

WARM SRAM: A Novel Scheme to Reduce Static Leakage Energy in SRAM Arrays

Mahadevan Gomathisankaran
Electrical and Computer Engineering
Iowa State University
Ames, IA 50011
Email: gmdev@iastate.edu

Akhilesh Tyagi
Electrical and Computer Engineering
Iowa State University
Ames, IA 50011
Email: tyagi@iastate.edu

Abstract—The increasing sub-threshold leakage current levels with newer technology nodes has been identified by ITRS [1] as one of the major fundamental problems faced by the semiconductor industry. Concurrently, the expected performance improvement and functionality integration expectations drive the continued reduction in feature size. This results in ever-increasing power per unit area and the accompanying problem of heat removal and cooling [2]. Portable battery-powered applications, fuelled by pervasive and embedded computing, have seen tremendous growth and have reached a point where battery energy and power density can't be increased further [3]. This raises the computational throughput per watt target for the future technology nodes. SRAM arrays which are used widely as a system component, such as *caches* and register files, in both high-performance and portable systems, are getting to be dominant power consumers because of their large capacity and area. Hence any reduction in *cache* energy can result in considerable overall power reduction. In this paper, we propose a novel circuit technique using *depletion mode devices*, to reduce the static energy of SRAM array in an on-chip cache by 90% without any performance impact.

I. INTRODUCTION

Static subthreshold leakage has emerged as one of the major impediments in CMOS scaling. The magnitude of the problem is reflected in the fact that the leakage current per unit transistor width is expected to increase a 1000-10000 fold in going from 180nm technology node to 70nm technology node. This appears to pose a problem only for state holding circuits such as memory arrays that are idle for extended periods until one considers the following. In the same period, 180nm to 70nm technology node, the on (drive) current is expected to stay constant at 750 $\mu\text{A}/\mu\text{m}$ in order to sustain the historical speed advantage. In the 180nm technology node, the ratio of on current to leakage current per unit transistor width is approximately 7.5×10^6 . This ratio is projected to shrink to 750 in 70nm technology node and to 75 in 32nm technology node! Some of the implications of this projection are as follows. In 32nm node, the leakage current is expected to play a major role in the design of some of the high fanout, computing logic as well. Consider a bus with 30 clients being driven by one client. The twenty-nine *off* transistors (drivers) from the inactive bus clients are able to leak about one half (38%) of the *on* drive current of the driving client! Similar concerns will surface for logic blocks such as decoders, multiplexors, and

gate arrays. This argues that we need to develop low-leakage design styles even for computing logic with high activity rate.

This paper proposes a CMOS design style with low leakage characteristics. This design style belongs to a family of CMOS design styles that we name *warmup CMOS*. These logic styles warmup initially to charge some nodes up to a certain potential. The steady state energy savings are delivered in the “warmed-up” state with an initial energy cost for the warmup. The specific warmup CMOS design style presented in this paper relies upon depletion mode transistors to provide proper biasing to reduce the leakage. Hence we name it *dep-warmup-CMOS*. During the active phase of logic, the depletion mode transistors are transparent (Figure 1). However, during the inactive phase, the depletion mode transistors leak enough charge to bias the transistors to a point favorable with respect to the leakage current. A low leakage current equilibrium state is forced in the inactive phase. We describe the *dep-warmup CMOS* design style in Section II.

An immediate application of *dep-warmup CMOS* logic design style is in static memory arrays characterized by caches and register files. Microprocessors attain significant performance improvement by increasing the size and associativity of on-chip caches. For example, Intel's latest processor family, Centrino [19] has a 1MB L2 cache on-chip. Both *dynamic* switching energy, and *static* subthreshold leakage current induced energy of on-chip caches are already significant factors in over-all power consumption of the processors. The static leakage energy would overwhelm the dynamic energy for these caches with the expected 1000-10000 fold increase in leakage current in the 70nm technology node.

This is why the static leakage power reduction is one of the most important considerations in both high-Performance and low standby power circuits' design. As the device size shrinks, to maintain the same *on* current, threshold voltage V_t has to be reduced. This causes sub-threshold current to increase exponentially. Table-I compares 0.18 μm technology and 70nm technology with respect to *on* and *off* currents. This illustrates the magnitude of the problem.

