

# Gated Decap: Gate Leakage Control of On-chip Decoupling Capacitors in Scaled Technologies

Yiran Chen, Hai Li\*, Kaushik Roy and Cheng-Kok Koh

Dept. of ECE, Purdue University  
1285 EE Bldg, West Lafayette, IN 47906, USA  
[yc, kaushik, chengkok}@ecn.purdue.edu](mailto:{yc, kaushik, chengkok}@ecn.purdue.edu)

\*Building BB, Qualcomm Inc.  
5775 Morehouse Drive, San Diego, CA 92121, USA  
[hail@qualcomm.com](mailto:hail@qualcomm.com)

## Abstract

A novel on-chip Decoupling Capacitor (Decap) design – Gated Decoupling Capacitor (GDecap) – is proposed to minimize the leakage power dissipation associated with present-day on-chip decoupling capacitors. Experiments on the application of GDecap in an 8-way clock-gated cluster pipeline show that on average, 41.7% Decap leakage power is improved, with only 0.037% worst-case performance degradation, at 70nm technology node. Around 5.36% area overhead in Decap area is incurred, compared to the conventional Decap deployment.

## 1. Introduction

With technology scaling,  $IR$  drop and  $Ldi/dt$  noise in power supply network become crucial issues in robust VLSI design [1]. Newly proposed power management techniques, such as clock-gating, supply-gating, dynamic scaling of supply voltage ( $V_{DD}$ ) and transistor threshold voltage ( $V_T$ ) aggravate power supply noise problem because of large current spikes when system switches back and forth between active and power saving modes.

On-chip MOS decoupling capacitors (Decap) are extensively adopted to minimize the power supply noise in VLSI design. However, with the technology scaling, the requirement of Decap becomes significant: in modern high-performance processors, MOS Decap's occupy 15% to 20% of the total chip area [2]. MOS Decap also introduces large leakage power consumption in scaled technologies: for example, 26W of possible Decap leakage power (about 12% of total power) has been reported in a high-performance microprocessor design [3]. Furthermore, the gate leakage current of MOS gate-oxide capacitor will increase exponentially following the reduction in gate-oxide thickness [2][4] at the future technology nodes.

To minimize the leakage power of on-chip MOS Decap, the Decap with thick gate-oxide can take the place of the Decap with thin gate-oxide [3]. However, as the authors pointed out, such a method is applicable only when the area constraint is loose and incurs extra manufacture cost.

One can notice that the Decap requirement dynamically changes over time in power-managed systems. For example, except for the static leakage current, a clock-gated (or a supply-gated) functional unit (FU) does not draw any peak current from power supply network when it is in the power saving mode (idle). Decap's assigned to idle FU's unnecessarily consume the leakage power. Hence, if Decap could somehow be *gated*, the Decap leakage power dissipation during power saving mode could be improved.

In this paper, we propose a novel on-chip Decap design named Gated Decoupling Capacitor (GDecap). *GDecap aims at reducing the unnecessary leakage power dissipation when the system or the sub-systems are idle (for example, during power saving mode).*

The rest of our paper is organized as follows: Section 2 illustrates the basic idea of GDecap design; Section 3 discusses the design methodologies of GDecap; Section 4 analyzes the requirements of GDecap application; Section 5 shows the experiments on the GDecap application in a clock-gated 8-way cluster pipeline and Section 6 concludes our work.

\*This research is partially supported by SRC Contract 1078.001 & 1122.001

## 2. Fundamentals of Gated Decap Design

### 2.1 GDecap structure

Fig. 1(a) shows the proposed GDecap design: A control transistor M2 is inserted between the MOS Decap M1 and Ground plane.  $I_{G1}$  and  $I_{G2}$  are the gate leakage currents of M1 and M2, respectively. Fig. 1(b) shows the conventional Decap design.

