Title

Decoupling capacitance design strategies for power delivery networks with power gating

Authors

Xu, Tong
Li, Peng
Sundareswaran, Savithri

Publication Date

2015

Peer reviewed

Decoupling Capacitance Design Strategies for Power Delivery Networks with Power Gating

TONG XU and PENG LI, Texas A&M University
SAVITHRI SUNDARESWARAN, Freescale Semiconductor

Power gating is a widely used leakage power saving strategy in modern chip designs. However, power gating introduces unique power integrity issues and trade-offs between switching and rush current (wake-up) supply noises. At the same time, the amount of power saving intrinsically trades off with power integrity. In addition, these trade-offs significantly vary with supply voltage. In this article, we propose systemic decoupling capacitors (decaps) optimization strategies that optimally trade-off between power integrity and leakage saving. Specially, new global decap and reroutable decap design concepts are proposed to relax the tight interaction between power integrity and leakage saving of power gated PDNs with a single supply voltage level. Furthermore, we propose a flexible decap allocation technique to deal with the design trade-offs under multiple supply voltage levels. The proposed strategies are implemented in an automatic design flow for choosing the optimal amount of local decaps, global decaps and reroutable decaps. The conducted experiments demonstrate that leakage saving can be increased significantly compared with the conventional PDN design approach with a single supply voltage level using the proposed techniques without jeopardizing power integrity. For PDN designs operating at two supply voltage levels, the optimal performance is achieved at each voltage level.

Categories and Subject Descriptors: B.7.1 [**Intesrated Circuits**]: Types and Design Styles

General Terms: Design

Additional Key Words and Phrases: Power gating, power delivery network, on-chip decaps

ACM Reference Format:
Tong Xu, Peng Li, and Savithri Sundareswaran. 2015. Decoupling capacitance design strategies for power delivery networks with power gating. ACM Trans. Des. Autom. Electron. Syst. 20, 3, Article 38 (June 2015), 30 pages.
DOI: http://dx.doi.org/10.1145/2700825

1. INTRODUCTION

Leakage power consumption has become a significant challenge for nanometer VLSI circuit designs. For example, the percentage of chips that is idle or significantly underclocked (dark silicon) increases as the process scales down [Taylor 2012; Esmaeilzadeh et al. 2012]. Dark silicon is estimated to take up 20% of the chip area at the 22nm technology node and it will take up 50% at the 8nm node [ITRS 2013]. To this end, controlling chip leakage power becomes increasingly important for modern IC designs for which power gating is an effective power management solution [Hu et al. 2004; Leverich et al. 2009; Chen et al. 2010; Intel 2008, 2013].

Power delivery networks have been the focus of a large body of research work dealing with efficient simulation and design [Kozhaya et al. 2002; Zeng et al. 2011; Feng and Li 2008; Lai et al. 2012]. A typical power-gated power delivery network (PDN) is shown

This material is based on work supported by the National Science Foundation under Grant No. 0747423.
Author's address: T. Xu; email: xutong85@gmail.com.

DOI: http://dx.doi.org/10.1145/2700825

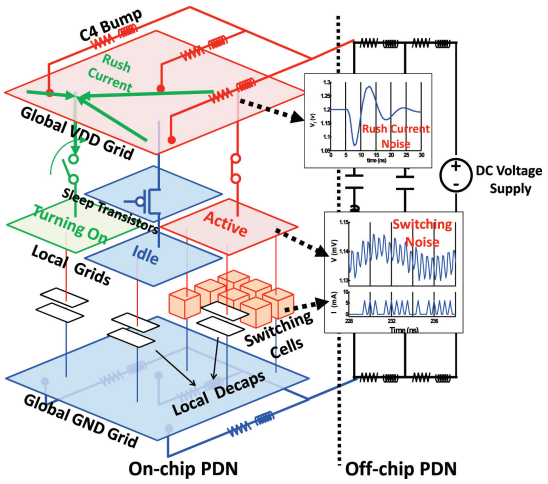


Fig. 1. Typical structure and supply noises of power-gated PDNs. The switching noise is due to switching currents of logic devices. The rush current noise is due to rush currents created to charge up the decaps of a local grid that is woken up.

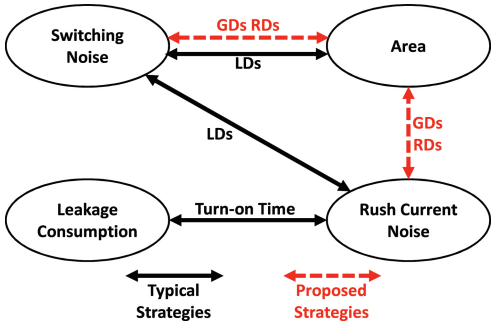


Fig. 2. Design trade-offs and strategies of power-gated PDNs with a single supply voltage. The oval-shapes indicate the concerns of a PDN design. The edges indicate the strategies to balance design concerns. Black solid edges are the typical strategies. Red dash edges are the strategies proposed in this article.

in Figure 1. The PDN is composed of an off-chip package and on-chip power grids. The on-chip part includes a global VDD grid, a global GND grid and multiple local power-gated grids (power domains). Each local power grid is connected to the global V_{DD} grid through switchable sleep transistors. Hence, they can be turned off to save leakage power. Such power delivery networks have been widely adopted in multi-core chips and SoCs to support the power gating of multiple power domains [Intel 2008, 2013].

Power integrity is a significant concern in power-gated PDN designs. Two types of supply noises exist in the power-gated PDNs: switching noise and rush current (wake-up) noise as shown in Figure 1. The first type is caused by switching activities of logic cells. When time variant switching currents flow through off-chip inductors and on-chip resistive grids, a voltage fluctuation is introduced to the circuit. The second one is due to currents that are created to charge up the decoupling capacitors in a local grid when it is woken up. The rush current noise is a unique source of supply noise for power-gated PDNs.

The primary design challenge of a power-gated PDN stems from the conflicting objectives of power integrity and power efficiency. We summarize the key design trade-offs and contrast between typical design strategies with the new strategies we exploit in

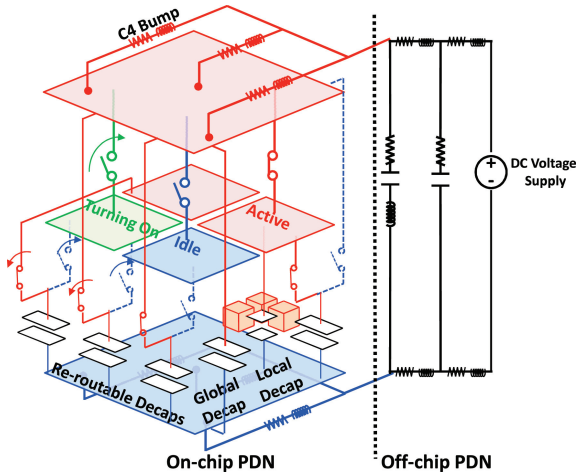


Fig. 3. Proposed structure of power-gated PDNs.

this article in Figure 2. The oval shapes of the diagram indicate the design concerns and the edges indicate the typical strategies (black solid edges) or the proposed strategies (red dash edges). Switching noise is typically suppressed by local decaps (LDs) that are connected between the local grids and the global GND grid [Su et al. 2003; Zhao et al. 2002; Jiang et al. 2005]. In this case, the suppressions of switching noise and rush current noise contradict each other since local decaps are the sources of rush current noise. Hence, it is hard to achieve the power integrity by only using local decaps. Extending the turn-on time of the sleep local grid is a common strategy to suppress the rush current noise [Kim et al. 2003; Agarwal et al. 2006; Kawasaki et al. 2008]. However, longer turn-on time inevitably reduces leakage saving, for there are fewer opportunities to launch power gating. As a result, the leakage saving of power gating is limited by the power integrity requirements.

Some existing works propose solutions to deal with the problem. Multiple sleep modes with different sleep depths have been proposed by Agarwal et al. [2006], and Singh et al. [2007]. Each sleep mode represents a trade-off between wake-up penalty and leakage saving through controlling the steady state potential in the sleep mode. Although the turn-on time of light sleep modes is shortened, the leakage saving of these modes is reduced correspondingly. The bypass power line and multi-size sleep transistors are used by Kawasaki et al. [2008]. But it is not very economic for core-level power gating since additional global power networks is required to implement the bypass power line.

In this article, we employ both global decaps (GDs) and reroutable decaps (RDs) to relax the tight interaction between power integrity and leakage saving as shown in Figure 3. Global decaps are allocated between global VDD and GND grids. They are mainly used to suppress the rush current noise by providing parts of charge required by local decaps. Reroutable decoupling is a recent design concept that was introduced in our recent work [Xu et al. 2011]. A reroutable decap is connected to the local grid and the global VDD grid via two switches. Reroutable decaps can work as local decaps or global decaps through controlling the switches. With reroutable decaps and global decaps, both switching noise and rush current noise can be suppressed without sacrificing the leakage saving. We also discuss the allocation problem of reroutable decaps. Different decap allocations are proposed in some existing works. The on-chip decaps are distributed allocated among switching cells [Zhao et al. 2002, 2007]. An optimization method is proposed in Su et al. [2003] to cluster the on-chip decaps and allocate them

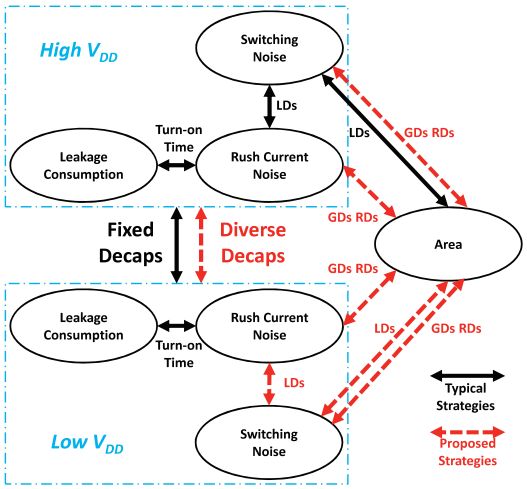


Fig. 4. Design trade-offs and strategies of a power-gated PDN with two supply voltages. The oval-shapes indicate the concerns of a PDN design. The edges indicate the strategies to balance the design concerns. Black solid edges are the typical strategies. Red dash edges are the strategies proposed in this article.

adjacent to logic blocks. However, the decaps discussed in most of these works are local decaps. Allocation of local decaps is improved in order to better suppress switching noise. However, reroutable decaps have two functions: switching noise suppression and rush current noise suppression. In this article, we propose the allocation of reroutable decaps with the considerations of these two functions.

Another important problem of power-gated PDN design has to do with supply voltage scaling. Dynamic Voltage Scaling (DVS) and Dynamic Voltage and Frequency Scaling (DVFS) are widely applied to modern processors. These strategies provide different supply voltages for a processor. The power-gated PDN design trade-off between leakage saving and supply noises highly depends on the circuit's supply voltage. On one hand, leakage current of a design exponentially decreases with decrease in supply voltage (V_{DD}). Hence, power gating at low V_{DD} requires longer break even time to compensate its energy cost. It means that there are fewer opportunities to launch power gating at low V_{DD}. On the other hand, both switching and rush current noises, when normalized with respect to the nominal supply voltage, have a tendency to decrease with V_{DD}. In summary, leakage saving is the dominant design concern at low V_{DD}, while power integrity is the dominant design concern at high V_{DD}.

Fixed decap configuration is a typical strategy of power-gated PDN designs [Chen and Lin 2009]. As shown in Figure 4, the amount of local decaps are determined based on the switching noise at high V_{DD}, for it is the worst case of power integrity. However, this amount of local decaps is overdesigned for the power integrity at low V_{DD} since the switching noise decreases with supply voltage [Xu and Li 2012]. Obviously, the decap configuration cannot be changed once circuit design is completed. Hence, extending turn-on time becomes the only method to suppress the rush current noise at low V_{DD}. As a result, the leakage saving at low V_{DD} is restricted by the overdesigned decaps.

