DRG-Cache: A Data Retention Gated-Ground Cache for Low Power[1]

Amit Agarwal, Hai Li, and Kaushik Roy
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47906, USA
<amita, hl, kaushik> @ ecn.purdue.edu

ABSTRACT

In this paper we propose a novel integrated circuit and architectural level technique to reduce leakage power consumption in high performance cache memories using single Vt (transistor threshold voltage) process. We utilize the concept of Gated-Ground [5] (NMOS transistor inserted between Ground line and SRAM cell) to achieve reduction in leakage energy without significantly affecting performance. Experimental results on gated-Ground caches show that data is retained (DRG-Cache) even if the memory are put in the stand-by mode of operation. Data is restored when the gated-Ground transistor is turned on. Turning off the gated-Ground transistor in turn gives large reduction in leakage power. This technique requires no extra circuitry; row decoder itself can be used to control the gated-Ground transistor. The technique is applicable to data and instruction caches as well as different levels of cache hierarchy such as the L1, L2, or L3 caches. We fabricated a test chip in TSMC 0.25μ technology to show the data retention capability and the cell stability of DRG-cache. Our simulation results on 100nm and 70nm processes (Berkeley Predictive Technology Model) show 16.5% and 27% reduction in consumed energy in L1 cache and 50% and 47% reduction in L2 cache with less than 5% impact on execution time and within 4% increase in area overhead.

Categories and Subject Descripters

B.3.2 [**Memory Structure**]: Design Styles --- *Cache memories;* B.3.1 [**Memory Structure**]: Semiconductor Memories --- *Static memory (SRAM)*; B.7.1 [**Integrated Circuits**]: Types and Design Styles --- *Memory technology.*

General Terms: Design, Performance and Experimentation.

Keywords: Gated-ground, SRAM, low leakage cache.

1 INTRODUCTION

Semiconductor devices are aggressively scaled each technology generation to achieve high integration density while the supply voltage is scaled to achieve lower switching energy per device. However, to achieve high performance there is need for commensurate scaling of the transistor threshold voltage [6]. Scaling of transistor threshold voltage is associated with exponential increase in sub-threshold leakage current [3]. To overcome the excessive leakage problem, advanced leakage control methods will become indispensable for future technologies.

State-of-the-art microprocessor designs devote a large fraction of the chip area to memory structures — e.g., multiple levels of instruction and data caches, translation look-aside buffers, and prediction tables. For instance, 30% of Alpha 21264 and 60% of StrongARM are devoted to cache and memory structures [4]. Caches account for a large (if not dominant) component of leakage energy dissipation in recent designs, and will continue to do so in the future. Recent energy estimates for 0.13μ process technology indicate that leakage energy accounts for 30% of L1 cache energy and as much as 80% of L2 cache energy [2]. To address the power inefficiency, we propose a single Vt (transistor threshold voltage) Data Retention Gated-Ground Cache (DRG-Cache) design and architecture, in which the unused portions of the memory core are set to low leakage mode to reduce power.

Many embedded designs [1], instead of gating the cell, use circuit only techniques [7] and primarily rely on a dual-threshold voltage (dual-Vt) process technology [8,9] to reduce leakage. Dual Vt makes sense till Vdd is 1V and Vt is around 0.25v due to Vt spread [8] and also it requires extra process cost. By providing an alternative solution, our single Vt integrated circuit/architecture approach to reduce leakage for high-performance designs offers a key advantage over the dual-Vt approach.

The Dynamically Resizable (DRI) I-cache presented in [5] varies the size of the L1 cache by turning off (using gated-Ground NMOS transistor between the Ground and SRAM cell) the unused section to reduce leakage. In this work we investigate the data retention capability of the DRG-cache having gated-Ground, its circuit level implications and architectural mechanisms. Our theoretical analysis and fabricated chip results suggest that data can be retained in DRG-Caches even when the gated-Ground transistor is turned-off. We propose that the data in the cache be in the data-retenting mode (gated-Ground transistor 'off') all the time unless we are reading/writing from that cell. Our approach is easily applicable to both data and instruction L1 as well as L2 cache with very little circuit overhead.

