reduce some of these negative effects. In [1], the authors propose a novel gating structure in which the sleep transistors are turned on in a non-uniform, step-wise manner to reduce the magnitude of the voltage fluctuations. In [5], the authors propose two clustering algorithms based on bin-packing and set-partitioning techniques for centralized sleep-transistor sizing, which improve the chip area and performance.

However, all those works have been done assuming that the power supply network had already been designed. There is no mention that the global power supply grid requires any special considerations related to power gating. Designing the global grid giving no consideration to power gating causes two problems. First, introducing power-gating to an existing grid causes a new power grid noise originating from the switching of sleep transistors. At the same time, gated macros shield the decap effect from the gated circuit and hinder the decap insertion in that area. Thus the power supply network should be appropriately redesigned to be immune to the gating noise. Second, the traditional P/G network optimization performs wire sizing and decoupling capacitor insertion [7][9] targeting at minimizing the total area. However, when power-gating is implemented, sleep transistors will consume power and more decaps will cause more dissipated leakage power, especially in more advanced technologies. If the power delivery network and sleep transistors are designed in such a way that their extra costs nullify the power savings, the power-gating technique becomes useless. Because of the mutual dependencies between the sleep transistors and the P/G network, we have to consider them together to achieve an optimal solution in terms of power. In this work, we present a general algorithm to optimize the power/ground supply network for power-gating. To the best of our knowledge, our work is the first one that addresses this problem.

The remainder of the paper is organized as follows. In section II, we point out the problems introduced by power-gating in the context of P/G network. In section III, we address the co-design of the sleep transistors and of the P/G network. In section IV, we present our method for optimizing the P/G network for power-gating. In section V we show the experimental results. Section VI summarizes our conclusions and suggests practical uses for the results.

II. ANALYSIS OF P/G NETWORK DESIGN WITH POWER-GATING

Traditionally, the on-chip power distribution network is constructed hierarchically. Power is delivered from the top-level metal layer (which is connected to the package), down through the internal layers, and finally to the active device. The entire chip is fully powered at all times. In a power-gating system, a circuit is divided into several power domains. Each domain is independently controlled by a sleep-transistor. The sleep transistors are turned off to reduce the dynamic power and leakage power in a standby mode, and turned on in an active mode for normal operation.

Power/ground network is traditionally optimized by wire width sizing and decap insertion, with an objective of minimizing the total area under the IR drop and electrical migration constraints. However, when power-gating is applied on a power supply network, the new technique will introduce extra effects which did not exist before but which should be considered now.

First, the sleep transistor needs to be carefully designed. Its size must be sufficient to guarantee that the current drawn by a macro does not cause an excessive voltage drop on a sleep transistor which is turned on. However, when large-size sleep transistors are turned on and off, considerable power supply currents are drawn in a short period of time, which may result in large voltage fluctuations on the power grid. This *power-gating noise (PGN)* affects the reliability of systems-on-chips (SoCs), especially for advanced CMOS technologies with narrow noise margins. We need to add extra decap to reduce the *PGN* and take into account the additional leakage currents associated with the intentionally introduced capacitors.

Second, in the power-gating system, the gated macros are disconnected from the power supply in their idle state, which prevents them from acting as decaps for the active circuits. Thus the non-gated macros adjacent to the gated circuitry may need extra decap to guarantee that power grid noise remains within tolerable margins. The gated macros not only stop providing decap to other macros, but also act as an obstacle for decap insertion. Extra decap will enlarge the chip size and again consume more leakage power.

So without taking these effects into account, the power supply network will not be robust enough to tolerate noise and needs to be redesigned. Now, the question is, should we design a P/G network with power-gating in one shot? Is the co-design of sleep transistors and P/G network necessary? In the following section, we will discuss the mutual dependencies between the sleep transistor design and P/G network design to answer these questions.

III. CAN THE SLEEP TRANSISTOR AND P/G NETWORK DESIGN BE DECOUPLED?

The most straightforward approach to design a robust P/G network with power-gating is to size the sleep transistors first, and then optimize the P/G network. This approach works to achieve the minimal total area consumed by the sleep transistors, decaps, and power mesh. This is so because the smaller sleep transistors will cause less noise, thus requiring fewer resources to counteract it. All the sizings go in the same direction. We can first minimize the total area of sleep transistor under the performance constraint (i.e. the maximal voltage drop across on the sleep transistor), then optimize the P/G network by wire-sizing and insert decap based on the given size of sleep transistors.