As we had argued earlier, with feature size reduction, the static energy component of the on-chip caches constitutes a sizeable fraction of the total processor energy. Circuit

TABLE I
LEAKAGE CURRENT TREND

Technology	W_{min} (nm)	L_{eff} (nm)	V_t (V)	I_{off} (nA/ μ m)
TSMC [7] 0.18 μ m	200	100	0.35	0.1
BPTM [4] 70nm	70	38	0.20	99.47

techniques such as DVS [14], ABB-MTCMOS [10], Gated- V_{dd} [11] have been proposed to reduce the leakage energy in the caches. Control algorithms deploying these techniques estimate the active footprint in the cache to selectively power it. This results in performance penalty as well as in energy penalty in switching the cache lines back and forth from active to dormant state. Thus an efficient control algorithm is essential to extract maximum possible energy savings.

Power consumption in any digital integrated circuit, is given by the equation,

$$P_{total} = I_o V_{dd} + \alpha C V_{dd}^2 f \quad (1)$$

where, I_o is the leakage current, which is governed by the diode equation $I_s(e^{qV/kT} - 1)$, V_{dd} is the power supply voltage, α is the average switching activity factor, C is the total capacitance of the circuit, and f is the frequency of operation. The first term in the equation corresponds to the leakage power and the second term corresponds to the dynamic switching power. With the reduction in feature sizes, V_{dd} has also decreased, forcing a reduction in the threshold voltage V_t of the transistors. Thus the leakage current I_o which depends on V_t , through the diode equation (presented in the preceding discussion), increases [15].

A more elaborate expression for the sub-threshold leakage current is given by [[16], page 201],

$$I_{sub} = A * \exp\left\langle \frac{q}{n'kT} (V_g - V_s - V_{th0} - \gamma'V_s + \eta V_{ds}) \right\rangle * B \quad (2)$$

where,

$$A = \mu_0 C_{ox} \frac{W_{eff}}{L_{eff}} \left\langle \frac{kT}{q} \right\rangle^2 e^{1.8}$$

$$B = 1 - \exp\left(\frac{-qV_{ds}}{kT}\right)$$

The leakage power decreases exponentially with respect to V_{ds} due to the Drain Induced Barrier Leakage effect [16]. This fact has been used by earlier researchers in DVS [14] and in Row-by-Row Dynamic V_{dd} Control (RRDV) Scheme [18]. We propose a novel technique using *depletion mode devices* to achieve a leakage reduction of 90% in SRAM arrays without any additional control mechanism. We call our SRAM as *warm SRAM* for the reasons explained in Section III.

II. PROPOSED CIRCUIT TECHNIQUE

In order to achieve leakage reduction we should be able to dynamically control the voltages V_{gs} , V_{ds} and V_s . We use depletion mode devices to achieve this. A depletion device works in same way as an enhancement mode device, except that the device is *ON* even when V_{gs} is zero. To understand how depletion mode devices can help in leakage current

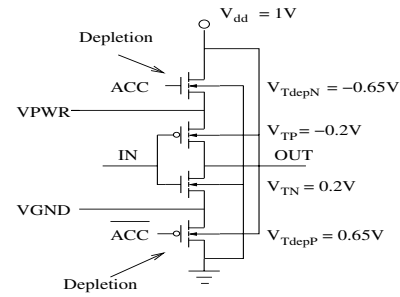


Fig. 1. Circuit with Depletion Device

reduction, consider Figure-1. Let, V_{TdepN} , V_{TdepP} , V_{TN} , and V_{TP} be the threshold voltages of the devices, such that