2. DRG-CACHE: CIRCUIT, ARCHITECTURE AND LAYOUT

To prevent the leakage energy dissipation in a DRG-Cache from limiting aggressive threshold-voltage scaling, we use a circuit-level mechanism called gated-Ground [5]. Gated-Ground enables a DRG-Cache to effectively turn 'off' the supply voltage and

[1]This research was funded in part by DARPA, MARCO GSRC, SRC, and by Intel and IBM.

DAC 2002. June 10-14,2001, New Orleans, Louisiana, USA.

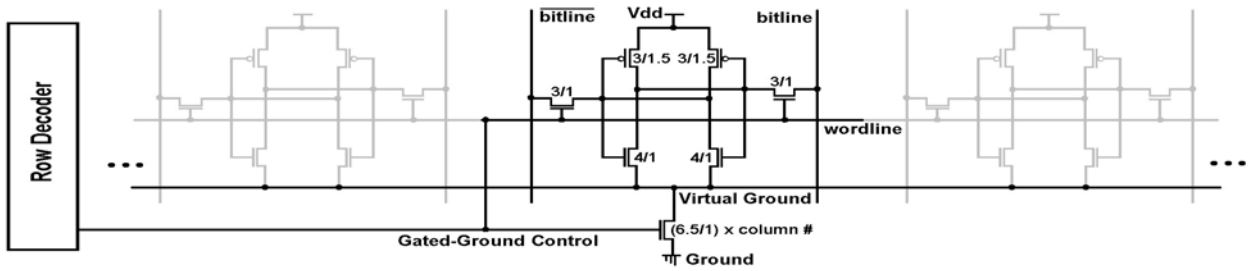

Figure 1. Anatomy of DRG-Cache : Data Retention Gated-Ground Cache.

virtually eliminate the leakage energy dissipation in the cache's unused (used section of the cache core is defined as the SRAM cells from which data is read/written) sections. The key idea is to introduce an extra NMOS transistor (Figure 1) in the leakage path from the supply voltage to the ground of the cache's SRAM cells; the extra transistor is turned 'on' in the used and turned 'off' in the unused sections, essentially "gating" the cell's supply voltage. Figure 1 shows the anatomy of DRG-Cache. Gated-Ground achieves significantly lower leakage because of the two off transistors connected in series reducing the leakage current by orders of magnitude; this effect is due to the self reverse-biasing of stacked transistors, and is called the stacking effect [10].

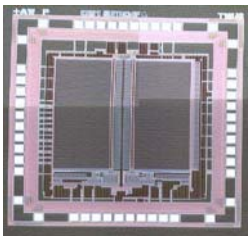

Figure 2. Die Photo of DRG-Cache.

Same as conventional gating techniques, the gated-Ground transistor can be shared among multiple SRAM cells from one or more cache blocks. It amortizes the overhead of the extra transistor. Because the size of gated-Ground transistor plays a major role in the data retention capability and stability (section 4) of the DRG-Cache, and also affects the power and performance savings (section 6), the gated-Ground transistor must be carefully sized (Figure 1) with respect to the SRAM cell transistors it is gating. While the gated-Ground transistor must be made large enough to sink the current flowing through the SRAM cells during a read/write operation in the active mode and to enhance the data retention capability of cache, too large a gated-Ground transistor may reduce the stacking effect, thereby diminishing the energy savings. Moreover, large transistors also increase the area overhead due to gating. In DRG-Cache the gated-Ground transistor is shared by a row of SRAM cells. The gated-Ground transistor is controlled by the row decoder logic of conventional SRAM. The cells are turned 'on' only when the row is being read from or when data is written into the row. However, this requires that the row decoder drives a larger gate capacitance associated with gated-Ground transistor unlike conventional caches. To maintain performance proper sizing of decoder is required.

Figure 2 shows the die photo of the fabricated DRG-cache along with a conventional cache (to the left) for comparison. The right most thin column in layout is the gated Ground transistor. To minimize the area overhead and optimize layout, we implemented gated-Ground transistor as rows of parallel transistors placed along length of the SRAM cells. This row of parallel transistor is placed at one end of row of SRAM cells. The metal line, which is used as a ground line in conventional cache, is used to connect the drain of gated-Ground transistor and the SRAM cell and acts as virtual ground. Wordline connecting each cell of a row to a row decoder is connected to the gate of the gated-Ground transistor. As it is clear from the layout, the area overhead due to gated-Ground transistor is about 4%. It is important to note that the DRG-Cache core is fully compatible with current cache design.