However, the purpose of introducing power-gating is to save power. If the cost of power overhead caused by power gating is greater than the savings it provides, power-gating is useless. So the task we face is how to minimize the power consumption caused by power-gating. We have observed a mutual dependency in terms of power between the sleep transistors' sizes and decap. When we decrease the sleep transistors' sizes, the switching noise on the power grid decreases, which calls for less decap and in turn less dissipated leakage power. However, smaller size transistors have larger resistance, which increases the IR drop across them, which increases the power consumed. So the simple approach outlined at the beginning of this section cannot guarantee the minimal power consumption. Optimizing the area would make no sense because it could erode the potential power saving. Moreover, because of the mutual dependency, the designs of sleep transistors and P/G networks cannot be decoupled. We need to find the trade-off between sleep transistors and decap to achieve the optimal solution in terms of power.

In this work, we present a general flow to optimize a P/G network with power-gating. Through this flow, we can achieve an optimal solution in terms of either area or power consumption. More generally, by adjusting the weighting parameters, we can determine the best trade-off between area and power, depending on design requirements.

IV. P/G NETWORK OPTIMIZATION CONSIDERING POWER-GATING

A. Modeling

a) Power-gating.

The power-gating model is shown in Fig.1. The upper power mesh can be modeled as a linear resistive network excited by constant voltage sources (we ignore the package inductance in this work). The lower power mesh is composed of a linear RC network excited by time-varying current sources which represent the switching currents drawn by the logic circuits. We use a simple triangular waveform to model the current profiles. C denotes decap from the on-chip non-switching capacitors C_{ckt} and/or purposely inserted oxide capacitors C_{ox}.

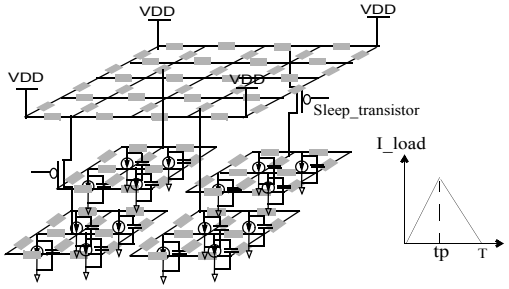


Fig. 1. Power gating modeling

b) Sleep transistor.

The model we use for the sleep transistor depends on its region of operation, see Fig.2. When the sleep transistor is switching, it works in the saturation region. In the saturation stage, the sleep transistor can be modeled as a current source with a triangle waveform. The peak current is expressed by Sakurai's alpha-power-law [8] formula:

$$I_D = \frac{\mu_p C_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS}-V_{th})^{\alpha} \quad (1)$$

where α is the velocity saturation index for modeling the short channel effects. It is close to one in modern deep submicron devices. We define the switching period as twice the falling time of the signal V_G: $t_2 - t_1 = t_1 - t_0$. (Refer to Fig.2 for definitions of t_0, t_1 and t_2.) When a gated macro is active, the sleep transistor which controls it works in the linear region and can be modeled as a resistor:

$$I_{ST} \approx \mu_p C_{ox}\left(\frac{W}{L}\right)(V_{dd} - V_{tH})V_{DS} \quad (2)$$

$$R_{ST} = \frac{1}{\mu_p C_{ox}(V_{dd} - V_{tH})}\left(\frac{L}{W}\right) \quad (3)$$

where I_{ST} is the switching current of the logic block, V_{tH} is the threshold voltage of the sleep transistor, and R_{ST} is the channel resistance of the sleep transistor in the linear operation region.

We know that voltage drop on a sleep transistor will degrade the chip's performance. We define the allowable performance loss (PL) as:

$$\delta = 1 - T_d / T_d^{ST} \quad (4)$$

where T_d is the delay of the circuit when sleep transistors are absent, and T_d^{ST} is the delay of the circuit when sleep transistors are present. From [5] we know that:

$$MAX(V_{ST}) = \delta(V_{dd} - V_{tL}) \quad (5)$$

$$MIN\left(\frac{W}{L}\right)_{ST} = \frac{I_{MAX}}{\delta\mu_p C_{ox}(V_{dd} - V_{tL})(V_{dd} - V_{tH})} \quad (6)$$

where I_{MAX} is the peak current of I_{ST} and V_{tL} is the threshold voltage of the gate in logic blocks. This formula indicates the lower bound of sleep transistor size.