In this article, we propose a flexible decap allocation configuration to adapt to supply voltage as shown in Figure 4. Reroutable decaps can act as local decaps or global decaps through controlling the switches. Hence, we can provide different decap configurations (LDs/GDs/RDs) for each V_{DD} level through the utilization of reroutable decaps. In this case, the design concerns (leakage saving and power integrity) at different voltage

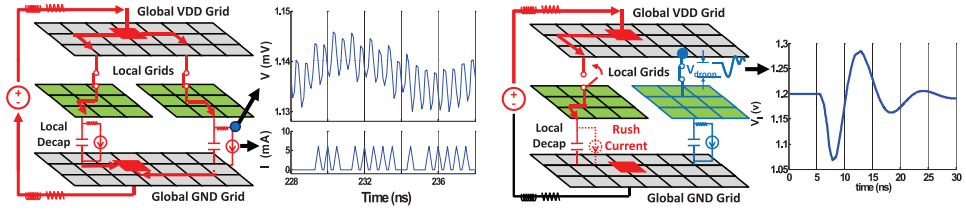


Fig. 5. Two types of supply noises associated with a power-gated PDN. (a) Switching noise is due to the switching activities of logic devices. (b) Rush current noise is due to the rush current created to charge the local decaps during the wake up process.

levels can be optimized separately. Therefore, the optimal design can be achieved for each supply voltage level.

The remainder of this article is organized as follows: In Section 2, we discuss the trade-offs and typical strategies of power-gated PDN designs. In Section 3, we propose LD&GD strategy and discuss related design issues. In Section 4, we propose LD&GD&RD strategy and discuss corresponding design issues. The flexible decoupling allocation strategy is proposed in Section 5 for a PDN design with two V_{DD}s. In Section 6, we present the simulation-based optimization flows for our proposed strategies. Section 7 shows the experimental results. The final section concludes our study.

2. TRADE-OFFS AND TYPICAL STRATEGIES OF POWER-GATED PDN DESIGNS

Power integrity and leakage saving are two key design considerations for power-gated PDN designs. In this section we analyze the design trade-offs pertaining to switching noise, rush current noise, and leakage saving.

2.1. Trade-off between Suppressions of Switching Noise and Rush Current Noise

Switching noise is due to the time variant switching current flowing through the power grids as shown in Figure 5(a). Switching noise is composed of high-frequency and mid-frequency components. The high-frequency component is due to the IR drop caused by resistive power grids. While the mid-frequency component is due to the resonance from the on-chip capacitance and the package inductance. The rush current noise appears when a local grid is turned on. During the wake up process, a rush current is created to charge up the decaps in the local grid. As shown in Figure 5(b), the other active local grid suffers the rush current noise and may generate logic errors. The power integrity of each logic device depends on the superposition of the two types of supply noises.

Local decaps are typically utilized to suppress switching noise. Parts of the current required by a switching devices is provided by its nearby local decaps. Hence, the high-frequency component of switching noise is suppressed. In addition, utilization of local decaps reduces the peak impendence of the PDN and thereby the mid-frequency component of switching noise. Local decaps are primary sources of rush current noise at the same time. Extending the turn-on time is a common method to reduce rush current noise in typical PDN designs. Turn-on time is related with leakage saving and performance delay that is discussed in the following section.

In addition, we have to admit that a current flows though global GND grid when one local grid's voltage goes to 0V. The current may bring voltage fluctuations to other active domains. However, we ignore these fluctuations based on the following reason. Power gating is an aggressive leakage saving strategy in modern processor designs. For example, power gating is usually the highest-level S-states (sleep states) of Intel's processors. Hence, after one local grid is turned off, most of switching cells (e.g.,

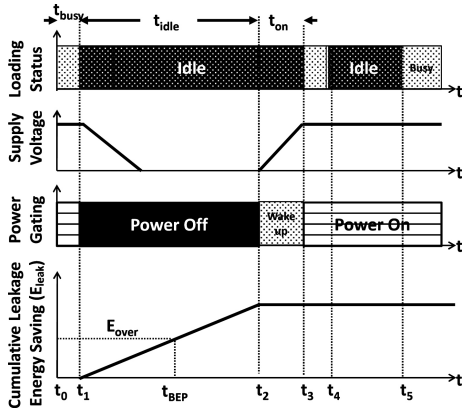


Fig. 6. Power gating process. The sleep transistor is supposed to be turned off as soon as the idle cycles arrive. t_{BEP} is the break even point at which time $E_{leak} = E_{over}$.

clock buffers) stop toggling. Loading capacitors and external capacitors are mainly discharged by leakage current. Therefore, we ignore the noise due to discharging since the peak current is limited.

2.2. Trade-Off between Rush Current Noise Suppression and Leakage Saving and Performance Delay

The power gating process is shown in Figure 6. The sleep transistors are turned off as soon as the idle cycles arrive at t_1. The supply voltage of the local grid gradually falls to 0. When the idle cycles end at t_2, the sleep transistors are turned on. It takes time $t_{on} = t_3 - t_2$ to recharge the local decaps to V_{DD}. After voltage recovery at t_3, the local domain starts to work again. Power gating saves leakage consumption during the idle cycles $t_{idle} = t_2 - t_1$. But the benefit obtained is at the cost of the performance delay and the energy overhead.

The total execution time for a single task without power gating is $t_{idle} + t_{busy}$. With power gating, the total execution time is extended to $t_{idle} + t_{busy} + t_{on}$. Therefore, the turn-on time t_{on} is the performance delay of the power gating technique.

In addition, the net energy saved by power gating is

$$E_{save} = E_{leak} - E_{over}, \tag{1}$$

where E_{leak} is the leakage energy saved by power gating during t_{idle} and $E_{over} = E_{ctrl} + E_{LD} + E_{on}$ is the energy overhead of the power gating, where E_{ctrl} indicates the energy spent on sleep transistor controlling, E_{LD} is the energy consumed to recharge the local decaps and E_{on} is the leakage energy consumption during turn-on time t_{on}. The time point at which the leakage saving compensates the energy overhead ($E_{leak} = E_{over}$) is the break even point t_{BEP}. If $t_{idle} < t_{BEP}$, the energy overhead overwhelms the leakage saving and thereby the power gating should not be applied to the idle time slot. For example, the idle slot from t_4 to t_5 in Figure 6 is too short to save energy through power gating. Hence, lots of leakage saving opportunities are missed due to the energy overhead (E_{over}).

Turn-on time plays a key role in determining the trade-offs between energy saving, performance delay, and rush current noise. Shortening turn-on time reduces energy overhead (E_{on}) as well as performance overhead (t_{on}). But, in order to reduce rush current noise, turn-on time is increased so that LDs are charged slowly thereby reducing rush current noise. An increase in turn-on time can eat into the leakage savings obtained through power gating.

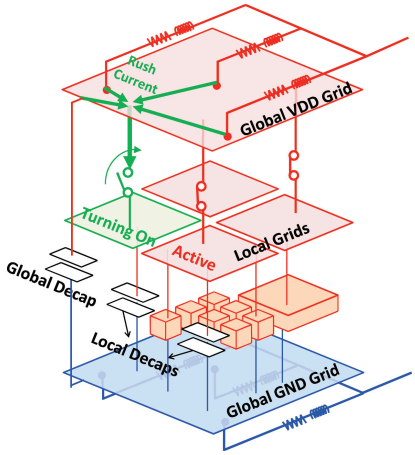


Fig. 7. Structure of global decaps. Global decaps are allocated between the global VDD grid and the global GND grid. The main utilization of global decaps is to suppress rush current noise through providing parts of rush current during local grids' wake up process.

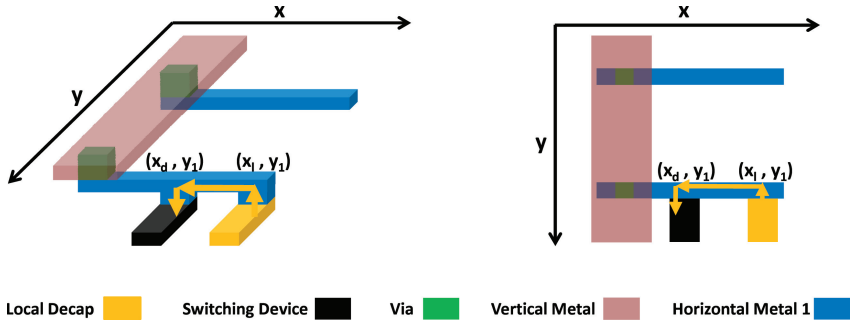


Fig. 8. The schematic layout and the top view of a typical PDN with a local decap. Only the global V_{DD} grid and the local grid are shown in the figure. The global GND grid is not depicted in the layout.

3. PROPOSED LOCAL/GLOBAL DECAP STRATEGY

In typical power-gated PDN designs, rush current noise is mainly suppressed by extending the turn-on time. However, as discussed in the last section, it reduces the leakage of power gating technique. In this section we propose a local/global decap strategy (LD&GD strategy) to further reduce supply noises especially the rush current noise. With the utilization of global decaps, more energy can be saved by shrinking the turn-on time.

3.1. Switching Noise Suppression

The LD&GD strategy utilizes both local decaps and global decaps to suppress switching noise. A global decap is connected between the global VDD grid and the global GND grid as shown in Figure 7. Global decaps are able to suppress switching noise (both high- and mid-frequency components), though they are not as efficient as equal amount of local decaps.

The schematic layout and the top view of a typical PDN with a local decap is shown in Figure 8. The schematic layout is based on a real industrial processor design with standard cells. The local grids are implemented by horizontal metal layer 1 (M_{H1}) and vertical metal layer (M_V). The local decap is located in the same row of the switching

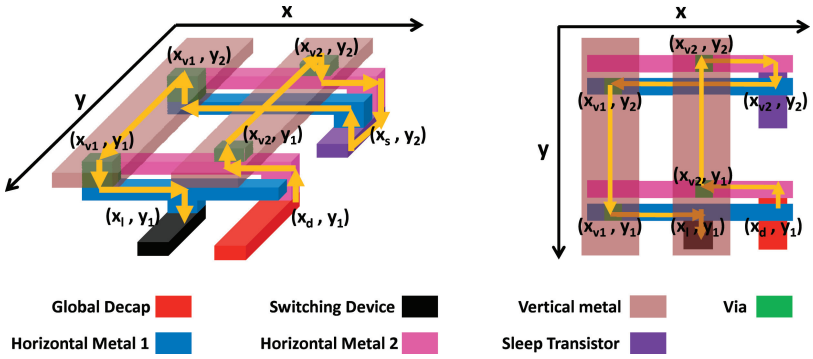
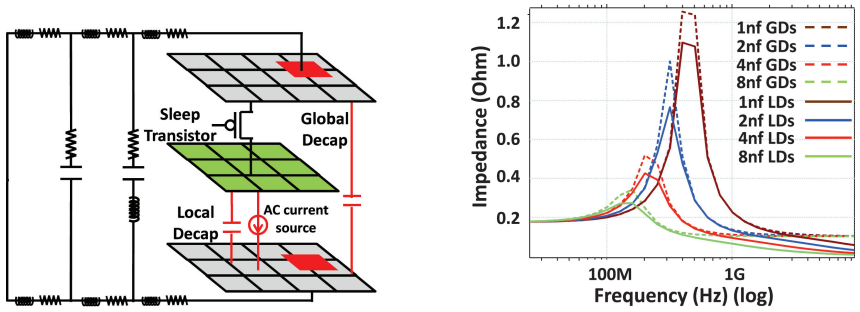


Fig. 9. The schematic layout and the top view of a PDN with a sleep transistor and a global decap. Only the global V_{DD} grid and the local grid are shown in the figure. The global GND grid is not depicted in the layout. Horizontal metal layer 1 and 2 are connected by a sleep transistor.