3. FACTORS AFFECTING THE STORED DATA IN DRG-CACHE

Conventional SRAM store the data as long as power supply is on. This is because the cell storage nodes at '0' and '1' are firmly strapped to power rails through conducting devices (by a pull-down NFET in one inverter and a pull-up PFET in the other inverter). Figure 3 shows a single cell schematic of our DRG-Cache. When the gated-Ground transistor is on, it behaves exactly like conventional SRAM in terms of storing data. Turning 'off' the gated-Ground nicely cuts-off the leakage path from the cell node that is at '1' to ground. However, it also cuts-off the opportunity to strap the cell node at '0' firmly to ground. This makes it easier for a noise source to write a '1' to that node. Node storing '1' remains firmly strapped to Vdd as long as input ($\overline{Q}$) to the pull-up PFET (M4) remains below the trip point of the inverter.

Secondly, there is an issue with data retention. If the sub-threshold current through pull-up PFET (M2) and through pass transistor (M6) is greater than sub-threshold current through gated-Ground via pull down NFET (M1) or in other words if resistance through the pull-up PFET (M2) and pass transistor (M6) be much less than the sub-threshold resistance of gated-Ground transistor, then a voltage divider may cause the node at '0' to rise high. This has a possibility of degrading the current drive of the PFET (M4) and increasing the leakage of the NFET (M3) in the inverter driven by '0' node. This case has potential, in worst case, to destroy data written into the cell and in the best case to degrade the stacking effect.

4. DATA RETENTION CAPABILITY OF DRG-CACHE

The leakage current in MOSFET depends on various process parameters, transistor size and the quiescent state of the circuit. We use leakage model in [10] for modeling our leakage current. This model uses the following simplified BSIM sub-threshold current equation.

$$I_{subth} = Ae^{\frac{q}{nkT}(V_G - V_S - V_{TH0} - \gamma' V_{SB} + \eta V_{DS})}(1 - e^{-\frac{qV_{DS}}{kT}}) \quad (1)$$

where $A = \mu_0 C'_{ox} \frac{W}{L_{eff}} (\frac{kT}{q})^2 e^{1.8}$

V_G, V_D and V_S are gate voltage, drain voltage, and source voltage of the transistor, respectively. Body effect is represented by the term $\gamma' V_S$, where γ' is the linearized body effect coefficient. η is the DIBL coefficient, representing the effect of V_{DS} on threshold voltage. C_{ox} is the gate oxide capacitance. μ_0 is the zero bias mobility and n is the sub-threshold swing coefficient of the transistor.

We do our leakage modeling at the time when gated-Ground transistor is turned 'off' and DRG-Cache is in low leakage mode (standby leakage mode). Initially '1' is written at Q and '0' at $\overline{Q}$. In the Figure 3, the light color transistors are 'on', and dark color transistors are 'off'. We assume that the voltage at $\overline{Q}$ goes to saturation after sometime and this saturation value is Vg. This Vg is small enough to keep the pull-up PFET (M4) 'on'. Because M4 remains 'on', it firmly straps the node Q to Vdd supply rail. This in turn keeps the NFET (M1) 'on'. Because M1 is 'on', voltage at virtual ground follows the voltage at $\overline{Q}$ and goes to Vg. This makes both gate voltage and source voltage of pull-down NFET (M3) at Vg, and turns 'off' M3 because V_{GS} is '0'.

From Figure 3 we observe that the transistors responsible for charging $\overline{Q}$ node are M2 and M6. For considering the worst-case condition we assume that the $\overline{bitline}$ remains at Vdd even if pre-charging is not applied for a long period of time, and the leakage associated with pass transistor M6 is maximum (Eq 1).The voltage at $\overline{Q}$ is discharged through capacitance leakage and leakage through gated-Ground transistor. Equating all the current at time of voltage saturation we get

$$I_{subM7} = I_{subM3} + I_{subM6} + I_{subM2} - I_{cap} \text{ (neglected)}$$

where, the subthreshold leakage currents are given as

$$I_{subM7} = A_7 e^{\frac{q}{nkT}(-V_{TH0} + \eta V_g)}$$

$$I_{subM3} = A_3 e^{\frac{q}{nkT}(-V_{TH0} - \gamma' V_g + \eta(V_{DD} - V_g))} \quad (2)$$