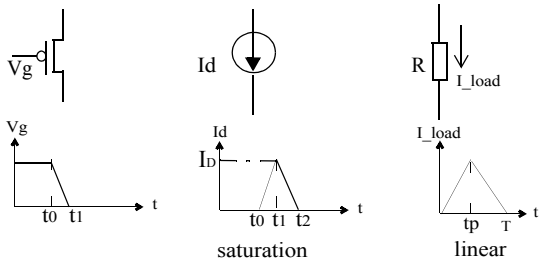


Fig. 2. Sleep transistor modeling

The model of sleep transistor in a saturation region will be used for noise analysis, and the model of sleep transistor in a linear region will be used for power estimation in our objective function. Note that the power consumption should also include the dynamic power consumed by the sleep transistors during the switching period. However, if the gating period is long enough, that part of power is relatively small and can be ignored in the objective function.

c) Power grid noise

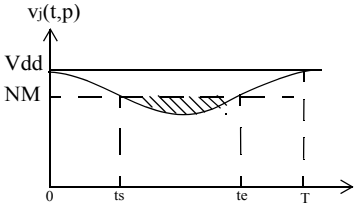


Fig. 3. Noise at typical grid node

To estimate the power grid induced noise at a node, we use the following efficient formulation [7]:

$$z_j(p) = \int_0^T max\{NM - v_j(t,p), 0\}\,dt = \int_{t_s}^{t_e}\{NM - v_j(t,p)\}\,dt \quad (7)$$

where p is the tunable circuit parameter and NM is the noise margin. This integral expression represents the shaded area in Fig.3. For the whole circuit, we have

$$Z = \sum_{j=1}^{N} z_j(p) \quad (8)$$

where N is the number of nodes whose noise goes beyond the margin. This metric is a summation of the integrals of the noise violations, which is zero if all constraints are satisfied.

B. Problem formulation

The optimization of P/G network with power-gating can be formulated as:

Objective: $MIN[\alpha \cdot Area + (1-\alpha)\cdot\beta\cdot Power]$

where $$Area = \sum_{i=1}^{I} l_{Ti} w_{Ti} + \sum_{j=1}^{J} A_{Cj} + \sum_{k=1}^{K} l_{Rk} w_{Rk}$$

$$Power = \sum_{i=1}^{I} (I_i^{max})^2 \cdot R_{Ti} + \sum_{j=1}^{J} P_{unit} \cdot A_{Cj}$$

Subject to: $$w_{Ti}^{min} \le w_{Ti} \le w_{Ti}^{max} \qquad i = 1, 2 \ldots I$$

$$0 \le A_{Cj} \le A_{Cj}^{max} \qquad j = 1, 2 \ldots J$$

$$w_{Rk}^{min} \le w_{Rk} \le w_{Rk}^{max} \qquad k = 1, 2 \ldots K$$

$$\text{and} \quad Z(w) < \varepsilon \quad (9)$$

where β is a scaling factor to make the two terms in the objective function of the same magnitude. w_{Ti} denotes width of the i-th transistor, w_{Rk} is the width of the k-th wire in the grid, and A_{Cj} is the area of the j-th inserted decap C_{ox}. These are the variables we need to adjust. l_T is the length of a sleep transistor and l_R denotes the length of a wire segment. w_{Ti}^{min} is the lower bound of the size on an i-th sleep transistor, which is determined based on Eq.(6). Other bounds are determined by the design. r_T is the resistance of a sleep transistor, which is defined in Eq.(3). I^{max} is the peak current of the current drawn from the gated blocks and P_{unit} is the leakage power per unit area of decap. Z is defined in Eq.(8) and ε is a very small number.