(a) (b)

Fig. 10. On-chip decaps' influence on circuit resonance. (a) The circuit model for analysis. (b) Impedances of the chip with different amount of local decaps and global decaps.

cell. The resistance between the local decap and the switching cell is a short metal segment on M_{H1}. Hence, the local decap can effectively suppress the high-frequency component of switching noise due to the small RC delay.

The schematic layout of a global decap is shown in Figure 9. The global grids are composed of horizontal metal layer 2 (M_{H2}) and vertical metal layer (M_V). M_{H1} and M_{H2} are connected by the cells of a sleep transistor. The resistance between the global decap and the switching cell is composed of the resistance of global grid (metal wires and vias), the equivalent resistance of the sleep transistor, and the resistance of local grid (metal wires and vias). The high resistance path introduces a large RC delay. Hence, the global decap is not as efficient as a local decap to suppress the high-frequency switching noise.

Global decaps are also able to suppress the mid-frequency component of switching noise. The mid-frequency switching noise is due to the resonance of the circuit that can be measured in the circuit shown in Figure 10(a). In this circuit, the DC supply voltage of the circuit is shorted. All the current loadings are removed. Only one AC current source is connected with the power grid. The amplitude of the current source is 1A. In this case, the impedance looking from the current source is shown in Figure 10(b). We compare the impedance with local decaps and the impedance with the same amount of global decaps. Both local decaps and global decaps are able to suppress the peak of the resonance. However, the resonance reduction with local decaps is more obvious than the one with the same amount of global decaps. This is because the resistance between

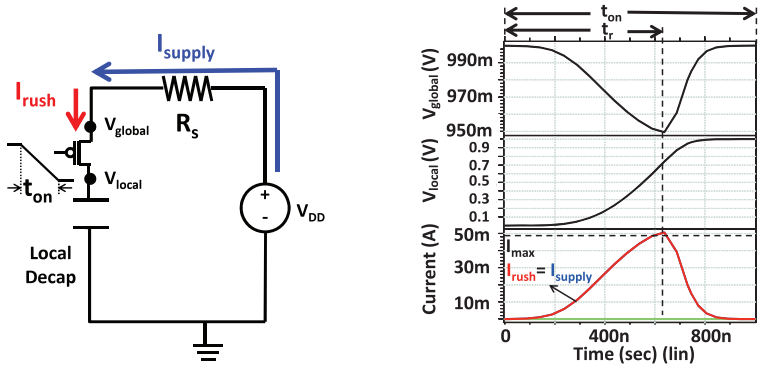


Fig. 11. Rush current noise suppression through extending turn-on: (a) simple circuit model with no global decap; R_s indicates the equivalent resistance between the supply voltage and the sleep transistor; (b) voltage drop observed for V_{global} and the corresponding rush current.

the local decaps and the current loading is much smaller and thereby they can provide a lower impedance path.

Although global decaps have the ability to suppress the switching noise, they are not as efficient as equal amount of local decaps. Therefore, local decaps are still the main technique to suppress the switching noise in our proposed PDN design.

3.2. Rush Current Noise Suppression

When a local grid is turned on, a rush current is created to charge the local decaps on that local grid. The local decaps include no-switching logic devices that act as decaps. The rush current leads to voltage drops in the global grid and the other active local grids. As a result, logic devices on the other active local grids may generate logic errors due to the voltage drops (rush current noise). The LD&GD Strategy takes use of global decaps to reduce the rush current and thereby turn-on time can be shortened to save more energy.

Extending the turn-on time t_{on} suppresses the noise by decreasing the peak of rush current. Sleep transistors and the local decaps are modeled as the source of the rush current as shown in the simple circuit example of Figure 11(a). R_s is the equivalent resistance between the supply voltage and the sleep transistor. $I_{supply}(t)$ is the current provided by the supply voltage. $I_{rush}(t)$ is the rush current drawn by the sleep transistor. Then, the current provided by the power supply must meet

$$I_{supply}(t) \leq I_{max} = \frac{r \times V_{DD}}{R_s}, \tag{2}$$

where r is the ratio of the maximum tolerable rush current noise to the supply voltage. In a typical PDN design, power supply is the only source to provide the rush current. Hence, we have $I_{rush}(t) = I_{supply}(t)$. In this case, the turn-on time of the sleep transistor (t_{on}) must be long enough to make sure the peak of the rush current $I_{rush}^{peak} \leq I_{max}$ as shown in Figure 11(b). The voltage of the local grid/decap V_{local} takes a long time to recover to V_{DD} since the charging process is slowed down.

To this end, without extending the turn-on time, global decaps can be used to suppress the noise by reducing I_{supply} instead of I_{rush}. As shown in Figure 12(a), both the power supply and the global decap are the sources to provide charging current. Hence, $I_{rush}(t) = I_{supply}(t) + I_{decap}(t)$, where $I_{decap}(t)$ is the current provided by the global decap. With the charge from the global decap, it is not necessary to slow down the charging

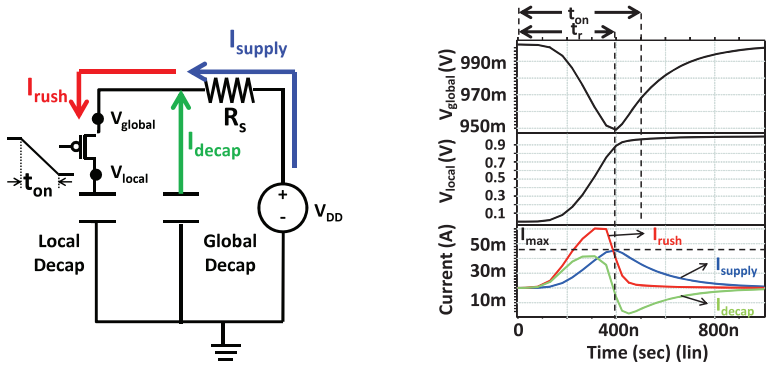


Fig. 12. Rush current noise suppression of global decaps: (a) simple circuit model with a global decap; (b) voltage drop of global grid and rush current

Table I. Impacts of On-Chip Decaps and Turn-On Time on the Design Concerns

Design option	Switching noise	Rush current noise	Power gating times	Energy overhead	Execution time
LD insertion	↓	↑	↓	↑	–
GD insertion	↘	↓	–	–	–
t_{on} shortening	–	↑	↑	↓	↓

↑increase ↓decrease ↘slightly decrease –no change

process in order to guarantee (2). Therefore, the voltage of the local grid can rise to V_{DD} quickly.

The utilization of global decaps relaxes the constraint of turn-on time. The turn-on time can be significantly shortened since the rush current noise is reduced by the global decaps.

3.3. Design Strategy

According to the analysis above, the impacts of local decaps (LDs), global decaps (GDs) and turn-on time (t_{on}) on the design concerns are summarized in Table I. Based on these impacts, the LD&GD design strategy uses local decaps and global decaps to suppress switching noise and rush current noise respectively. After the power integrity specification is met, turn-on time is further shortened to apply power gating for shorter idle time, reduce the energy overhead and thereby save more leakage power.

The power integrity specification may be specified as follows. First, total supply noise (superposition of switching noise and rush current noise) should be smaller than the maximum tolerable voltage drop. Second, switching noise and rush current noise should be respectively smaller then their own tolerance. In practice, one may set up a tighter tolerance for one of the two noises, say, rush current noise, as it may lead to an overall smaller budget for decoupling capacitance. In practice, the total decaps budget is limited due to fixed on-chip white space. Therefore, the total supply noise and each type of noises is tuned by the proportion between local decaps and global decaps. In Figure 13, the total decap budget (100nf) is divided into local decaps and global decaps. Rush current noise is reduced though increasing the ratio of GDs to LDs, while switching noise is reduced by decreasing the ratio.

Besides the decap configuration, turn-on time is another design parameter that determines the total supply noise. As shown in Figure 14, with the use of global decaps,

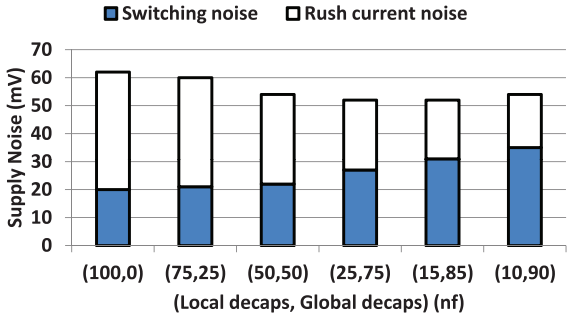


Fig. 13. Trade-off between switching noise and rush current noise. The power-gated PDN utilized for simulation is shown in Figure 3. Total decap budget (100nf) is divided into local decaps and global decaps. Local decaps and global decaps are uniformly distributed on local grids or global grids. The switching devices are modeled as triangular current sources [Kozhaya et al. 2002]. Turn-on time is 1000ns.

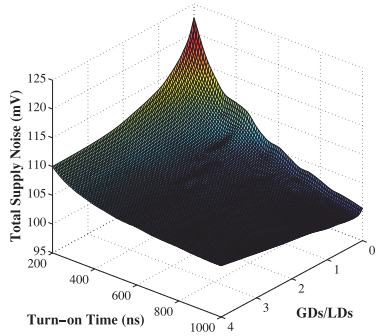


Fig. 14. Total supply noise is controlled though the LD&GD design strategy. Total decap budget (100nf) is divided into local decaps and global decaps. Local decaps and global decaps are uniformly distributed on local grids or global grids. The switching devices are modeled as triangular current sources [Kozhaya et al. 2002].

the proposed LD&GD Strategy is to exploit an optimal split between LDs and GDs for a given total decap budget to adjust the ratio between rush current noise and switching noise and maximize the overall power integrity.

The drawback of LD&GD design strategy is that a large amount of global decaps is needed. Assume that the maximum tolerable voltage droop is $0.1V_{DD}$, the total charge to recharge the local decaps is given by

$$Q_{rush} = 0.9V_{DD}C_{local}, \tag{3}$$

where C_{local} is the amount of local decaps. If all the charge is provided by the global decaps, we need approximately

$$C_{global} = \frac{Q_{rush}}{0.1V_{DD}} = 9C_{local}. \tag{4}$$

In most of the cases, it is hard to meet this requirement of global decaps. In order to explain the drawback of LD&GD, we use a highly simplified model that is only global decaps are used to charge local decaps. In practice, the voltage supply has to provide parts of charges. Hence, Equation (4) estimate the upper bound of required global decaps. The goal of this equation is not to provide extremely accurate analysis but to illustrate the key design problems.

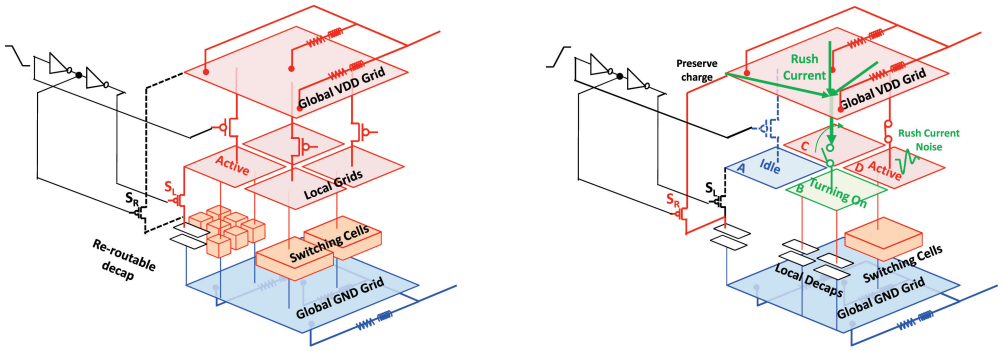


Fig. 15. Reroutable decap functions. (a) Function 1: when the local grid is active, the reroutable decap acts as a local decap to suppress the switching noise of its own power domain. (b) Function 2: when the local grid is turned off, the reroutable decap is routed to the global VDD grid. It acts as a global decap to suppress the supply noises of other local domains. In addition, the significant charge on reroutable decap is preserved by the global grid.