$$I_{subM6} = A_6 e^{\frac{q}{nkT}(-V_{TH0} - (1+\gamma'+\eta)V_g + \eta V_{DD})}$$

$$I_{subM2} = A_2 e^{-\frac{q}{nkT}(-V_{TH0P} + \eta(V_g - V_{DD}))}$$

We use the parameters from TSMC 0.25µ technology library and sizes of the transistors from our cache layout in our leakage model. In our case, V_g turns out to be around 0.4V. Here we have neglected the capacitance leakage. However, if we consider the capacitance leakage, it will reduce the voltage at $\overline{Q}$. This in turn, because of the feedback, affects positively in turning on transistor M4. Because Vg is small enough to turn on pull-up PFET (M4), our initial assumptions are correct. Q is strapped to V_{DD} through low resistance, and hence it always remains at '1'. Turning on M7 will restore the data by pulling $\overline{Q}$ to '0' through M3 and M7.

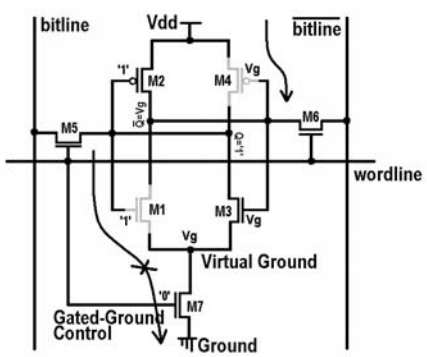

Figure 3. DRG-Cache: Data Retention Capability.

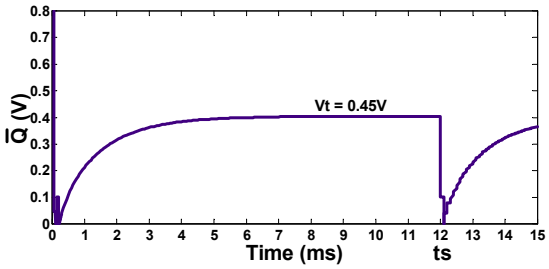

Figure 4. Voltage variation at node storing '0'.

We simulated the extracted net-list from layout of the DRG cache. After time t_s gated-Ground is again turned on. The voltage at node storing '1' always remains at Vdd. Figure 4 shows the voltage variation at $\overline{Q}$ (using TSMC 0.25µ process technology and supply voltage of 2.5V). As the gated-Ground is turned off the voltage at $\overline{Q}$ starts rising and gets saturated to 400mv. Turning on the gated-Ground transistor restores the data at $\overline{Q}$ to zero. This result shows that data is not lost even if we turn off the gated-Ground transistor for indefinite time.

Table 1 shows the simulation results for saturation voltage at node storing '0' using TSMC 0.25µ technology. Unless otherwise, mentioned the temperature for all simulations are considered as 25˚C, which is worst case for data retention. Here Gsize is the relative size of gated-Ground transistor normalized with respect to the size of NMOS M3 shown in figure 3. Vtn_0 is the zero bias threshold voltage of the NMOS transistors of SRAM cell. Vtp_0 is kept at –0.50V. Sat $\overline{Q}$ is the saturation voltage at node storing 0($\overline{Q}$).

4.1 Impact of Widening the Gated-Ground Transistor

First three rows of Table 1 indicate that the saturation voltage of node storing '0' can be made as small as possible by increasing the size of gated-Ground transistor (Gsize). Increasing the size of gated-Ground increases the discharging current for node storing '0', which reduces the saturation voltage. This also degrades the power savings and improves the performance. Hence, there is a trade off between performance, stability and power dissipation. However, increasing the size of gated-Ground only affects leakage linearly. Recall, the effect of gated-Ground on leakage is exponential. Hence, we can get higher stability without losing much in terms of leakage reduction.

Table 1. Data Retention Capability of DRG-Cache

Temp (˚C)	Gsize	Vtn_0 (V)	Sat $\overline{Q}$ (Vg) (V)	ΔVt_{max} (mV)
25	1.0	0.45	0.40	150
25	2.0	0.45	0.22	157
25	4.0	0.45	0.08	165
50	1.0	0.45	0.11	111
100	1.0	0.45	0.06	97
25	1.0	0.35	0.08	85
25	1.0	0.25	0.07	78

4.2 Impact of Temperature

The temperature of a chip can vary depending on the workload and power consumption. The leakage current increases exponentially with temperature (Eq. 1). However, this current increment is more in NFETs as compared to PFETs, because NFET parameters (for example μ, η, and γ) are stronger than PFET for contributing to leakage current. Hence, the discharging current is stronger than charging current in high temperature case. This makes Vg to saturate at a lower voltage. First, fourth and fifth rows of Table 1 indicate the impact of temperature on saturation voltage.