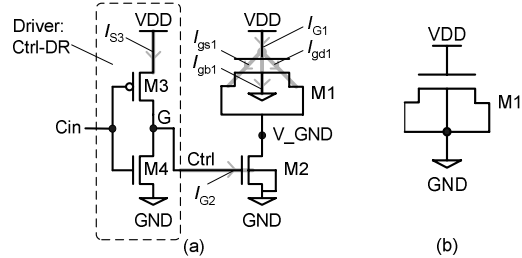


Fig. 1 GDecap structure (a) GDecap design with control scheme (b) Conventional Decap design

As the discussions in [4],  $I_{G1}$  includes three major components: gate-to-source current  $I_{gs1}$ , gate-to-drain current  $I_{gd1}$  and gate-to-substrate current  $I_{gb1}$ . Under the normal biasing condition of Decap ( $Cin = \text{logic ZERO}$ ):  $|V_{gs1}|$  and  $|V_{gd1}|$  are high,  $I_{gd1}$  and  $I_{gs1}$  dominate the total gate leakage current.  $I_{gb1}$  can be negligible.

The control scheme of GDecap is also shown in Fig. 1(a): Based on input  $Cin$ , which is equal to logic ONE in power savings mode, signal  $Ctrl$  is generated by control signal driver  $Ctrl-DR$  to drive M2. Here  $I_{S3}$  denotes the current flowing through the source of transistor M3 in  $Ctrl-DR$ .

### 2.2 Active and power saving modes of GDecap

#### A. Active mode

The active mode of GDecap is defined when the control transistor M2 is on (GDecap is un-gated and  $Cin$  is logic ZERO). The voltage at node  $V\_GND$  (see Fig. 1(a)) is pulled down to Ground. GDecap functions as a conventional Decap. The main contributors to the total GDecap leakage power, which is denoted by  $P_{un-gated}$ , are the gate leakage powers of M1 and M2, i.e.,  $P_{un-gated} = V_{DD} \cdot I_{G1} + V_{DD} \cdot I_{G2}$  ( $I_{G2} \approx I_{S3}$ ). In Section 3, we will show that the size of M2 is much smaller than M1 and the total leakage power of a GDecap mainly contributed by  $I_{G1}$ . As a result,  $P_{un-gated}$  is only slightly higher than the leakage power of a conventional Decap, denoted as  $V_{DD} \cdot I_{G1}$ .

#### B. Power saving mode

The power saving mode of GDecap is defined when the control transistor M2 is off (GDecap is gated and  $Cin$  is logic ONE). The leakage power of a gated GDecap, denoted by  $P_{gated}$ , is  $P_{gated} = V_{DD} \cdot I_{G1}$ . Due to the "stacking effect", the voltage at node  $V\_GND$  increases to a saturated voltage level that is determined by the equivalent series resistances (ESRs) of M1 and M2. The potential drop across the gate oxide of M1 decreases with the reductions of  $V_{gs1}$  and  $V_{gd1}$ . As a result, the M1's gate leakage current  $I_{G1}$ , including the gate-to-source/drain currents  $I_{gs1}$  and  $I_{gd1}$ , reduces exponentially [4].

## 3. Design Methodologies of Gated Decap

In this section, we use an example with BPTM 70nm technology [5] to illustrate the GDecap design methodologies. Based on [6],

we set  $V_{DD}$  at 1.1V and the power supply noise threshold as 10% of  $V_{DD}$  (or 0.11V). The power mesh model in [1][3] and the gate leakage model in [4] are adopted in our simulation. Every power grid node supplies current to a  $20 \times 20$  sq.  $\mu\text{m}$  area (say) where the current demand of FU is simulated as a triangular current waveform with the duration of half clock cycle (167ps, or 6GHz [6]). All simulations are conducted by using HSPICE. For more details of the simulation setup, we refer readers to [1] and [3].

As the Decap design guidance in [7][8], Decap's are usually designed as an array of single MOS capacitor where the channel length of a single Decap cell is set to about 10 times the minimal channel length, or 700nm at 70nm technology node. The column "Conv. Decap" in Table 1 shows the parameters of conventional Decap's that are required by the power supply node supplying current to a  $20 \times 20$  sq.  $\mu\text{m}$  area, in order to keep the power supply noise under the threshold (10% of  $V_{DD}$ ). Here the channel width of each Decap is set to 968.75nm, based on the Decap design rule in [7][8].