4. PROPOSED LOCAL/GLOBAL/REROUTABLE DECAP STRATEGY

As discussed in the previous section, large amounts of decaps are needed in order to achieve a short turn-on time. It can be very hard to find a feasible decap configuration when the decap budget is very limited. To deal with this problem, we propose the Local/Global/Reroutable Decap Strategy (LD&GD&RD strategy) that uses reroutable decaps (RDs), a new design concept proposed in our recent work [Xu et al. 2011], to further relax the tight interaction between power integrity and power leakage saving.

4.1. Structure and Functions of Reroutable Decaps

The structure of reroutable decaps is shown in Figure 15(a). Reroutable decaps are essentially programmable decoupling devices. For each reroutable decap, two switches S_L and S_R are used to control the decap routing. The functionalities of reroutable decaps are described below.

4.1.1. Function 1. The first function of a reroutable decap is to act as a local decap for its own local grid as shown in Figure 15(a). When the local grid is active, S_R is off and S_L is on. The reroutable decap is connected to local grids as a local decap to suppress switching noise. The equivalent resistance of S_L can impact the efficiency of a reroutable decap to suppress switching noise. Next section discusses design requirements for S_L.

4.1.2. Function 2. The second function of reroutable decap is to act as a global decap and preserve the charge on itself as shown in Figure 15(b). When local grid A goes to sleep, S_L is turned off and S_R is turned on. The reroutable decap is routed to global VDD grid. During this time, it acts as a global decap that aids in suppressing both switching noise on global grid and rush current noise on neighboring local grids For example, local grid B creates rush current during its wake up process. The rush current brings rush current noise to active local grid C and D. The reroutable decap provides current required by local grid B and thereby reduces the rush current noise of C and D. Most of the charge on reroutable decaps is preserved by the global VDD grid. Hence, when the reroutable decap is routed back to local grid A (Function 1), it creates much less rush current noise than a local decap during A's wake up process. As a result, the rush current noise created by local grid A is significantly reduced.

A special case is that the power integrity specification of a local grid can be met by GDs and its own LDs and RDs. In this case, it is not necessary to have reroutable

Table II. On-Chip Decaps Comparison

Type	Switching Noise Suppression	Rush Current Noise Suppression	Energy Overhead	Area Overhead
LDs	Excellent	Negative	√	–
GDs	Poor	Good	–	–
RDs	Good	Excellent	–	√

decaps work as global decaps to suppress the supply noises of other active local grids. Hence, S_R is only used to preserve charge on the reroutable decap. Since only leakage current flows through S_R, it can be made small to reduce the area overhead.

4.2. Advantages of Reroutable Decaps

We summarize and compare different types of on-chip decaps in Table II. Reroutable decaps avoid the disadvantages of LDs and GDs. First, reroutable decaps are more efficient than global decaps to suppress the switching noise. Compared with global decaps, reroutable decaps are allocated on the same metal layer of the switching cells. Hence, they are closer to the sources of switching noise than global decaps. Second, reroutable decaps reduce rush current and energy overhead of power gating. The charge of a reroutable decap is preserved by the global VDD grid. Hence, they require little charge during the wake up process. This means turn-on time can be shortened and leakage energy consumed during wake up E_{ton} is reduced. By replacing parts of LDs with RDs, the energy overhead E_{LD} is decreased. Therefore, the total energy overhead of power gating E_{over} is significantly reduced. Compared with same amount of LDs or GDs, reroutable decaps occupy more on-chip area due to switches S_L and S_R.

4.3. Design Strategy

The LD&GD&RD Strategy exploits reroutable decaps to reduce rush current noise and the energy overhead. Two design issues emerge with this strategy: allocation of reroutable decaps and the size of the S_L and S_R switches.

4.3.1. Allocation of Reroutable Decaps. We firstly discuss the influence of routing metal. The routing metal inevitably increases the equivalent resistance between decap and local/global node. The increased resistance may cause two problems. First, the resistance of routing metal may increase the RC delay of charging decaps. A decap may loss parts of charges to suppress supply noises. Then, the voltage supply should recharge the decap to V_{DD} before next supply noise appears. However, the voltage of decap may not have enough time to recover to V_{DD} if the RC delay is too large. This problem can lead to power integrity issue since the decap may finally loss its effects. Second, the efficiency of decap is decreased due to the increased RC delay. For example, when the voltage of a local grid node fluctuates, long-routing decaps can provide less charges during the same time than short routing decaps. In order to reduce the influence from routing metal, we avoid long-distance routing in this article. For a reroutable decap, the controlling transistors S_L and S_R are placed adjacent to the MOS-based decap. S_L and S_R are respectively connected to the nearest local grid metal layer and global grid metal layer through vias.

We discuss the allocation of reroutable decaps on local grid next. Unlike typical on-chip decaps, reroutable decaps are reused by more than one local grids. On one hand, a reroutable decap acts as a local decap to suppress the switching noise of its own power domain. On the other hand, when the local grid is turned off, it acts as a global decap to suppress supply noises of other power grids. Hence, the allocation of reroutable decaps should consider both of these cases.

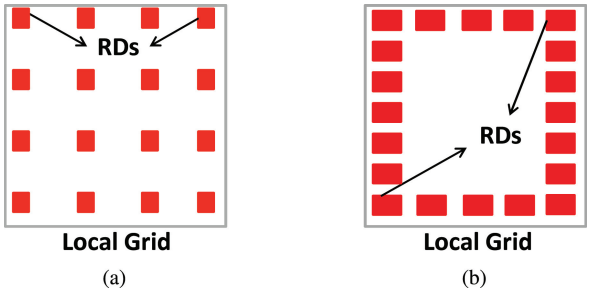


Fig. 16. Two different allocations of reroutable decaps: (a) distributed allocation; (b) clustered allocation.

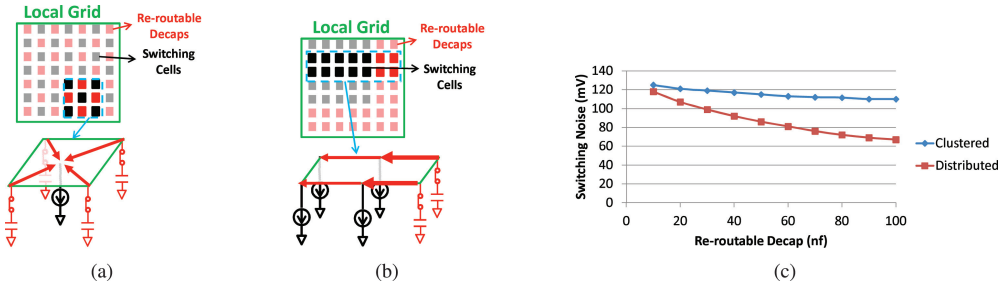


Fig. 17. Switching noise suppression with reroutable decaps through different allocations. The simulations are based on the circuit model shown in Figure 3. Only reroutable decaps are utilized in the circuit (no LD or GD). The amount of RDs is taken as a tuning parameter. The switching noises of the circuit with different amounts of RDs are monitored. (a) Distributed allocation. (b) Clustered allocation. (c) Switching noises under different reroutable decaps allocation.

Reroutable decaps can be allocated in two different ways. The first one is distributed allocation. As shown in Figure 16(a), reroutable decaps are uniformly distributed on local grid *A*. The other one is clustered allocation that is shown in Figure 16(b). Reroutable decaps are densely located at the boundaries of local grid *A*. The advantages and disadvantages of each allocation are discussed as follows.

Distributed allocation is advantageous to suppress switching noise. The resistance between a reroutable decap and a switching cell determines the efficiency of switching noise suppression. Through distributed allocation, reroutable decaps are located among the switching cells of local grid *A* as shown in Figure 17(a). Hence, the switching noise of each switching cell is suppressed by the reroutable decaps nearby. In contrast, the reroutable decaps are located along the boundaries of local grid *A* in clustered allocation as shown in Figure 17(b). Since they are allocated far away from most of the switching cells, large resistance weakens the suppression of switching noise. As shown in Figure 17(c), the switching noise under distributed allocation is smaller than the one under clustered allocation with same amount of reroutable decaps. It indicates that RDs in distributed allocation are more efficient than the ones in clustered allocation to suppress switching noise.

On the other hand, clustered allocation has an advantage over distributed allocation to suppress rush current noise. When a local grid is turned off, its reroutable decaps are routed to the global VDD grid. As shown in Figure 18(a) and 18(b), local grid *A* is turned off and the reroutable decaps of *A* act as global decaps to suppress the rush current noises of other active local grids (*C* and *D*). The noise is due to the rush current created by local grid *B* during its wake up process. For distributed allocation in Figure 18(a), reroutable decaps are allocated far away from local grid *B* that is the source of rush

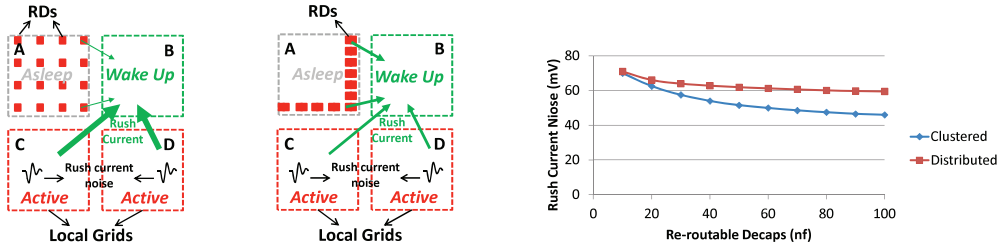


Fig. 18. Rush noise suppression with reroutable decaps through different allocations. The circuit model is shown in Figure 3. Only local decaps and reroutable decaps are utilized in the circuit (no GD). The amount of local decaps allocated in each local domain is 25nf. Reroutable decaps are only allocated in local grid *A*. (a) Distributed allocation. (b) Clustered allocation. (c) Rush current noises under different allocations.

current. Hence, such kind of allocation is disadvantageous to the suppression of rush current noise. In contrast, reroutable decaps are allocated along the boundaries of local grid *A* under clustered allocation. When the reroutable decaps are routed to the global grid, they are closer to local grid *B* than distributed allocation and thereby more current can be provided by these reroutable decaps. As shown in Figure 18(c), with the same amount of reroutable decaps, clustered allocation is more efficient to suppress rush current noise.

As discussed above, a reroutable decaps is reused to suppress switching noise (own local grid) and to suppress rush current noise (other local grids) at different time. In order to enhance the efficiency, distributed allocation and clustered allocation can be utilized together. We divide reroutable decaps into two groups. Reroutable decaps of the first group are uniformly distributed on the local grid to improve the efficiency of switching noise suppression. Reroutable decaps of the second group are allocated at local grid boundaries to improve the efficiency of rush current noise suppression. In order to determine the RD amount of each group, we propose a simulation based optimization flow that is discussed in Section 6.

For the special case discussed in Section 4.1, reroutable decaps are only used to preserve charge when the local grid is turned off. In this case, reroutable decaps are not used to suppress the rush current noises introduced by other local grids. Therefore, all the reroutable decaps can be allocated through distributed allocation.