4.3 Impact of Lowering Vt

It is evident from the first, sixth and seventh rows of Table 1 that voltage at node storing '0' gets saturated at lower voltage with lower Vt process technology. This is due to fact that lowering the Vt increases the leakage current of all the transistors and since the discharging current is stronger than the charging current in the low Vt case, the saturation voltage Vg is lower.

4.4 Stability Analysis Under Vt Variation

We observed in the previous section that the voltage at the node storing '0' changes as we change the Vt of the cell. If this value crosses the trip point of the cell (which is close to Vdd), the cell may flip at the time of reading. In current process technology and on upcoming technology generations it is very difficult to

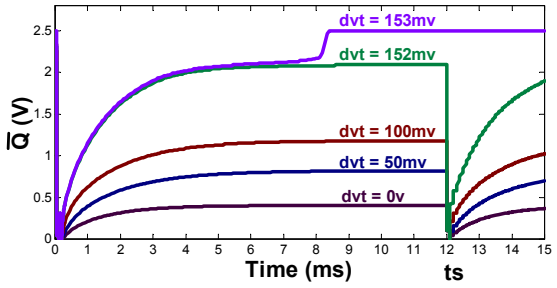

Figure 5. Stability analysis under Vt variation

fabricate transistors all having exact same Vt. The Vt mismatch may cause destructive read in the worst case. We simulated the DRG-cache to determine the effect of Vt variation on stored data. For worst-case analysis we assigned transistors M2, M3, M5 and M6 (Figure 3) low Vt (Vt - ΔVt) and transistors M1, M4 and M7 high Vt (Vt + ΔVt) transistors. We let the data to saturate for a long time (ts) and after that we try to read it (Figure 5). The saturation voltage rises as ΔV_t approaches close to 0.152V. However, we are able to read the data. Changing the ΔV_t to 0.153V flips the cell, i.e. the data cannot be read correctly from the cell ($\overline{Q}$ remains at Vdd after time ts).

The last column of Table 1 shows the limitations on V_t variation on various cases to avoid destructive read. Comparing the first row with the third and fourth row we observe that increasing the size of gated-Ground transistor has positive effect on stability. The stability of the cell is sensitive to lowering the Vt_0 of all the transistors including the gated-Ground transistor. We observe that ΔVt_{max} is more than 75mV (baseVtn_0 = 0.25V) in worst case and up to 165mv (base Vtn_0 = 0.45V) in best case.

4.5 Impact of Technology Scaling

We simulated DRG-cache using Berkeley Predictive Technology Model (BPTM) [11] for 70nm and 100nm technology after scaling our netlist. The Vdd used for these technologies are 1.0 and 1.2 V, respectively. Table 2 shows that even for these technologies, voltage at node storing '0' gets saturated to a value which is small compared to Vdd. This saturation voltage depends on supply voltage and leakage currents for respective technology.

Table 2. Impact of technology scaling on data retention capability of DRG-cache

Supply Voltage (V)	Technology	Sat $\overline{Q}$ (Vg) (V)	ΔVt_{max} (mV)
1.2	100nm	0.21	103
1.0	70nm	0.16	79

5. ENERGY AND DELAY ANALYSIS

A DRG-cache reduces leakage energy by gating ground to cache unused sections (gated-Ground turned off). However, it increases dynamic energy at the time of reading/writing due to overhead associated with extra gated-Ground transistor. The row decoder has to drive the gate capacitance associated with gated-Ground transistor, which in turn increases the delay and power.

In conventional cache, at the time of reading/writing, only one row (selected by row decoder) contributes to dynamic energy and the rest of the rows remain inactive (high temperature) leakage mode. If the cache is not accessed, all the rows leak actively. In our DRG-Cache architecture, only gated-Ground transistor associated with accessed row remains on, and all others remain off. Hence, all the rows except the accessed one remains in the standby leakage mode. When the cache is not accessed all the rows remains in standby leakage mode.