**Table 1 Conventional Decap and GDecap parameters**

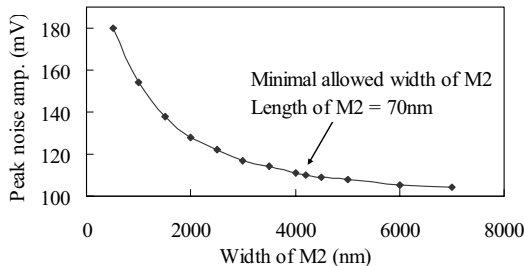
		Conv. Decap	GDecap
M1	Total Width	11625nm	11625nm
	Length	700nm	700nm
M2	Width	N/A	4200nm
	Length	N/A	70nm
Conv. Decap or Un-gated GDecap	$I_{G1}$	23.479 $\mu\text{A}$	23.203 $\mu\text{A}$
	$I_{S3}$	N/A	2.658 $\mu\text{A}$
	Power	25.827 $\mu\text{W}$	26.254 $\mu\text{W}$
Gated GDecap	$I_{G1}$	N/A	137.71nA
	$I_{S3}$	N/A	11.316nA
	Power	N/A	154.67nW
Area*		41.61 $\mu\text{m}^2$	44.43 $\mu\text{m}^2$
Area overhead			6.78%
leakage power penalty (un-gated)			1.65%
Leakage power saving (gated)			99.40%

\*Based on the estimation under 90nm technology.

### 3.1 Sizing-up of the control transistor M2

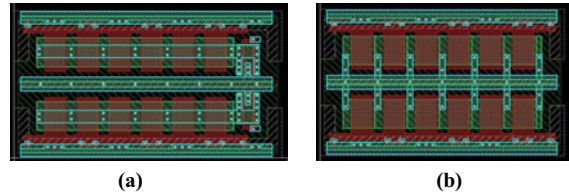
The additional ESR of control transistor M2 in GDecap degrades the effectiveness of Decap. However, the gate capacitance of M2 also contributes to the total capacitance of GDecap and enhances the effectiveness of GDecap. Sizing-up M2 can reduce the ESR of M2 as well as increase the capacitance of GDecap. Fig. 2 shows the amplitude of the minimized power supply noise under different channel widths of M2 with a 70nm channel length. Our simulations show that a channel width of 4200nm (with a 70nm channel length) is *sufficient* to maintain the effectiveness of the GDecap in power supply noise reduction while improving leakage power dissipation. The corresponding saturated voltage level at  $V_{GND}$  is 688.5mV. Increasing the size of M2 will degrade the stacking effect and increase the gate leakage currents  $I_{G1}$  and  $I_{G2}$  in the power saving mode.

Because the 70nm layout design rule is not available yet, we used the layout implementation with the closest technology: TSMC



**Fig. 2 Power supply noise amplitude under different sizes of M2**

90nm technology, to estimate the area overhead of GDecap. The dimensions of devices have been converted from 70nm Technology to 90nm Technology. The layout implementations of conventional Decap and GDecap in Fig. 3 show a 6.78% area overhead of GDecap over conventional Decap design.



**Fig. 3 Layout implementation of (a) GDecap (b) Conventional Decap**

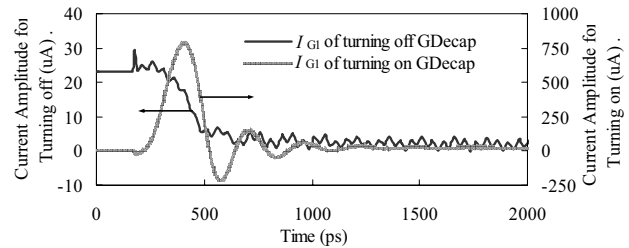
The area, the leakage currents and the leakage power dissipations of a GDecap in active and power saving modes, are shown in the column "GDecap" in Table 1. Note the power dissipation of Ctrl-DR has been considered (see more details in Section 3.3).

### 3.2 Transient analysis of GDecap

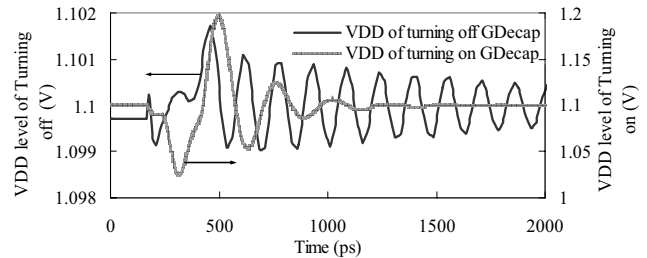
In this section, we roughly discuss the transients of power supply network when gating and un-gating GDecap. A simple scheme to minimize the transient in power supply network is also proposed.