4.3.2. Sizes of Switch S_L and S_R. Switch S_L connects a reroutable decap with a local grid. It determines the charge that can be provided by the reroutable decap for switching noise suppression. The size of S_L is constrained by two issues: area overhead and capacitance overhead. The area overhead is due to the addition of switch that is given by

$$A_0 = \frac{Area\ of\ S_L}{Area\ of\ decap}.$$

The capacitance overhead is another constraint of S_L. The series resistance of S_L reduces the efficiency of capacitance. Due to reduced efficiency, more capacitance is required to meet the power integrity requirement if we replace local decaps with reroutable decaps. The capacitance overhead is given by

$$C_0 = \frac{capacitance\ of\ RD}{equivalent\ capacitance\ of\ LD}.$$

Figure 19 shows capacitance overhead and switch area overhead of the reroutable decaps required to reduce switching noise to 10% V_{DD}. As the width of S_L increases,

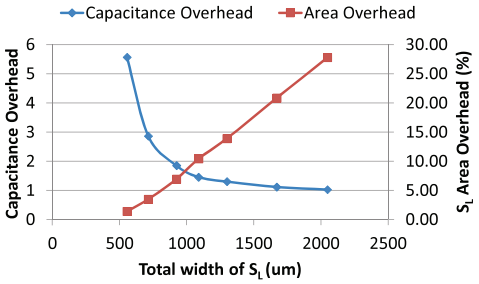


Fig. 19. Capacitance overhead and switch area overhead of the reroutable decaps required to reduce switching noise to tolerable value. The maximum tolerable switching noise is 10% of V_{DD}. The circuit model is shown in Figure 1. Only reroutable decaps are utilized in the circuit (no LD or GD). The reroutable decaps are allocated through distributed allocation. The width of S_R is 0.5um for each reroutable decap.

the area overhead increases while the capacitance overhead decreases. In Section 6, we propose simulation based design optimization to determine the width of S_L in order to balance between capacitance overhead and area overhead.

Similar issues should be considered in the design of switch S_R. On one hand, the size of S_R should be large enough to suppress the rush current noise introduced by other local grids. On the other hand, the area overhead of the switch should be controlled to save limited on-chip white space.

5. FLEXIBLE DECAP ALLOCATION STRATEGY FOR A PDN WITH MULTIPLE SUPPLY LEVELS

Power-gated PDNs with dynamic voltage and frequency scaling (DVFS) are also worth considering. DVFS technique dynamically changes the operation point (supply voltage and clock frequency pair) to improve performance and power saving. But the combination of power gating and DVFS makes a PDN design more complex since the design trade-offs vary with supply voltage.

5.1. Leakage and Noises at Different Voltage Operation Points

As discussed in Sections 3 and 4, the decap configuration (LD/GD/RD) at an operation point is determined based on leakage saving and power integrity. The trade-off between these two design concerns varies with supply voltage provided by the DVFS technique.

Leakage Saving: Power gating is exploited to reduce leakage power consumption that includes the subthreshold leakage and gate tunneling leakage. Subthreshold leakage becomes the dominant as the process technology scales down. The subthreshold current has an exponential relationship with threshold voltage (V_{TH}). A reduction of V_{TH} occurs at higher drain-source bias (V_{ds}) due to drain induced barrier lowering (DIBL). As a result, when the supply voltage of logic devices decreases linearly, the leakage current I_{leak} is reduced exponentially [Calhoun and Chandrakasan 2003]. Figure 20 shows the normalized leakage current of an inverter at different supply voltages. When the supply voltage decreases from 1.2V to 0.6V, the leakage current is reduced by about 20 times. As mentioned in Section 2, break even time (t_{BEP}) of power gating is the time during which leakage saving compensates energy overhead. Since the every overhead is mainly used to recharge the capacitance of the local grids, we have

$$E_{over} \propto C_{eq} V_{DD}^2, \tag{5}$$

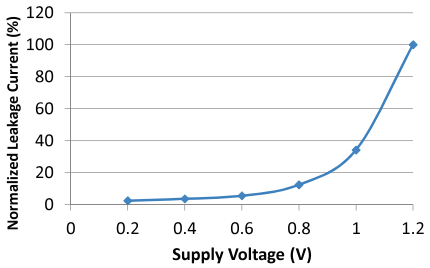


Fig. 20. Normalized leakage current of an inverter increases with the supply voltage (V_{DD}). Leakage current is normalized to the value when $V_{DD} = 1.2$V. The technology node is 45nm.

Table III. Trade-offs at Different Operating Points

V_{DD}	Swi. Noise/V_{DD}	Rush. Noise/V_{DD}	t_{BEP}
high	↑	↑	↓
low	↓	↓	↑

↑increase ↓decrease

where C_{eq} is the equivalent capacitance of the turning-on local grids. The break even time can be estimated as

$$t_{BEP} = \frac{E_{over}}{P_{leak}} \propto \frac{V_{DD}}{I_{leak}}, \tag{6}$$

where P_{leak} is the leakage power and I_{leak} is the average leakage current that exponentially decreases with V_{DD}. Hence, the break even time increases as the supply voltage decreases. It means that leakage consumption is harder to be saved through power gating at lower V_{DD}.

Supply Noises: The switching current created by switching cells and the rush current created during wake-up process both superlinearly increase with V_{DD}. As a result, the ratio of switching noise or rush current noise to V_{DD} increases as supply voltage linearly increases. In other words, it is harder to meet the power integrity specifications at higher V_{DD}.

The trade-offs with different supply voltages are summarized in Table III. It indicates that the design trade-offs change as the system switches between different voltage operating points. Power integrity is the dominant design concern at high-V_{DD} while leakage saving is the critical design concern at low-V_{DD}.

5.2. Flexible Decoupling Design

For typical PDN designs, only local decaps are used (LD only strategy). In this case, the required amount of local decaps varies with the supply voltage. As shown in Figure 21, the required local decaps decrease as V_{DD} scales down when the total supply noise tolerance is kept as a fixed percentage of the nominal V_{DD}. In order to meet the power integrity requirement in the worst case, local decaps are designed to suppress the switching noise at highest V_{DD}. However, such amount of local decaps is superfluous at lower V_{DD}. Superfluous local decaps create extra rush current noise and thereby limit turn-on time shrinking. As a result, power gating has fewer opportunities to save leakage at low V_{DD}.

Obviously, a fixed decap configuration cannot adapt to the design trade-off changes with V_{DD}. A flexible decoupling strategy is proposed here using RDs. Two types of RDs are used: (a) regular RDs illustrated in Figure 22(a) and (b) global RDs illustrated in Figure 22(b). When the local grid is active, regular RDs are connected to the local grid and global RDs are connected to the global VDD grid. When the local grid is idle,

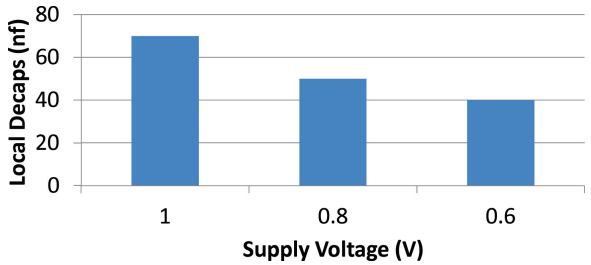


Fig. 21. The decaps required at different supply voltages for the LD Only Strategy. The maximum tolerable supply noise is 10% of V_{DD}.

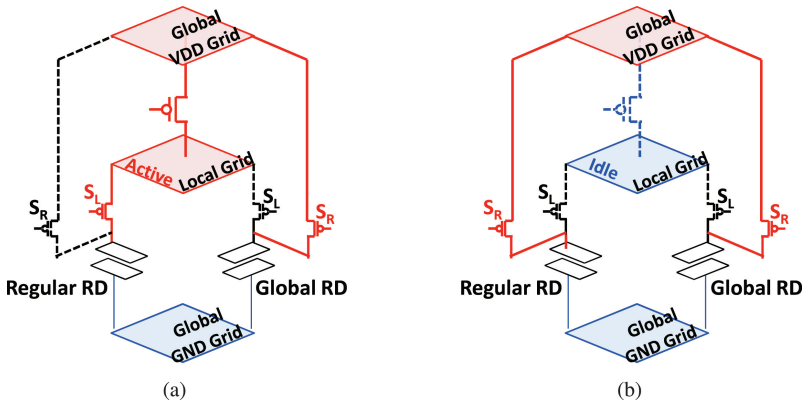


Fig. 22. Usages of regular RD and global RD. (a) When the local grid is active, regular RD is connected to the local grid and global RD is connected to the global VDD grid. (b) When the local grid is idle, both regular RD and global RD are connected to the global VDD grid.

both regular RDs and global RDs are connected to the global VDD grid. Regular RDs make sure that the design has enough decaps to suppress switching noise. Global RDs are used to further reduce rush current noise and thereby increase leakage saving. A flexible decap configuration is provided through tuning the proportion between these two types of RDs. At high V_{DD}, all RDs are used as regular RDs since it is the worst case for switching noise. As V_{DD} decreases, leakage saving becomes the main design concern. Hence, the proportion of global RDs is increased to further reduce rush current noise. As a result, the PDN can be optimized at each V_{DD} level through this flexible decap allocation configuration.

6. OPTIMIZATION OF POWER-GATED PDN DESIGN

In this section, we propose a simulation based optimization flow to design a power-gated PDN automatically. Global decaps, local decaps, reroutable decaps and the turn-on time are taken as design parameters. Supply noises, leakage saving and area overhead are taken as components of the objective function.

6.1. Optimization with Single Supply Voltage

We propose a simulation based optimization flow in Figure 23 to implement the LD&GD&RD Strategy that is discussed in Section 4.

The design parameters of the strategy include the amount of LDs, GDs, RDs in distributed allocation, RDs in clustered allocation, turn-on time, and total width of S_L and S_R. These design parameters are constrained as follows. The descriptions of related

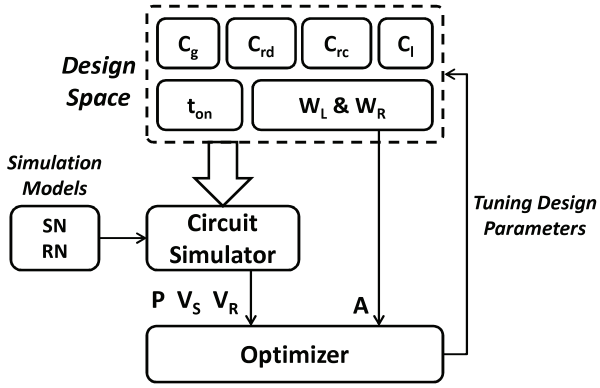


Fig. 23. Simulation based optimization flow with single supply voltage.

Table IV. Design Parameters for PDNs with Single Supply Voltage

V_S	maximum switching noise
V_R	maximum rush current noise
P	leakage power consumption
A	area overhead of the RDs' switches
C_l	amount of local decaps[1]
C_g	amount of global decaps[2]
C_{rd}	amount of reroutable decaps in distributed allocation
C_{rc}	amount of reroutable decaps in clustered allocation
C_{tot}	total on-chip decap budget
W_L	total width of switch S_L
W_R	total width of switch S_R
W_m	maximum width of reroutable decaps

[1]Local decaps are allocated in distributed allocation.
[2]Global decaps are allocated in distributed allocation.

parameters are listed in Table IV. The local decaps are allocated between local grids and global GND grid in distributed allocation. In this case, the local decaps are close to the switching cells and thereby the suppression of switching noise is enchanted. The global decaps are allocated between global VDD and GND grids in distributed allocation. This is because that the sleep transistors are allocated in distributed allocation, which are the drains of rush currents for global VDD grid. Hence, distributed allocation of global decaps is more effective to suppress rush current noise than clustered allocation.

$$\begin{cases} C_g + C_l + C_{rd} + C_{rc} & \leq C_{tot} \\ W_L + W_R & \leq W_m \\ C_g, C_l, C_{rd}, C_{rc}, W_L, W_R & \geq 0 \end{cases} \tag{7}$$

Two circuit models (SN and RN) are provided for the simulation. These two models share the same PDN structure. In model SN, all local grids are active. In model RN, only one local grid is active while the other local grids are asleep or waking up. Based on the design parameters selected from the design space and model SN, the maximum switching noise (V_S) can be obtained from the circuit simulation. The maximum rush current noise (V_R) and leakage consumption (P) can be obtained from the simulation of model RN. The area overhead (A) of reroutable decaps' switches is estimated based on the total width of S_L and S_R.