Figure 6. Breakdown of delay in cache.

Figure 6 shows the percentage contribution of all the four parts: decoder, wordline, bitline/sense amplifier, databus, in overall delay. This estimation is based on our simulation. The delay of overall cache is dominated by decoder and bitline/sense amplifier delay. The wordline contributes only 3% to overall delay.When we read from conventional cache, bitline gets discharged through pull-down NFET and pass-transistor NFET connected to node storing '0'. Putting an extra transistor (gated-Ground transistor) increases the resistance of discharging path, which in turns degrades the bitline/sense-amplifier delay in DRG-Cache compared to conventional cache. Secondly, the gated-Ground transistor is controlled by row decoder, which poses extra

overhead on wordline driver. This increases the wordline delay of DRG-Cache compared to conventional cache. However, the contribution of wordline delay to overall cache delay is very small [12,14]. Hence, overall performance of DRG-Cache does not get much affected by increase in wordline delay.

6. RESULTS

In this section, we present experimental results on the energy and performance trade-off of a DRG-Cache as compared to a conventional-cache. Due to ease of fabrication we utilize the 0.25µ technology results to validate the data retention capability, stability and performance-leakage saving trade-off in DRG-Cache. We scale down our net-list for 100nm and 70nm technology for showing energy savings achieved by DRG-Cache.

6.1 Energy Performance Trade-off

Table 3 shows the leakage savings and relative read time for different gated-Ground implementations. Here relative read time is the read time of DRG-Cache with respect to corresponding conventional cache. Conv (conventional) cache leakage and DRG-Cache leakage shown are leakage energy normalized with respect to leakage energy of conventional cache at $Vtn_0 = 0.45V$.

Table 3. Energy Performance Trade-off

Gsize	Vtn_0 (V)	Normalized Conv. Cache leakage	Normailized DRG Cache Leakage	Relative Read Time
2.0	0.45	1.0	0.68	1.028
1.0	0.45	1.0	0.54	1.046
1.0	0.35	9.0	5.2	1.044
1.0	0.25	80.8	47.8	1.044

First two rows of Table 3 and Table 1 indicate that increasing the width of the gated-Ground transistor improves the read time, data retention capability and stability of cell but decreases energy savings and increases the area. The read time improves due to increase in drive current because of large sized gated-ground transistor. However, at the same time decoder driver has to drive large capacitance due to increase in capacitance associated with gated-Ground transistor, which increases the wordline delay and dynamic energy of cache. Because wordline delay is a small fraction of overall delay the overall read time is improved. Last two rows of Table 3 indicate that lowering the threshold voltage increases the leakage energy by several orders of magnitude. The DRG-Cache leakage energy is orders of magnitude smaller than the conventional cache leakage energy. The motivation behind lowering the threshold voltage is to get high performance cache, which at the same time improves the data retention capability but degrades the stability of the cell (Table 1).

6.2 Energy Savings and Power-Delay Products for 70nm and 100nm Technology

6.2.1 Utilization Factor

We simulated simplescalar [13] for getting the information about the cache utilization factor. In the context of aggressive modern out-of-order microprocessors, which exploit instruction level parallelism [15] we assume that L1 cache is accessed in each and every cycle (conservative assumption). The idle time of L2 cache depends on both number of load instruction in a cache and miss rates of both L1 and L2 cache. L2 accesses are generated by L1 misses, so lower the miss rate of L1, higher the idle time of L2 cache. L2 cache utilization is also affected by L2 miss rate and the number of load instruction in an application. Our simulation for spec95 benchmarks show that on an average L2 cache is not accessed 80% of time. This indicates that most of the time L2 remains in the idle mode, which in turn makes leakage energy of L2 to dominate over its dynamic energy. Hence, the DRG-Cache is ideal for reducing leakage in L2 cache.

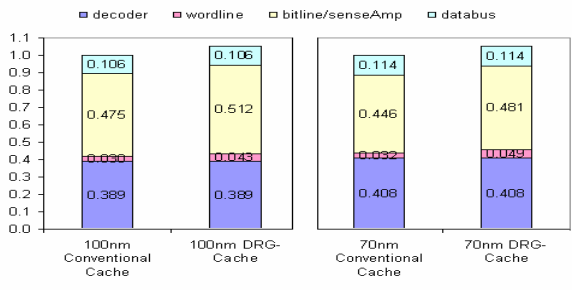

Figure 7. Performance loss in DRG-Cache.