#### A. Turning off (gating) GDecap

When M2 is turned off, the gate leakage current  $I_{G1}$  reduces quickly, as shown in Fig. 4. Here the direction of  $I_{G1}$  is defined in Fig. 1(a). The induced VDD oscillation, which is shown in Fig. 5, lasts for a long time after GDecap's have been turned off. Fortunately, due to the small amplitude of  $I_{G1}$ , the maximal induced power supply voltage droop is normally very small.



**Fig. 4 Current transients of GDecap**



**Fig. 5 Voltage transients of GDecap**

After turning-off M2, the voltage at  $V_{GND}$  slowly approaches the saturated voltage level. In this period, the charge on the gate of M1 is slowly released by the gate leakage current  $I_{G1}$  until the GDecap reaches the steady state of power saving mode.

#### B. Turning on (un-gating) GDecap

When M2 is turned on, the charge loss of the gate of M1 during the power saving mode is compensated by  $I_{G1}$  in a short time. The corresponding transient amplitudes of  $I_{G1}$  and VDD oscillation are much higher than those in turning off GDecap, as shown in Fig. 4 and Fig. 5 respectively. Fortunately again, the damping time of the VDD oscillation is much shorter than the one in the transient of turning off GDecap because the active (un-gated) GDecap's function as conventional Decap's.

### 3.3. Minimization of GDecap transients

To ensure the correct functionality when the circuits are not in the power saving mode, the maximum amplitude of the GDecap transients must be limited under the power supply noise threshold (say, 10% of  $V_{DD}$ ). Hence, when turning on GDecap, the signal Ctrl should be slow enough to smooth the spike of  $I_{G1}$  and minimize the induced VDD droop. However, Ctrl cannot be too slow since it also prolongs the GDecap turning-on/-off transient.

Ctrl can be slowed down by reducing the size or increasing the load capacitance of the control signal driver Ctrl-DR. Fig. 6 shows the maximal VDD droops under different driver sizes (1X-4X minimal size) and fanouts (1-8). The parameters of minimal size Ctrl-DR and the fanout of Ctrl-DR (the control transistor M2) are shown in Table 2 and Table 1, respectively.

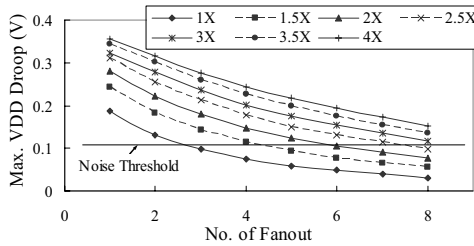


Fig. 6 Max. VDD droops with diff. driver sizes and fanout numbers

Table 2 Parameters of M3 and M4 in Ctrl-DR

	Length (nm)	Width (nm)
M3	70	340
M4	70	170

Any configurations whose maximum VDD droop under 0.11V (10% of  $V_{DD}$ ) can be considered as the control scheme of GDecap. Our design adopts a minimal size Ctrl-DR (Table 2) with a fanout of 4 control transistors. In the floorplan, the Ctrl-DR could be located at the center of (say), the  $20 \times 20$  sq.  $\mu\text{m}$  area that we experimented with, and drive the control transistors of the GDecap's located at the four corners of the square. We have considered a  $20\mu\text{m}$  interconnect between Ctrl-DR and control transistor M2 in our simulation.

To be conservative, the GDecap turning-off and turning-on transient periods (denoted by  $T_{\text{turn-off}}$  and  $T_{\text{turn-on}}$ , respectively) can be defined to start when Cin begins to switch and end when the amplitude of VDD fluctuation is less than 0.5% of  $V_{DD}$ . Under the control scheme adopted in our simulation, the transient periods of turning-on and turning-off are 1.333 ns ( $T_{\text{turn-on}}$ ) and 1.833 ns ( $T_{\text{turn-off}}$ ), respectively. To ensure the correct functionality of FU, GDecap may be required to be un-gated about 1.333ns (or 8 cycles for 6GHz operation) before the corresponding idle FU's switch back to the active mode.