α is a weight factor to trade off between area and power; its range is between 0 and 1. If we set $\alpha = 1$, we simultaneously size the wire segments, decap, and sleep transistors to achieve the minimal area under noise constraint. As we discussed in section III, the result should be the same as that obtained from sizing the sleep transistor first, followed by P/G network optimization. If we set $\alpha = 0$, we can size the sleep transistor and decap to achieve the optimal power for the initial power mesh. α can be adjusted depending on design requirement. Experiments show that the grid power consumption constitutes only 5% of the combined decaps and sleep transistors power consumption, so we do not include the grid power in the formula.

This is a non-linear problem with non-linear objective and constraint functions. We can apply a standard sequence-of-quadratic-programs (SQP) [3] to solve this optimization problem. The calculation of objective function and its derivatives with respect to each decision variable is straightforward. The evaluation of constraint function Z can be done by HSPICE simulation. We use adjoint sensitivity analysis [2] to calculate the derivatives of Z with respect to varying parameters. In the following sub-section we discuss this point in more detail.

C. Sensitivity calculation

The adjoint-network-based sensitivity analysis is a standard technique for circuit optimization used when sensitivities of the cost function with respect to many parameter values are required. In [7] the authors use this method to find the optimal solution for decap sizing and placement.

The basic idea of the adjoint sensitivity analysis is as follows. By applying Tellegen's theorem, we have

$$\int_0^T \sum_V [\hat{v}_V \delta i_V - \hat{i}_V \delta v_V]\, dt + \int_0^T \sum_I [\hat{v}_I \delta i_I - \hat{i}_I \delta v_I]\, dt = \int_0^T \sum_{NS} [\hat{i}_{NS} \delta v_{NS} - \hat{v}_{NS} \delta i_{NS}]\, dt \tag{10}$$

where v and i represent the vectors of branch voltages and currents in the original network N, and where $\hat{v}$ and $\hat{i}$ represent the branch voltages and currents in the adjoint network $\hat{N}$. The networks N and $\hat{N}$ have the same topology; both satisfy the Kirhchoff's current (KCL) and voltage (KVL) laws. V denotes the terms corresponding to voltage sources; I denotes the terms corresponding to the current sources; NS denotes the terms corresponding to branches with no sources.

All the interesting performance metrics (power, delay, voltage, etc.) can be expressed with the following general formula:

$$\Phi = \int_0^T f(v_I, i_V)\, dt \tag{11}$$

whose variation is:

$$\delta\Phi = \int_0^T \left[\sum_V \frac{\partial f}{\partial v_I} \delta v_I + \sum_I \frac{\partial f}{\partial i_V} \delta i_V \right] dt \tag{12}$$

So, for a given performance function Φ, the excitations $\hat{v}_V$ and $\hat{i}_I$ in adjoint network can be chosen so as to make the left-hand side of Eq.(10) equal to the performance variation $\delta\Phi$. Since the choice of the element types in the adjoint circuit is arbitrary, they can be chosen so as to cancel the terms depending on δi_{NS} or δv_{NS} in the right-hand side of Eq.(10) leaving only the terms depending on parameter variation.

In the original network we attach a zero valued current source to those nodes which do not tap to any source (either voltage or current). Z is then the function of v_I:

$$Z = \sum_{j=1}^{N} \int_{t_s}^{t_e} \{NM - v_I^j(t)\}\, dt \tag{13}$$

Then the variation of Z can be expressed as:

$$\delta Z = \sum_{j=1}^{N} \int_{t_s}^{t_e} \frac{\partial f}{\partial v_{Ij}} \delta v_{Ij}\, dt = \sum_{j=1}^{N} [u(t - t_e) - u(t - t_s)] \cdot \delta v_{Ij} \tag{14}$$

We assign the adjoint circuit excitations such that:

$$\hat{i}_I = \frac{\partial f}{\partial v_I} = u(t - t_f) - u(t - t_0), \quad \hat{v}_V = \frac{\partial f}{\partial i_V} = 0$$

Then the Eq.(10) is rewritten as:

$$-\delta Z + \int_0^T \sum_I \hat{v}_I \delta i_I\, dt - \int_0^T \sum_V \hat{i}_V \delta v_V\, dt = \int_0^T \sum_{NS} [\hat{i}_{NS} \delta v_{NS} - \hat{v}_{NS} \delta i_{NS}]\, dt \tag{15}$$