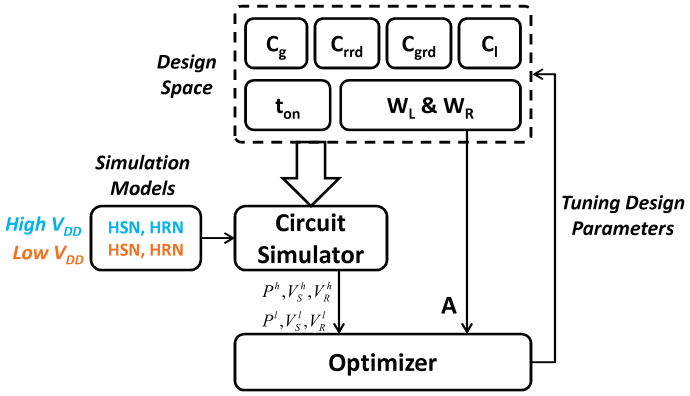


Fig. 24. Simulation based optimization flow with two supply voltages.

Based on V_S, V_R, P, and A, the optimizer evaluates the current design through an objective function and tune the parameters to improve the design for the next iteration. The objective function is given as

$$min \quad f = f_s(V_s) + f_r(V_r) + f_p(P) + f_a(A), \tag{8}$$

where f_s and f_r are respectively the penalty functions of switching noise and rush current noise, f_p is the penalty function of leakage power consumption, and f_a is the penalty function of switch area overhead. The optimization flow is not restricted to any specific objective function but can use any generic function of V_s, V_r, P, and A. The formulation of each penalty function can be selected by designers.

The penalty function we used in this article is given by

$$\begin{aligned} f &= f_s(V_s) + f_r(V_r) + f_p(P) + f_a(A) \\ &= w_s e^{\frac{V_s}{V_{s0}}} + w_r e^{\frac{V_r}{V_{r0}}} + w_p e^{\frac{P}{P_0}} + w_a e^{\frac{A}{A_0}}, \end{aligned} \tag{9}$$

where V_{s0}, V_{r0}, P_0 and A_0 are the normal values (design specifications), w_s, w_r, w_p and w_a are the penalty weights. Our particular formulation is just one way of solving the design problem.

6.2. Optimization with Multiple Supply Voltages

For a PDN design with multiple supply voltages, we only consider the special case where RDs are used to suppress the rush current noise created by their own local grid. Since the RDs are not used as global decaps, all of them are allocated in distributed allocation.

The optimization flow for a PDN design with two supply voltages (V_{DD}^h and V_{DD}^l) is shown Figure 24 which can be extended to handle a larger number of supply levels. V_{DD}^h and V_{DD}^l respectively indicate the high and low supply voltage.

The design parameters referred include the amount of LDs, GDs, regular RDs, global RDs, turn-on time, and total width of S_L and S_R. The descriptions of these parameters are listed in Table V. The constraints of the parameters are given as follows.

$$\begin{cases} C_g + C_l + C_{rrd} + C_{grd} & \leq C_{tot} \\ W_L + W_R & \leq W_m \\ C_g, C_l, C_{rrd}, C_{grd}, W_L, W_R & \geq 0 \end{cases} \tag{10}$$

Four simulation models are used for the PDN design with two supply voltages. HSN and LSN are respectively for the switching noise simulations at V_{DD}^h and V_{DD}^l. HRN

Table V. Design Parameters for PDNs with two Supply Voltages

$V_S^{h(l)}$	switching noise with $V_{DD}^{h(l)}$
$V_R^{h(l)}$	rush current noise with $V_{DD}^{h(l)}$
$P^{h(l)}$	leakage power consumption $V_{DD}^{h(l)}$
A	area overhead of the RDs' switches
C_l	amount of local decaps[1]
C_g	amount of global decaps[2]
C_{rrd}	amount of regular reroutable decaps[3]
C_{grd}	amount of global reroutable decaps[3]
C_{tot}	total on-chip decap budget
W_L	total width of switch S_L
W_R	total width of switch S_R
W_m	maximum width of reroutable decaps

[1]Local decaps are allocated in distributed allocation.
[2]Global decaps are allocated in distributed allocation.
[3]Reroutable decaps are allocated in distributed allocation.

and LRN are respectively for the rush current noise simulations at V_{DD}^h and V_{DD}^l. Based on the design parameters and simulation models, the maximum switching noise and rush current noise, the leakage consumption at V_{DD}^h and V_{DD}^l are obtained from circuit simulations.

Based on the outputs of circuit simulations, the optimizer evaluates the current design through an objective function and tune the parameters to improve the design for the next iteration. The objective function is given as

$$min \quad f = f_s^h + f_s^l + f_r^h + f_r^l + f_p^h + f_p^l + f_a, \tag{11}$$

where $f_s^{h(l)}$ is the penalty function of switching noise at $V_{DD}^{h(l)}$, $f_r^{h(l)}$ is the penalty function of rush current noise at $V_{DD}^h(l)$, $f_p^{h(l)}$ is the penalty function of leakage power consumption at $V_{DD}^{h(l)}$, and f_a is the penalty function of switch area overhead. The detailed formulation of each penalty function can be selected by designers. The parameters referred in the flow are described in Table V.

7. EXPERIMENTAL RESULTS

In this section, we present the experimental results of the PDN designs with single V_{DD} and two V_{DD}s.

The settings of the experiments are listed in Table VI. The interface to the optimizer and the optimization flow are implemented in C++. The package model parameters are from Gupta et al. [2007]. The power grids including four local grids are generated according to IBM power grid benchmarks [Nassif 2008]. The multicore processor workloads are generated based on benchmark suit PARSEC [Bienia 2011]. In this article, Asynchronous Parallel Pattern Search package (APPSPACK) [Griffin et al. 2008] are employed as the optimizer. APPSPACK is serial or parallel, derivative-free optimization software for solving nonlinear unconstrained, bound-constrained, and linearly-constrained optimization problems, with possibly noisy and expensive objective functions.

The models used for simulation is shown in Figure 25. The PDN structure includes 4 local grids. For the simulation of switching noise, all the local grids are active as shown in Figure 25(a). For the simulation of rush current noise, local grid A is asleep, local grid D is active, local grids B and C are turning on as shown in Figure 25(b).

Table VI. Experimental Setting

Single supply voltage	1V
High supply voltage (V_{DD}^h)	1V
Low supply voltage (V_{DD}^l)	0.6V
Technology node	45nm
Average power	12W
On-chip Decap budget (C_{tot})	100nf
Maximum RD switch overhead(W_m)	1000μm
Maximum tolerable switching noise	9.5% of V_{DD}
Maximum tolerable rush current noise	0.5% of V_{DD}
Number of power domains	4
Size of PDN	120K Nodes
Circuit simulator	HSPICE C-2009-0.9
Optimizer	APPSPACK [Griffin et al. 2008]

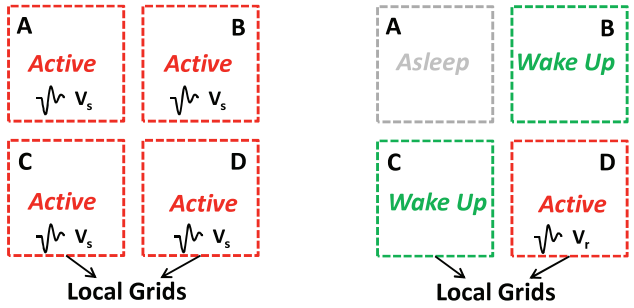


Fig. 25. Simulation models: (a) model for switching noise simulation; (b) model for rush current noise simulation.

We choose the simulation model of Figure 25(b) based on following considerations. Reroutable decaps' suppression of rush current noise actually includes three cases.

(1) Reroutable decaps' own domain is the source of rush current noise. For example, in Figure 25(b), the rush current noises are caused by local grid B and C. The reroutable decaps of B and C reduce the rush current noises through preserving their charges during idle time.
(2) Reroutable decaps' own domain is the victim of rush current noise. For example, in Figure 25(b), D suffers the rush current noises. The reroutable decaps of D stabilize the nodal voltage through acting as local decaps.
(3) Reroutable decaps' own domain is neither the aggressor nor the victim of rush current noise. For example, in Figure 25(b), local grid A neither causes nor suffers any rush current noise. The reroutable decaps of local grid A suppress the rush current noises through acting as global decaps to provide parts of rush current.

Our proposed flow is not limited to any specific simulation model. Users can choose one or more models based on different situations or purposes. In this article, we choose the model of Figure 25(b) in order to present a relatively common case in which reroutable decaps are used in three different ways to suppress rush current noise.

We compare three different design strategies: LD only strategy, LD&GD strategy and LD&GD&RD strategy. For the LD only strategy, only local decaps are utilized in the PDN design. For the LD&GD strategy, both local decaps and global decaps are utilized.

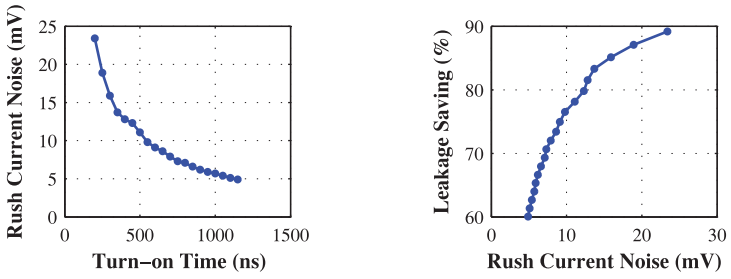


Fig. 26. Rush current noise and leakage saving through the LD only strategy. Switching noise is reduced to 9.5% of V_{DD}. (a) Rush current suppression fully depends on extending turn-on time. (b) The interaction between leakage saving and rush current noise. Leakage saving is restricted by rush current noise. The leakage saving is normalized to the leakage power consumed without power gating.

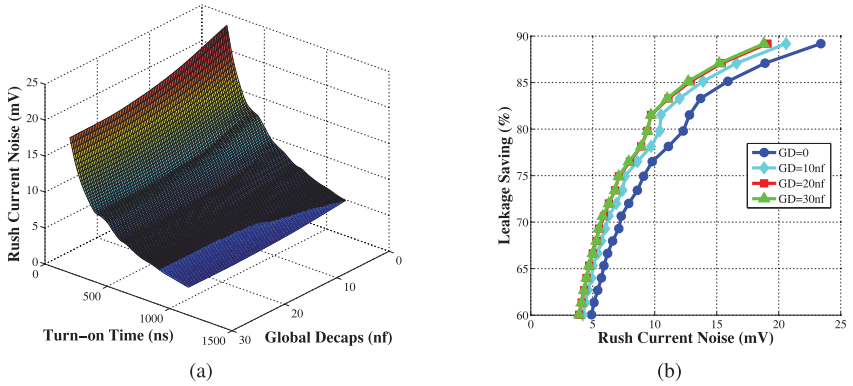


Fig. 27. Rush current noise and leakage saving through the LD&GD strategy. Switching noise is reduced to 9.5% of V_{DD}. (a) Rush current noise is suppressed by both turn-on time and global decaps. The gray zone in Figure 27(a) covers the designs with rush current noise under 0.5% of V_{DD}. (b) Global decaps relax the interaction between leakage saving and rush current noise.

For the LD&GD&RD strategy, local decaps, global decaps, and reroutable decaps are all used.

7.1. PDN Design with Single Supply Voltage

For the LD only strategy, rush current noise is mainly suppressed through extending the turn-on time. In Figure 26(a), all designs meet the requirement of switching noise suppression (9.5% of V_{DD}). In order to reduce the rush current noise to 0.5% of V_{DD}, turn-on time has to be extended to 1000ns. Since turn-on time determines the opportunities of power gating, leakage saving is restricted by rush current noise. As shown in Figure 26(b), rush current noise dramatically increases as more leakage is saved. In this figure, the leakage saving is normalized to the leakage power consumption without power gating. Therefore, the LD only strategy has limited leakage saving due to the tight interaction between rush current noise and leakage saving.