## 4. Application of GDecap

In general, the application of GDecap must satisfy two conditions: First, the energy saving achieved by GDecap in the power saving mode must exceed the energy overheads of GDecap in active mode and turning-on/turning-off transients; Second, the switching (between active and power saving modes) of GDecap should not affect the functionality of the circuits that are in active mode.

### 4.1. Energy overheads of GDecap transients

The energy overhead of GDecap in active mode is only 1.65% (Table 1). Here we analyzed only the energy overheads of GDecap transients. The energy overhead  $E_{\text{turn-off}}$  of GDecap turning-off transient is defined as  $E_{\text{turn-off}} = (E_{1 \rightarrow 0} - P_{\text{gated}} \cdot T_{\text{turn-off}})$ .  $E_{1 \rightarrow 0}$  is the total energy consumed by a GDecap during turning-off transient period  $T_{\text{turn-off}}$ . Similarly, the energy overhead  $E_{\text{turn-on}}$

of the GDecap turning-on transient is defined as  $E_{\text{turn-on}} = (E_{0 \rightarrow 1} - P_{\text{un-gated}} \cdot T_{\text{turn-on}})$ .  $E_{0 \rightarrow 1}$  is the total energy consumed by a GDecap during turning-on transient period  $T_{\text{turn-on}}$ .

The turning-on energy overhead  $E_{\text{turn-on}}$  heavily depends on the time duration for which a GDecap is gated (gated period): the longer a GDecap is gated, the higher  $E_{\text{turn-on}}$  is, because the charge compensation to the gate of M1 is higher. Fig. 7 shows how the turning-on energy overhead  $E_{\text{turn-on}}$  varies as the number of clock cycles a GDecap stays gated increases. The node " $\infty$ " indicates the result when a GDecap has been gated for a sufficiently long time: 128.33fJ. In our simulation, the rising/falling time of Cin is set at 16.7ps, which is 1/10 of one cycle for 6GHz operation.

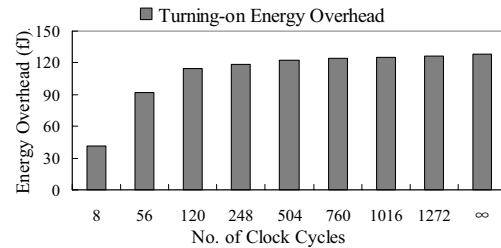


Fig. 7 Turn-on energy overhead with various gated periods

Our simulation also show that the turning-off energy overhead  $E_{\text{turn-off}}$  is insensitive to the time duration for which a GDecap is un-gated. To turn off a GDecap that has stayed un-gated for a sufficiently long time,  $E_{\text{turn-off}}$  is 9.21fJ.

### 4.2 Minimal gated time period and minimal gated ratio

Fig. 8 shows the energy overhead, which includes both  $E_{\text{turn-off}}$  and  $E_{\text{turn-on}}$ , amortized over various gated periods (clock cycles). The plot also depicts the absolute leakage energy saving, which is the difference between the energies consumed by a conventional Decap and a gated GDecap during one clock cycle. The net leakage energy saving is derived by subtracting the amortized energy overhead from the absolute leakage energy saving. The net leakage energy saving ratio is the ratio between net leakage energy saving and the energy consumed by a conventional Decap during one clock cycle. Here node " $\infty$ " denotes the case when GDecap is gated for infinitely long time.

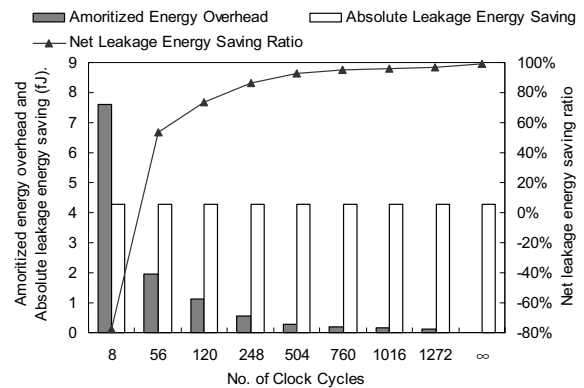


Fig. 8 Amortized energy overhead, absolute leakage energy saving and net leakage energy saving ratio of a GDecap

The gated period of the GDecap must be long enough to achieve a power saving higher than the turning-on/-off energy overheads. Simulation shows that the minimal gated period of GDecap for positive power saving is around 2000ps for 70nm technology.