The voltage sources are constant, so $\delta v_V = 0$. Unlike in [7], we need to adjust the sleep transistor which is modeled as a current source in the original network. That means i_I is a varying parameter and $\delta i_I \neq 0$. Eq.(15) becomes:

$$\delta Z = \int_0^T \sum_{NS} [\hat{v}_N \delta i_N - \hat{i}_N \delta v_N]\, dt + \int_0^T \sum_I \hat{v}_I \delta i_I\, dt \tag{16}$$

From [2], we know that if the corresponding R(C) branches of the adjoint and original networks are the same, the first term on the right-hand side of Eq.(16) depends only on parameter variation, which can be stated as:

$$\delta Z = \sum_G \left[\int_0^T \hat{v}_G(\tau) v_G(t)\, dt \right] \cdot \delta G + \sum_C \left[\int_0^T \hat{v}_C(\tau) \dot{v}_C(t)\, dt \right] \cdot \delta C + \sum_I \left[\int_0^T \hat{v}_I dt \right] \cdot \delta i_I \tag{17}$$

We have: $G = \frac{w_R}{\rho \cdot l_R}$, $C = C_0 + \frac{\varepsilon_{ox}}{T_{ox}} \cdot A_C$ and

$$I = \frac{\mu_p C_{ox}}{2} \left(\frac{w_T}{l_T} \right) (V_{Gs} - V_{th}) = \lambda \cdot w_T$$

where C_0 is the capacitance of the on-chip decap. By applying chain rules, we finally get:

$$\delta Z = \sum_G \left[\int_0^T \hat{v}_G(\tau) v_G(t)\, dt \right] \cdot \frac{1}{\rho \cdot l_R} \cdot \delta w_R + \sum_C \left[\int_0^T \hat{v}_C(\tau) \dot{v}_C(t)\, dt \right] \cdot \frac{\varepsilon_{ox}}{T_{ox}} \cdot \delta A_C + \sum_I \left[\int_0^T \hat{v}_I dt \right] \cdot \lambda \cdot \delta w_T \tag{18}$$

D. Optimization flow

The complete P/G network with power-gating optimization flow is summarized in Fig.4. First we set up a floorplanned chip with the initial power grid, define the gated

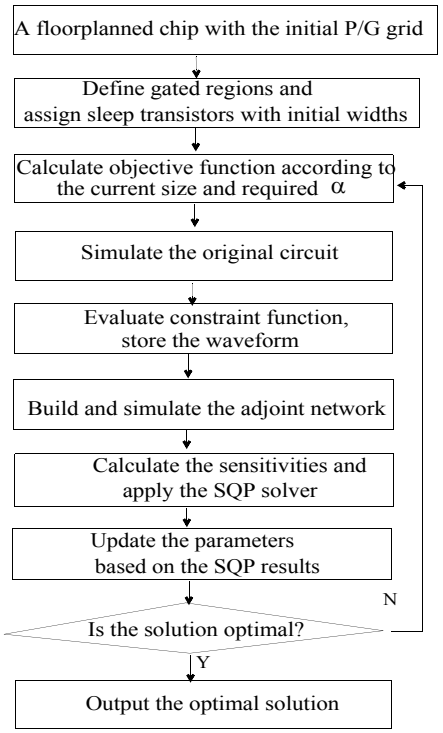


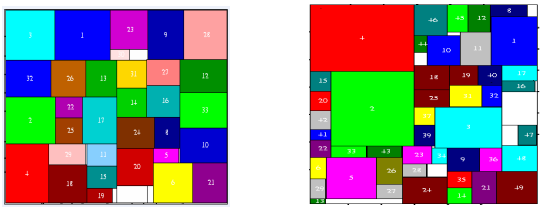
Fig. 4. Optimizing P/G network for power-gating

regions, and assign the initial sizes of sleep transistors to them. Then we apply SQP solver to size the wire segments, decap, and sleep transistor simultaneously to achieve the optimal solution according to the objective function. The sensitivity calculation is done by the adjoint method described in section IV. It must be repeated in every iteration.

V. EXPERIMENTS

We integrated C++ with TCL to develop our optimization flow. HSPICE was used to simulate the circuit and SOL/QPSOL[3] was evoked to solve SQP. We ran all the experiments on a P4 2.4GHz PC with Linux. The test circuits were implemented in $0.13\mu m$ technology with $Vdd = 1.2$. The temperature was $25°C$.