6.2.2 Performance Loss

The DRG-Cache loses performance against conventional cache due to increase in wordline and bitline/sense amplifier delays. Figure 7 depicts the contribution of four stages of cache in overall delay. It also depicts the increase in delay in different stages with respect to conventional cache and its effect on overall delay. The graph shows that wordline delay increases by more than 43% in 100nm and 53% in 70nm technology, while the increase in bitline/sense amplifier delay is only 7.8% in both technologies. However, the contribution of wordline delay is only 3% to the overall delay causing overall performance degradation to be 5.2 – 5.3% in respective technology.

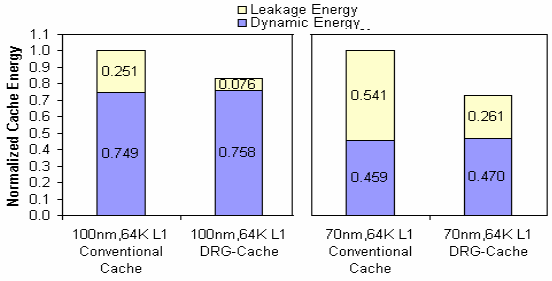

Figure 8. The energy savings for a 64K L1 DRG-Cache.

6.2.3 Energy Savings

Figure 8 shows the contribution of leakage and dynamic energy in L1 cache based on energy analysis given in section 5 and measured utilization factor. Assuming total energy of conventional cache is 1, we normalized leakage and dynamic energy consumption of DRG-Cache to the conventional cache. The graph indicates that leakage energy accounts for 25% in 100nm and 54% in 70nm technology for L1 caches. DRG-Cache saves around 70% leakage in 100nm and 51% in 70nm process. The energy overhead associated with decoder results in 1.2% and 2.3% increase in total dynamic energy of L1 cache for respective process technologies. The overall energy reduction achieved by DRG-Cache is as much as 27% in 70nm and 16.5% in 100nm process technology.

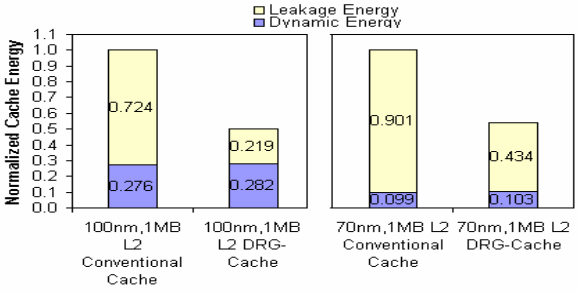

Figure 9. The Energy saving for a 1M L2 DRG-Cache.

Figure 9 shows the normalized (similar to Figure 8) energy results corresponding to 1 MB L2 Cache. Because L2 Cache utilization is small as compared to L1 cache, leakage energy accounts for as much as 90% in 70nm and 72.4% in 100nm process. The increment in dynamic energy is 2.1% and 4% in 100nm and 70nm process respectively. This increase in dynamic energy loss as compared to L1 cache is due to large decoder associated with L2 cache. The overall energy saving achieved by DRG-Cache in L2 is 50% in 100nm and 46.4 % in 70nm process technology.

In both L1 and L2 caches, the leakage power saving for 100nm technology is larger than 70nm technology, though it is shown in [16] that the leakage savings due to stacking is expected to increase with scaled technologies. However, for the Berkeley Predictive Technology Model [11] the subthreshold swing S is smaller in 100nm technology compared to 70nm, and hence we observe more leakage savings in 100nm. For L1 cache, the leakage energy contribution to the overall energy for 70nm is larger than the leakage contribution for 100nm process. For L2 it is approximately same for both the process technologies, which makes the overall energy savings to be larger for 70nm than for 100nm process in L1 cache. Hence, for L2 cache, the overall energy savings using DRG with 100nm process is larger than the energy savings using 70nm process, unlike L1 cache.

DRG-Cache achieves significant reductions in the power-delay product for both in L1 and L2 caches. This reduction ranges from as much as 43%-47% in L2 and 12%-23% in L1 depending on the process technology used. This demonstrates the effectiveness of DRG-Cache in reducing energy without significantly affecting the performance. However, one should note that for scaled technology one may not be able to use same low Vt for both logic and cache because of excessive leakage in large memory arrays. High Vt SRAMs in turn will increase the cache access time and process cost. Hence, our technology is a usable alternative for deep submicron design.