The gated ratio of GDecap is defined as the ratio of the gating time and the total execution time. Because of the leakage power overhead of GDecap in the active mode (1.65%) and the turning-on/turning-off energy overheads, a minimal gated ratio is required

for positive power saving of GDecap. In general, the minimal gated ratio is very small: when average gated period is 56 clock cycles, the minimal gated ratio for positive power saving is only 3.0%. With the increase of average gated period, the minimal gated ratio will further decrease.

### 4.3 The “isolation gap” in floorplan

Due to the RC delay along the power supply network, Decap must be deployed close to the active FU to effectively reduce power supply noise [9]. We define the Decap effective distance  $D_{\text{eff}}$  (from the current source) as the minimum distance between GDecap and current source such that: when all GDecap's located beyond that distance are gated, the increase of power supply droop associated with that current source is minimal (say, 0.5% of  $V_{\text{DD}}$ ). Note the negligible difference between the power supply noise when no GDecap is gated and when GDecap's beyond  $D_{\text{eff}}$  are gated in Fig. 9). Hence, as long as all FU's located within the effective distance of a GDecap switch to power saving mode, the GDecap can be *practically* turned off without affecting the functionality of the rest of active FUs. Fig. 10 shows the increased amplitudes of VDD droop when all GDecap's located at various distances away from current source are gated. For 70nm technology node, the effective distance is 120 $\mu\text{m}$ .

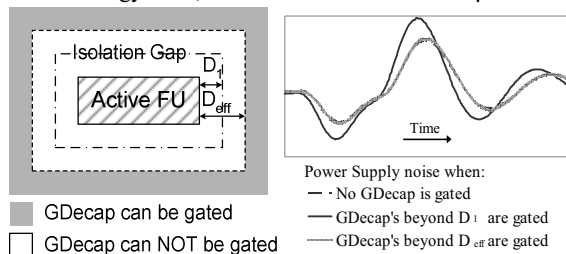


Fig. 9 Effective distance and Isolation gap in the floorplan

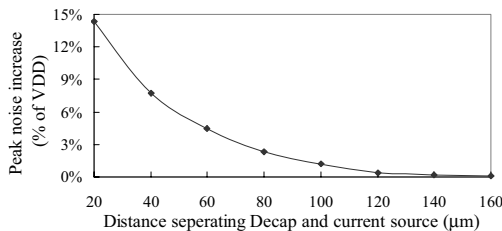


Fig. 10 Effective distance of Decap at 70nm technology node

To effectively minimize the power supply noise associated with the active FU's, an “isolation gap” is reserved between the borders of active FU and the gated GDecap, as shown in Fig. 9. The width of isolation gap is the effective distance of GDecap.

## 5. GDecap application in PLB architecture

Pipeline balancing (PLB) is a clock-gating scheme for multi-issue cluster pipeline [10]. For example, in an 8-way clustered PLB pipeline, each single pipeline has its own FU and instruction issue logic. An IPC (instructions per cycle) monitor counts the IPC for every 256-cycle period, which is called a “window”. Based on the IPC in the current window, the IPC required by the program in the next window is predicted. As a result, the FU's, issue logics and the corresponding latches of 4, 2 or no pipelines are clock-gated in the next window to save dynamic power. Consequently, the issue width of pipeline is shaped to be 4, 6 or 8, respectively.

To apply GDecap to PLB pipeline, we use the IPC of the first 248 cycles in a window, instead of the IPC over total 256 cycles, to predict the IPC in the next window. To ensure the functionality of FU's, the rest 8 cycles (see Section 3.3) are used to turn on GDecap before un-gating the corresponding pipeline. The signal  $C_{\text{in}}$  can be borrowed from the clock-gating signals and no extra

global signal wire is needed. The control signal of Ctrl-DR is delayed by 8 cycles to un-gate FU's. The delay logic is very simple (i.e., 3-bit counter) and the switching activities of the clock-gating signals are very infrequent. Hence, the routing congestion and the power/area overheads introduced by GDecap control scheme are negligible.