For the LD&GD strategy, global decaps are used to suppress the rush current noise. Figure 27(a) shows how rush current noise is influenced by turn-on time and the amount of global decaps. In this experiment, the switching noise is reduced to 9.5% of V_{DD}. The gray zone in the figure covers all feasible designs of which rush current noises are under 0.5% of V_{DD}. Compared with the LD only strategy, the feasible designs provided

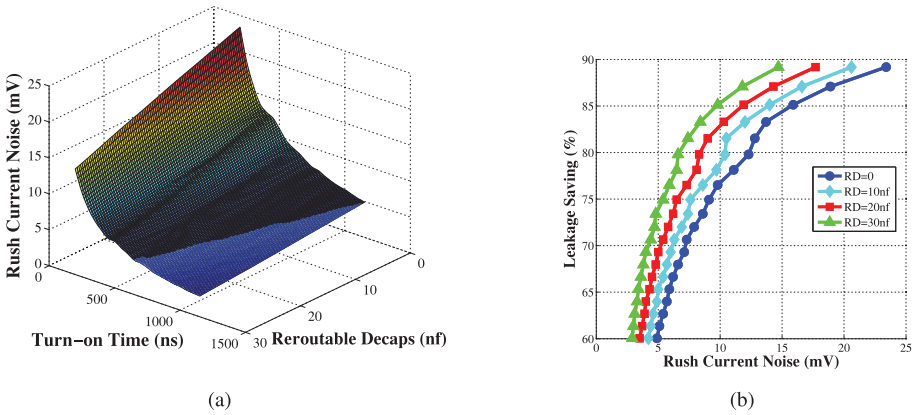


Fig. 28. Rush current noise and leakage saving through the LD&GD&RD strategy. No GD is used in order to evaluate the influence of reroutable decaps. Switching noise is reduced to 9.5% of V_{DD}. (a) Rush current noise is suppressed by both turn-on and reroutable decaps. The gray zone covers the designs whose rush current noises are under 0.5% of V_{DD}. (b) Reroutable decaps obviously relax the interaction between leakage saving and rush current noise.

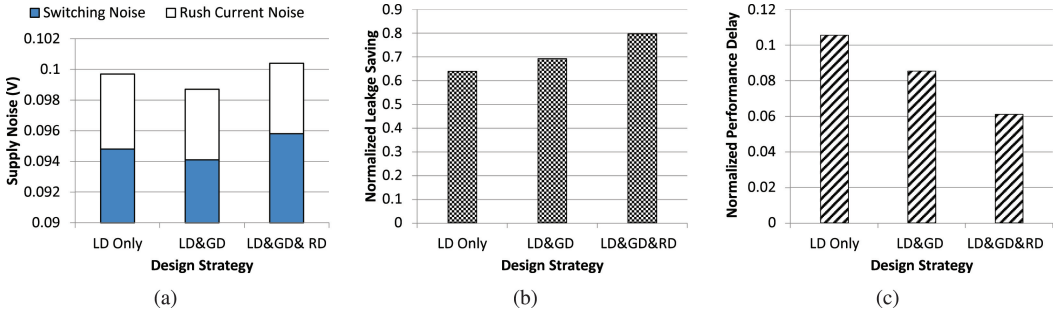


Fig. 29. Comparison of optimization results obtained from the LD only strategy, the LD&GD strategy and the LD&GD&RD strategy. (a) Comparison of supply noises. (b) Comparison of normalized leakage savings. The leakage savings through different design strategies are normalized to the leakage consumption without power gating. (c) Comparison of normalized performance delays. The performance delays through different design strategies are normalized to the execution time without power gating.

by the LD&GD strategy have shorter turn-on time. This is because the constraint of turn-on time is relaxed by global decaps. As shown in Figure 27(b), the interaction between leakage saving and rush current noise is relaxed by global decaps. In other words, the LD&GD strategy can save more leakage power than the LD only strategy upon the same specification of supply noises.

The LD&GD&RD Strategy exploits reroutable decaps to further reduce rush current noise. Figure 28 shows rush current noise and leakage saving of the PDN designs under the LD&GD&RD strategy. In this experiment, only reroutable decaps and local decaps are used. Compared with the LD&GD strategy, the zone of feasible designs in Figure 28(a) obviously extends. It indicates that reroutable decaps are more efficient to suppress rush current noise than the same amount of global decaps. Figure 28(b) shows that the interaction between leakage saving and rush current noise is further relaxed by the utilization of reroutable decaps.

Figure 29(a) presents the optimized supply noises obtained from the LD only strategy, the LD&GD strategy and the LD&GD&RD strategy. Supply noises are important design concerns of a PDN design. The three strategies have similar performance of sup-

Table VII. Comparison between Different Strategies for Single Supply Voltage

Strategy	LD only	LD&GD	LD&GD&RD
Decap Budget(nf)	100	100	100
Local Decap(nf)	70	45	39
Global Decap(nf)	0	55	49
RD in distributed allocation (nf)	0	0	7
RD in clustered allocation (nf)	0	0	5
Switching Noise(mV)	94.8	94.1	95.8
Rush Current Noise(mV)	4.9	4.6	4.6
Total Supply Noise(mV)	99.7	98.7	100.4
Turn-on Time(ns)	1150	800	450
Leakage Saving[1]	64.0%	69.3%	79.8%
Performance Delay[2]	10.6%	8.5%	6.1%
Decap Area[3]	70%	100%	103%
Runtime (hour)	N.A.	10.5	16.2

[1]Normalized to the leakage power consumed without power gating.
[2]Normalized to the execution time without power gating.
[3]Normalized to the area of total decap budget (100nf).

ply noises suppression. The maximum tolerable switching noise and rush current noise are respectively set as 9.5% and 0.5% as listed in Table VI. In this article, we choose a weighted sum of supply noises, power consumption, and area overhead as the objective function for design optimization. The penalty function we used is given by (9). Hence, based on the specific optimization package used (APPSPACK which can only solve unconstrained optimization problems in this case), our unconstrained optimization may not guarantee that the supply noises strictly meet the requirements. But practically, we can get pretty close to the tolerance ($\pm 2\%$). Our particular formulation is just one way of solving the design problem. We can also use other optimization formation where noise can be set as a hard constraint. There, this problem will go away completely.

Figure 29(b) and 29(c) respectively show the leakage saving and performance delay of optimization results obtained from the three strategies. The leakage saving is normalized to the total leakage consumption of the PDN design without power gating. The performance delay is normalized to the total execution time without power gating. The LD only strategy has no other means but extending the turn-on time to suppress the rush current noise. Turn-on time of the LD only strategy is extended long enough in order to meet the specification of rush current noise. Power gating with long turn-on time cannot be applied to short idle intervals that take up a large proportion of idle time. As a result, the normalized leakage saving of the LD only strategy achieves 60%. On the other hand, long turn-on time leads to a long delay of each power gating. Therefore, the performance delay is about 11% of the total execution time without power gating.

The LD&GD Strategy relaxes the interaction between rush current noise and turn-on time through the utilization of global decaps. Hence, the normalized leakage saving increases to 70% and the performance delay is reduced to 8.5%.

For the LD&GD&RD strategy, reroutable decaps are exploited to further reduce the rush current noise. Compared with global decaps, reroutable decaps are more efficient to suppress the rush current noise. In this case, the turn-on time can be significantly reduced. As a result, the tight interaction between the rush current noise and the leakage saving is relaxed by the reroutable decaps. In this case, this strategy saves about 80% leakage consumption that is the most leakage saving among the three strategies. The performance delay is reduced to 6.1%.

As shown in Table VII, the total decap budget (100nf) is not fully utilized in the design obtained though the LD only strategy. It is because that increasing local decap may lead to soaring rush current noise. Hence, the decap area only takes up 70% of the area of decap budget (100nf). However, this area saving is at the cost of leakage saving. The decap budget is fully used for both LD&GD strategy and LD&GD&RD strategy. This is because that they both have an effective mechanism (GDs or RDs) to suppress rush current noise. In practice, the control circuit for LD&GD&RD strategy, should be designed as a balanced network to deliver the control signal to each reroutable decap, which is very similar to a clock tree. We use a two-level buffer chain to control switches S_L and S_R of a reroutable decap, as shown in Figure 15. The input of the buffer chain is the power gating control signal. The two buffers respectively generate control signals for S_L and S_R. The control circuits are designed based on the following considerations.

First, we discuss about the delay of the control circuit. The dominant area overhead of the control circuit is the area of S_L and S_R. For a reroutable decap, the area of switches is about 10 20% of the area of the MOS based capacitor. In a typical design, the amount of reroutable decaps is about 10 20% of the total decap budget. Hence, the total area of S_L and S_R is 1 4% of the area of total decap budget. The area overhead is estimated in the optimization flow. The final obtained design reflects the area overhead.

Second, we discuss about the delay of the control circuit. The main capacitive loading of the control circuit is the gate capacitance of S_L and S_R, which is 1 4% of total decap budget as discussed above. The typical turn-on time of our power delivery network model is 1000s clock cycles when only local decaps are used. The typical idle time is several thousands of cycles. The delay of the control circuit is targeted at 2% of turn-on time that is about 10s 100s cycles. Compared with the turn-on time, this delay does not influence the total performance delay too much.

Finally, we discuss the energy overhead of the control circuit. The total energy consumed by the control circuit is mainly used to charge the gate capacitance of S_L and S_R. As discussed above, the equivalent capacitance is about 1 4% of the total decap budget. According to $E = CV_{DD}^2$, the total energy consumption of the control circuit can be estimated. The energy consumed by the control circuit is estimated in the optimization flow. The total energy consumed by the final obtained design includes this part of energy that typically takes up 2 5% of the total power consumption during idle time.

For the single supply voltage PDN design, the area overhead of the control circuits is 3% of the area of total decap budget (100nf). The power consumption of the control circuit is 4% of the total power consumed during idle time. The average delay of the buffer chain is 9.1ns.

7.2. PDN Design with Two Supply Voltages

Through the optimization flow proposed in Section 6.2, the optimal decap configurations of the three strategies are obtained as shown in Figure 30. For the LD only or the LD&GD strategies, the decap configuration is fixed at the two supply voltages. The LD&GD&RD strategy provides flexible decap configurations for two supply voltages. The total reroutable decaps in the design is 18nf. At the low voltage level ($V_{DD} = 0.6$V), these reroutable decaps work as 4nf regular RDs and 14nf global RDs. Regular RDs act as local decaps when the local grid is active and act as global decaps when the local grid is idle. Global RDs are connected to the global VDD grid no matter the local grid is active or idle. At the high voltage level ($V_{DD} = 1V$), all reroutable decaps are used as regular RDs to enhance the suppression of supply noises.

Supply noises and leakage saving of the LD only strategy are shown in Figure 31. For the LD only strategy, the amount of local decaps is determined by the switching noise at high V_{DD}. Hence, the supply noises meet the power integrity specification at $V_{DD} = 1V$. On the other hand, the total supply noise is much smaller than the maximum tolerable

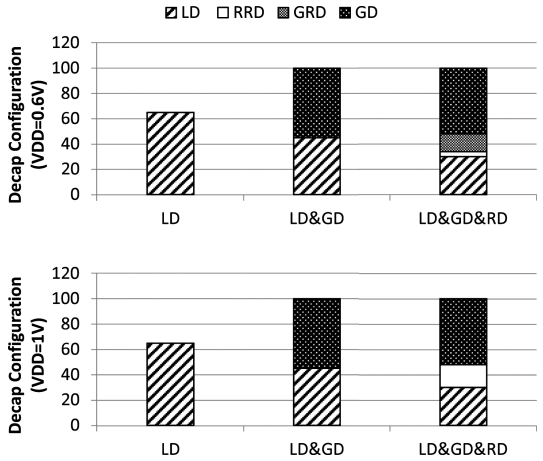


Fig. 30. Decap configurations with two supply voltages. The total decap budget is 100 nf.