We used circuits from MCNC and ISCAS benchmarks as the basic logic blocks. The switching current profiles and on-chip decap are extracted by simulating each circuit. We mapped those blocks to MCNC floorplan benchmarks *ami33* and *ami49*, respectively. The initial floorplanned chips are shown in Fig.5.



(a) ami33 (b) ami49

Fig. 5. Initial floorplanned chips

We first optimized the power network without power-gating and then applied the power-gating on the circuits in such a way that the gated area occupied half of the chip, consisting of the top-left and bottom-right quadrants. Table.1 shows the effect of power-gating. The first two columns list the data of chip and grid size, respectively. Power mesh reported in this experiment is global power mesh which is mainly affected by power-gating, because sleep transistors are connected to global power mesh. After applying power-gating, there were 96(219) violated nodes in non-gated region which would affect the non-gated circuits. The worst case voltage V_{min} was 0.93(0.72). So we concluded that not considering power-gating could lead to a wrong decision in designing a power supply network.

Table 1. Effect of power-gating

	chip_size (um^2)	grid_size	# of violated nodes	Z (v.ns)	V_{min} (v)
ami33	1232x936	30x30	96	0.58	0.93
ami49	1723x1748	40x40	219	10.07	0.72

Then we applied our flow to optimize the power network with power-gating. We first set $\alpha = 1$, which implies that we target the optimal solution only in terms of area(OptA). Then, based on this optimal solution, we set $\alpha = 0$ to determine the optimized solution in terms of power(OptP). We compared our results with the traditional two-step method (sleep transistor sized first, followed by power network optimization).

Table 2. Comparison of optimizations

		power mesh area	decap&sleep-trans.		V_{min}	# of iters	run time(s)
			area	power (mw)			
ami33	Trad	98739	16476	15.8	1.082	23	581
	OptA	98739	16476	15.8	1.080	26	625
	OptP	98862	22647	14.0	1.081	27	627
ami49	Trad	210631	62323	39.2	1.078	33	2139
	OptA	210631	62323	39.2	1.081	39	2725
	OptP	211064	78389	35.7	1.079	39	2869
cost & saving			0.6%	10%			

The comparison results are shown in Table.2. The area of power-mesh, the area of decaps and sleep transistors, as well as the power consumed by decap and sleep transistor, are reported. From this table, we note that we achieve the same results by applying the traditional method and by applying OptA, which confirms that separately optimizing sleep transistor and power networks will achieve the optimal solution in terms of area. However, it's not the optimal solution in terms of power. Comparing the data in the fifth column, we note that around 10% power saving from decaps and sleep transistors achieved by applying our optimization flow. Because the leakage power is dominant in advanced technologies, the power saving from decap has a significant impact on the total chip power consumption. The area overhead is about 0.6% in chip area, which is negligible. We also notice that the power mesh does not change much comparing OptP with OptA. After the power network with power-gating is re-optimized, worst case voltage is within or close to margin, as shown in the sixth column. Column seven lists the iterations to achieve the optimal solution. The last column lists run time. Although we employed HSPICE inside the optimization flow, run time is still affordable because the global power mesh has relative coarse grids and few nodes.

In the second experiment, we studied the impact of various power-gating schedules on optimization results. We applied four power-gating schedules on each benchmark:

- *g1*: gated area occupied 30% of the chip;
- *g2*: gated area occupied 70% of the chip;
- *g3*: gated area occupied half of the chip and gated blocks have higher current density;
- *g4*: gated area occupied half of the chip and gated blocks have lower current density.

Fig. 6. Different gating schedules

Fig.6 illustrates the gating schedules respectively. Gated circuits are marked grey. In *g3* and *g4*, we purposely re-floorplanned the chip in such a way that circuits with

Table 3. Optimizations vs. power-gating schedules

		before opt V_{min}	OptA			OptP				
			area (um^2)	power (mw)	run_time (s)	area (um^2)	power (mw)	run_time (s)	power saving%	area cost (%)
ami33	g1	1.03	13835	11.2	476	18447	10.3	497	8	0.4
	g2	0.89	26675	20.7	997	37052	19.2	1020	7	0.9
	g3	0.80	37225	28.2	1023	48756	26.0	1151	8	1.0
	g4	0.96	15108	13.8	798	23180	12.3	786	11	0.7
ami49	g1	0.72	59107	33.2	2070	74166	30.5	2113	8	0.5
	g2	0.58	98136	47.5	3070	128254	44.2	3127	7	1.0
	g3	0.51	112790	55.2	3154	145919	51.3	3018	7	1.1
	g4	0.70	51764	31.7	2349	69836	28.8	2413	9	0.6

higher current density are placed in the left half of the chip and circuits with lower current density are placed in the right half of the chip.