The modified SimpleScalar tool [11] is used in our simulations. The dimensions of components in each pipeline are scaled from Alpha21264B, which was fabricated with 0.18 $\mu\text{m}$  technology [12]. With the consideration of isolation gap in the floorplan, the Decap area overhead introduced by the GDecap technique is only 5.36%. Among 23 SPEC2000 benchmarks, on average, GDecap technique achieves 41.7% Decap leakage power reduction in the 4 clock-gated pipelines. The worst-case IPC-based performance loss is only 0.037%, compared to original PLB technique in [10].

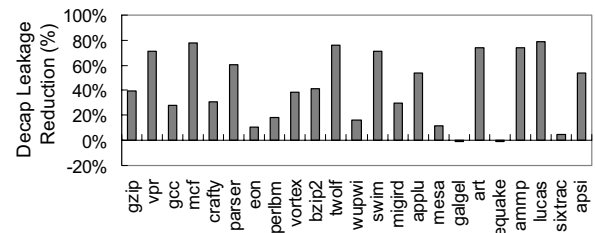


Fig. 11 Decap leakage power savings in PLB pipeline

## 6. Conclusion

We proposed a Gated Decoupling Capacitor (GDecap) Design technique to reduce the leakage power dissipation of on-chip decoupling capacitor (Decap). When functional units (FU's) are idle, the corresponding GDecap's can be turned off to achieve low gate leakage current without undue impact on the system performance. Experiments on the GDecap application in an 8-way PLB microprocessor show that GDecap technique can effectively minimize the unnecessary Decap leakage power with marginal area overhead and negligible performance loss.

## References

- [1] H. H. Chen and J. S. Neely, Interconnect and circuit modeling techniques for full-Chip power supply noise analysis. *IEEE Trans. Compon., Pack. and Manuf. Tech.*, Part B, vol. 21, pp. 209-215, Aug. 1998.
- [2] M. K. Gowan, L.L. Biro and D. B. Jackson, Power considerations in the design of the Alpha 21264 microprocessor, *Proc. of Design Automation Conf.*, pp. 726-731, Jun. 1998.
- [3] H. H. Chen, J. S. Neely, M. F. Wang and G. Co, On-chip Decoupling capacitor Optimization for Noise and Leakage Reduction, *Proc. of the 16<sup>th</sup> Symp. on Integ. Ckts. and Sys. Design*, pp. 251-255, Sept. 2003.
- [4] S. Mukhopadhyay, C. Neau, R. T. Cakici, A. Agarwal, C. H. Kim and K. Roy, Gate Leakage Reduction for Scaled Devices using Transistor Stacking, *IEEE Trans. on Very Large Integ. (VLSI) Sys.*, Vol. 11-4, pp. 716-730, Aug. 2003.
- [5] <http://www-device.eecs.berkeley.edu/~ptm>.
- [6] International Technology Roadmap for Semiconductor, 2003.
- [7] P. Larsson, Parasitic Resistance in an MOS Transistor Used as On-chip Decoupling Capacitance, *IEEE Jour. of Solid-State Ckts*, Vol. 32-4, pp. 574-576, Apr. 1997.
- [8] R. J. Baker, CMOS: Mixed-Signal Circuit Design, *John Wiley and Sons Publishers*, 2002.
- [9] J. Choi, et al., Modeling of Realistic On-Chip Power Grid using the FDTD Method, *Proc. of IEEE Intl. Symp. on Electromag. Compat. 2002*, Vol. 1, pp. 238-243, Aug. 2002.
- [10] R. I. Bahar and S. Manne, Power and energy reduction via pipeline balancing, *Proc. of the 28<sup>th</sup> Annual Intl. Symp. Comp. Arch.*, pp. 218-229, Jul. 2001.
- [11] D. Burger and T. M. Austin. The simplescalar tool set, version 2.0. TR1342, University of Wisconsin, Jun. 1997.
- [12] K. Skadron, et al., Temperature-Aware Computer Systems: Opportunities and Challenges, *IEEE Micro*, Vol. 23-6, pp. 52-61, Nov.-Dec. 2003.