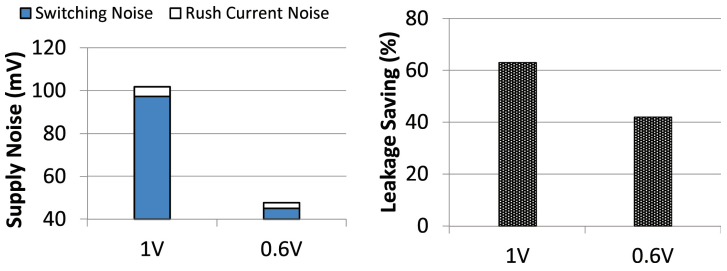


Fig. 31. Supply noises and leakage saving of the LD only strategy.

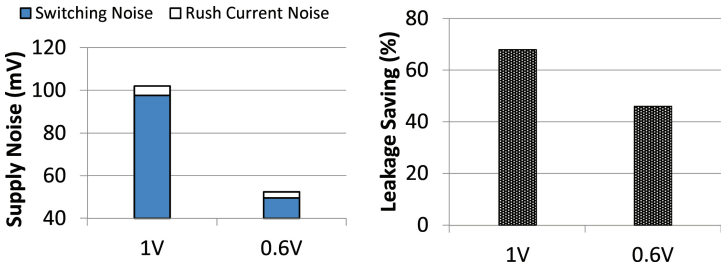


Fig. 32. Supply noises and leakage saving of the LD&GD strategy.

voltage drop ($10\%V_{DD}$) at $V_{DD} = 0.6V$. Although this result is advantageous to power integrity, it indicates that parts of the local decaps are unnecessary. These unnecessary local decaps increase rush current noise that impairs the leakage saving. As a result, the power gating at low V_{DD} only saves 40% of leakage consumption.

Supply noises and the leakage saving of the LD&GD strategy are shown in Figure 32. This strategy is similar to the LD only strategy of which the decap configuration is fixed. As a result, the leakage saving at low V_{DD} is still limited by a large amount of local decaps that is unnecessary at low voltage level.

Figure 33 shows the supply noises and leakage saving with LD&GD&RD strategy. The LD&GD&RD Strategy provides different decap configurations for two supply voltages. At the high voltage level, all the reroutable decaps are used as regular RDs to

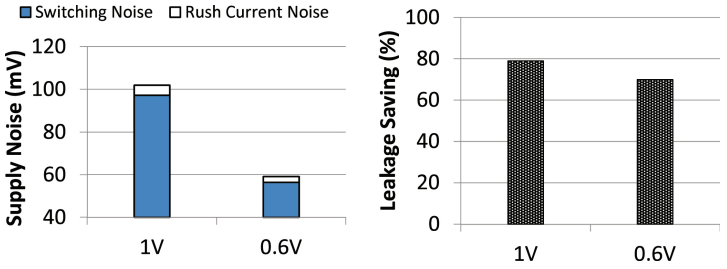


Fig. 33. Supply noises and leakage saving of the LD&GD&RD strategy.

Table VIII. Comparison between Different Strategies

	LD only		LD&GD		LD&GD&RD	
Strategy	0.6V	1.0V	0.6V	1.0V	0.6V	1.0V
L. Decaps(nf)	65	65	45	45	30	30
G. Decaps(nf)	0	0	55	55	52	52
Regular RDs(nf)	0	0	0	0	4	18
Global RDs(nf)	0	0	0	0	14	0
S. Noise(mV)	45.0	97.3	49.6	97.6	56.4	97.2
R. Noise(mV)	2.7	4.5	2.8	4.4	2.7	4.7
Tot. Noise(mV)	47.7	101.8	52.4	102	59.1	101.9
t_{on}(ns)	850	1000	700	800	400	450
Leak. Saving[1]	42%	63%	46%	68%	70%	79%
Decap Area[2]	65%		100%		105%	
Runtime (hour)	N.A.		18		31	

[1]Normalized to the leakage power consumed without power gating.
[2]Normalized to the area of total decap budget (100nf).

make sure that supply noises meet the power integrity specification. As V_{DD} decreases, parts of reroutable decaps are used as global RDs to suppress the rush current noise. Hence, as shown in Figure 33, the turn-on time is further shortened and thereby the leakage saving increases to 70%.

As shown in Table VIII, the decap area of LD only strategy takes up 65% of the area of decap budget (100nf). This is because that local decaps may increase the rush current noise. The decap budget is fully used for both LD&GD strategy and LD&GD&RD strategy. Compared with the LD&GD strategy, the LD&GD&RD strategy consumes 5% more area due to the switches of re-routable decaps. At higher VDD level, the power consumption of the control circuit is 5% of the total power consumed during idle time. The average delay of the buffer chain is 9ns. At lower VDD level, the power consumption of the control circuit is 2% of the total power consumed during idle time. The average delay of the buffer chain is 7.8ns.

8. CONCLUSIONS

In this article, on-chip decaps design strategies are proposed to deal with power-gated PDN design dilemma between power integrity and power efficiency and balance between two kinds of noises. Reroutable decaps are exploited to relax the tight interaction between supply noises and leakage saving. A special case for the PDN with multiple supply voltages is discussed. The LD&GD&RD strategy provides flexible decap configurations for different supply voltages. A simulation-based optimization flow is utilized to design PDNs through proposed strategies . The experimental results have shown that leakage saving is increased by 30% based upon the proposed methodology compared

with conventional PDN design with single supply voltage. For a PDN with two supply voltages, flexible decap configurations provided by the proposed techniques allow the optimal performance to be achieved at each voltage level.

REFERENCES

Kanak Agarwal, Kevin Nowka, Harmander Deogun, and Dennis Sylvester. 2006. Power gating with multiple sleep modes. In *Proceedings of the 7th International Symposium on Quality Electronic Design*. IEEE, 633–637.

Christian Bienia. 2011. Benchmarking modern multiprocessors. Ph.D. Dissertation. Princeton University.

B. Calhoun and A. Chandrakasan. 2003. Standby voltage scaling for reduced power. In *Proceedings of the IEEE Custom Integrated Circuits Conference*. IEEE, 639–642.

Shi-Hao Chen and Jiing-Yuan Lin. 2009. Implementation and verification practices of DVFS and power gating. In *Proceedings of the International Symposium on VLSI Design, Automation and Test (VLSI-DAT'09)*. IEEE, 19–22.

Sin-Yu Chen, Rung-Bin Lin, Hui-Hsiang Tung, and Kuen-Wey Lin. 2010. Power gating design for standard-cell-like structured ASICs. In *Proceedings of the Conference on Design, Automation and Test in Europe*. European Design and Automation Association, 514–519.

Hadi Esmaeilzadeh, Emily Blem, Renee St Amant, Karthikeyan Sankaralingam, and Doug Burger. 2012. Dark silicon and the end of multicore scaling. *IEEE Micro* 32, 3, 122–134.

Zhou Feng and Peng Li. 2008. Multigrid on GPU: Tackling power grid analysis on parallel SIMT platforms. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD'08)*. 647–654.

Joshua D. Griffin, Tamara G. Kolda, and Robert Michael Lewis. 2008. Asynchronous parallel generating set search for linearly constrained optimization. *SIAM J. Sci. Comput.* 30, 4, 1892–1924.

Meeta S. Gupta, Jarod L. Oatley, Russ Joseph, Gu-YeonWei, and David M. Brooks. 2007. Understanding voltage variations in chip multiprocessors using a distributed power-delivery network. In *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE'07)*. IEEE, 1–6.

Zhigang Hu, Alper Buyuktosunoglu, Viji Srinivasan, Victor Zyuban, Hans Jacobson, and Pradip Bose. 2004. Microarchitectural techniques for power gating of execution units. In *Proceedings of the International Symposium on Low Power Electronics and Design*. ACM, 32–37.

Intel. 2008. First the tick, now the tock: Next generation Intel microarchitecture (Nehalem). Intel Whitepaper.http://www.intel.com/technology/architecture-silicon/next-gen/whitepaper.pdf.

Intel. 2013. Mobile 4th gen Intel core processor family: Datasheet, Vol. 1. Intel Whitepaper. http://www.intel.com/content/www/us/en/processors/core/4th-gen-core-family-mobile-m-h-processor-lines-vol-1-datasheet.html.

ITRS. 2013. The International Technlogy Roadmap for Semiconductors 2013 Edition. http://public.itrs.net/. (2013).

Hailin Jiang, Malgorzata Marek-Sadowska, and Sani R. Nassif. 2005. Benefits and costs of power-gating technique. In *Proceedings of the IEEE International Conference on VLSI in Computers and Processors*. IEEE, 559–566.

Ken-ichi Kawasaki, Tetsuyoshi Shiota, Koichi Nakayama, and Atsuki Inoue. 2008. A sub-ms wake-up time power gating technique with bypass power line for rush current support. In *Proceedings of the IEEE Symposium on VLSI Circuits*. IEEE, 146–147.

Suhwan Kim, Stephen V. Kosonocky, and Daniel R. Knebel. 2003. Understanding and minimizing ground bounce during mode transition of power gating structures. In *Proceedings of the International Symposium on Low Power Electronics and Design*. ACM, 22–25.

Joseph N. Kozhaya, Sani R. Nassif, and Farid N. Najm. 2002. A multigrid-like technique for power grid analysis. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst*. 21, 10, 1148–1160.

Suming Lai, Boyuan Yan, and Peng Li. 2012. Stability assurance and design optimization of large power delivery networks with multiple on-chip voltage regulators. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD'12)*. 247–254.

Jacob Leverich, Matteo Monchiero, Vanish Talwar, Parthasarathy Ranganathan, and Christos Kozyrakis. 2009. Power management of datacenter workloads using per-core power gating. *Computer Architecture Lett.* 8, 2, 48–51.

Sani R. Nassif. 2008. Power grid analysis benchmarks. In *Proceedings of the Asia and South Pacific Design Automation Conference*. IEEE, 376–381.

Harmander Singh, Kanak Agarwal, Dennis Sylvester, and Kevin J. Nowka. 2007. Enhanced leakage reduction techniques using intermediate strength power gating. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst*. 15, 11, 1215–1224.

Haihua Su, Sachin S. Sapatnekar, and Sani R. Nassif. 2003. Optimal decoupling capacitor sizing and placement for standard-cell layout designs. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst*. 22, 4, 428–436.

Michael B. Taylor. 2012. Is dark silicon useful? Harnessing the four horsemen of the coming dark silicon apocalypse. In *Proceedings of the 49th Annual Design Automation Conference*. ACM, 1131–1136.

Tong Xu and Peng Li. 2012. Design and optimization of power gating for DVFS applications. In *Proceedings of the 13th International Symposium on Quality Electronic Design*. 391–397.

Tong Xu, Peng Li, and Boyuan Yan. 2011. Decoupling for power gating: Sources of power noise and design strategies. In *Proceedings of the 48th ACM/EDAC/IEEE Design Automation Conference*. 1002–1007.

Zhiyu Zeng, Zhou Feng, Peng Li, and Vivek Sarin. 2011. Locality-driven parallel static analysis for power delivery networks. *ACM Trans. Des. Autom. Electron. Syst.* 16, 3, Article 28.

Min Zhao, Rajendran Panda, Ben Reschke, Yuhong Fu, Trudi Mewett, Sri Chandrasekaran, Savithri Sundareswaran, and Shu Yan. 2007. On-chip decoupling capacitance and P/G wire co-optimization for dynamic noise. In *Proceedings of the 44th ACM/IEEE Design Automation Conference (DAC'07)*. IEEE, 162–167.

Shiyou Zhao, Kaushik Roy, and Cheng-Kok Koh. 2002. Decoupling capacitance allocation and its application to power-supply noise-aware floorplanning. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst* 21, 1, 81–92.

Received May 2014; revised November 2014; accepted December 2014