We applied our optimization flow on each chip with various power-gating schedules and compared OptA with OptP, in terms of area and power consumed by decap and sleep transistors. Table.3 shows the results. From the table, we note that in both gating schedules *g1* and *g2*, our optimization flow did not save as much power as in the first experiment (refer to Table.1). We explain that result as follows: If the gated area is small, the sleep transistor is small, so adjusting its area does not save much power. When the gated area is large, the sleep transistor is also large, and so the cost of power on the large amount of extra decap nullifies the saving on the sleep transistor. So, our optimization flow favors power-gating with comparable areas of gated and non-gated circuits. However, we note that for the two corner cases, the power consumed by sleep transistor and decap might be equal or exceed the power saving from power-gating, in which case power-gating is not practical.

Comparing the data for *g3* and *g4*, we notice that the advantage of OptP is diminished when gated circuits have higher current densities. This behavior is consistent with the sleep transistor's size increase when current density increases. A small change of sleep transistor will require a large amount of decap to ameliorate the noise. More decap consumes more leakage power, which makes the power saving on sleep transistor relatively smaller. So our optimization flow is more effective when power-gating is applied on circuits with smaller current densities. Again, we note that in case of large current density of the gated circuits, the amount of power consumed by the sleep transistor and decap will be larger, which may nullify the saving from power-gating. In such a case, we would not choose to apply the power-gating technique.

VI. CONCLUSION

In this paper, we have studied the effects on P/G network introduced by a power-gating technique and analyzed mutual dependencies between the sleep transistor and the P/G network. We presented a general flow to optimize the P/G supply network for power-gating in terms of both power and area. Experimental results show that sizing a sleep transistor and a power network separately cannot achieve an optimal solution in terms of power. By compromising only 1% of the total area, our optimization method allows us to save 10% of power dissipated on decaps and sleep transistors. This offers a practical solution for a power-gated system. Our optimization flow is more effective when gated circuits are of smaller current density, and when the areas of gated and non-gated circuit are comparable. We believe that our results could be used as reference by designers working on power-gating systems.

ACKNOWLEDGEMENT

This work was supported in part by SRC grant # 1069 and in part by MICRO through IBM. We gratefully acknowledge equipment grant from Intel.

REFERENCES

[1] K.Choi, Y.Xu and T.Sakura, "An effective yet feasible power-gating scheme achieving two orders of magnitude lower standby leakage", *Symposium on VLSI Circuits*, pp. 312-315, 2005.

[2] P.Feldmann, T.V.Nguyen and et al, "Sensitivity computation in piecewise approximate circuit simulation", *IEEE Trans on Computer Aided Design*, pp. 171-183, 1991.

[3] P.E.Grill, W.Murray, and et al, User's Guide for SOL/OPSOL: A Fortran Package for Quadratic Programming. Stanford, CA. Dept. Oper. Res.

[4] H.Jiang, M.Marek-Sadowsks and S.R.Nassif, "Benefits and costs of power-gating technique", *International Conference on Computer Design,* pp 559-566, 2005

[5] C.Long and L.He, "Distributed Sleep Transistor Network for Power Reduction", IEEE Trans of VLSI, pp. 937-946, 2004

[6] S.Mutoh and et al, "A 1V power supply high-speed digital circuit technology with multi threshold voltage CMOS", *IEEE Journal of Solid-State Circuit,* pp. 847-854, 1995.

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

[8] T.Sakurai and A.Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas" *IEEE Journal of Solid-State Circuits,* vol25, pp. 584-594, 1990.

[9] K.Wang and M.Marek-Sadowska, "On-chip power-supply network optimization using multigrid-based technique", *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems,* pp 407-417, 2005.