7 CONCLUSIONS

In this paper we explored an integrated circuit and architectural level approach to reduce leakage energy dissipation in deep-submicron cache memories while maintaining high performance. DRG-Cache utilizes inherent idleness of cache to save leakage by turning off the idle sections of the SRAM core. The key observation in this paper is that data is not lost when the gated-Ground transistor is turned off in the unused sections of the cache. A fabricated test chip in 250nm technology verified the feasibility of our approach.

Our simulations for 100nm and 70nm technologies using single V_t approach show that DRG-Cache saves around 16.5% and 27% energy in L1 and 50% and 47% energy in L2 cache, with approximately 5.2% and 5.3% impact on performance for 100nm and 70nm process, respectively, compared to conventional caches with same low Vt. The power-delay products show 12% and 23% improvement for L1 and 47% and 43% improvement for L2 compared to conventional design, for respective process technologies. This shows the effectiveness of DRG-Cache in saving leakage for contemporary as well as future cache designs.

8 REFERENCES

[1] J. Montanaro et. al. A 160-MHz, 32-b, 0.5-W CMOS RISC microprocessor. IEEE Journal of Solid-State Circuits, 31(11), 1703–1714, 1996.

[2] V. De. Private communication.

[3] S. Borkar. Design challenges of technology scaling. IEEE Micro, 19(4), 23–29, July 1999.

[4] S. Manne, A. Klauser, and D. Grunwald. Pipline gating: Speculation control for energy reduction. In Proceeding of the 25th Annual Int. Symp. on Comp. Archi., 32–141, 1998

[5] M. D. Powell, S.-H. Yang, B. Falsafi, K. Roy, and T. N. Vijaykumar. Gated-Vdd: A circuit technique to reduce leakage in cache memories. In Proceedings of ISLPED, (July 2000), ACM Press, 90–95.

[6] J. M. Rabaey. Digital Integrated Circuit. Prentice Hall, 1996.

[7] I. Fukushi, R. Sasagawa, M. Hamaminato, T. Izawa, and S. Kawashima. A low-power SRAM using improved charge transfer sense. In Proceedings of the 1998 Int. Symp. on VLSI Circuits, pages 142–145, 1998.

[8] L. Wei, Z. Chen, M. Johnson, K. Roy, and V. De. Design and optimization of low voltage high performance dual threshold CMOS circuits. In Proceedings of the 35th Design Auto. Conf., pages 489–494, 1998.

[9] F. Hamzaoglu, Y. Ye, A. Keshavarzi, K. Zhang, S. Narendra, S. Borkar, M. Stan, and V. De. Dual-Vt SRAM cells with full-swing single-ended bit line sensing for highperformance on-chip cache in 0.13um technology generation. In Proceedings of the 2000 Int. Symp. on Low Power Elect. and Design (ISLPED), July 2000.

[10] Z. Chen, L. Wei, M. Johnson, K. Roy. Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks. Int. Symp. on Low Power Electronics and Design, 1998, pp.239-244.

[11] http://www-device.eecs.berkeley.edu/~ptm/

[12] N. Shibata, M. Watanabe and Y. Sato. A 2-V 300-MHz 1-Nb Current-Sensed Double-Density SRAM for Low-Power 0.3-µm CMOS/SIMOX ASICs. IEEE Journal of Solid State Circuits, Vol. 36, No. 10, pages 1524-1537, Oct. 2001.

[13] D. Burger and T. M. Austin. The SimpleScalar tool set, version 2.0. Technical Report 1342, Computer Sciences Department, University of Wisconsin–Madison, June 1997

[14] T. Wada and S. Rajan. An Analytical Access Time Model for On-Chip cache Memories. IEEE Journal of Solid State Circuits, Vol. 27, No 8, pages 1147-1156,August 1992.

[15] J. L Hennessy and D. A. Patterson. Computer Architecture A Quantitative Approach. Morgan Kaufmann, 2nd edition

[16] Narendra S., et al. Scaling of Stack Effect and its Application for Leakage Reduction. in Proceeding of ISLPED'01 (Hiltington CA, Aug 2001), ACM Press, 194-200.