$$V_{TdepN} < 0, V_{TN}; V_{TdepP} > 0, V_{TP}.$$

When the access signal *ACC* is *HIGH*, both the depletion devices are *ON* hence the virtual power and ground nodes $V_{PWR} = V_{dd}$ and $V_{GND} = GND$. The circuit is in normal operation mode. When *ACC* is made *LOW* both depletion mode devices are switched-off and their V_{gs} is heavily reverse biased, i.e., for depletion NMOS $V_{gs} = -V_{dd}$ and for Depletion PMOS $V_{gs} = V_{dd}$. With $V_{ds} = 0$ both of these depletion mode devices will be in deep sub-threshold region, hence will have very minimal leakage. One of the two enhancement mode devices will be leaking depending on the *OUT* state. Hence the charge stored in V_{PWR} node starts getting accumulated in the node V_{GND} . As V_{PWR} decreases, (and simultaneously V_{GND} increases), the depletion mode NMOS (and PMOS) current increases. Note that the leakage of charge from V_{PWR} to V_{GND} reduces the V_{ds} and increases the V_{sb} for the enhancement mode devices. The leakage current in the enhancement mode device reduces exponentially as the V_{ds} reduces, V_{gs} is reverse biased and the V_{sb} increases. Thus all the three factors in Equation 2 come in to play and reduce the leakage current dramatically. The circuit reaches a stable point when the current supplied by depletion mode device is equal to the leakage current of the enhancement devices. If the threshold voltages are carefully chosen this equilibrium can be achieved at deep sub-threshold region of the depletion mode device, hence reducing the leakage current in the circuit dramatically.

Intuitively, we can see that this equilibrium is reached when V_{PWR} is closer to $-V_{TdepN}$ and V_{GND} is closer to $V_{dd} - V_{TdepP}$. Thus V_{ds} (consider *OUT* as *HIGH* and hence $V_{out} = V_{PWR}$) across the NMOS enhancement transistor will be $V_{TdepP} - V_{TdepN} - V_{dd}$. In order to differentiate V_{out}^{high} from V_{out}^{low} , this $V_{ds} = V_{TdepP} - V_{TdepN} - V_{dd} > 0$, resulting in $V_{TdepN} < V_{TdepP} - V_{dd}$. The equilibrium condition of the circuit can be found by solving Equation 2, substituting appropriate threshold voltages and equating all the currents. To have at-least a voltage difference of 0.3V between *HIGH* and *LOW* and for $V_{dd} = 1V$, this equilibrium condition requires (for the case $|V_{TdepN}| = |V_{TdepP}|$), $|V_{TdepN}| = 0.65V$. We used 70nm technology model files provided by BPTM [4] to perform the HSPICE [5] simulation. We chose the following threshold voltages for the HSPICE simulation: $|V_{TN}| = |V_{TP}|$

$= 0.2V$; $|V_{TdepN}| = |V_{TdepP}| = 0.65V$; $V_{dd} = 1V$. The steady state currents and voltages of different nodes are as given in Table-II. In these measurements, the depletion mode transistors were opened (ACC made 1) for 1 ns and then closed. All the three signals, i.e., IN , ACC and \overline{ACC} were driven by minimum sized inverters. Hence, they all have the same delay, rise and fall times.

TABLE II
STEADY STATE RESPONSE OF CIRCUIT IN FIGURE-1

$IN(V)$	$OUT(V)$	$V_{PWR}(V)$	$V_{GND}(V)$	$I_{off}(pA)$
0.0	0.949	0.949	0.148	10
1.0	0.052	0.852	0.052	01

Table I shows the leakage current of a minimum sized transistor to be $(99.47nA/\mu m) * 38nm$ which is $3779.86pA$. Compared to $10pA$ leakage current from our scheme in Table II, we have achieved a leakage current reduction of 377 times, but with a penalty in performance. The performance impact of the extra NMOS access transistor in the charging path and PMOS access transistor in the discharging path could be high. Since these devices are depletion mode devices, this impact can be managed to a great extent. Various delay parameters are listed in Table-II. There is a 54.5% increase in the average propagation delay which may not suit high-speed logic circuits. We can reduce the delay by several means. The easiest one is to make the ACC signal rise time slope higher (faster) compared to the input signal slope, or equivalently pre-raise ACC earlier than the input transitions. In this inverter, just by making the ACC rise 10 times faster than the input, we reduced the propagation delay penalty to 18%. But the increase in fall time, which is limited by the PMOS, is still 76%. We can increase the width of the depletion mode transistor, which unfortunately will also increase the leakage current. This is one of basic limitations of this circuit design style. The other limitation is the energy spent in switching V_{PWR} node. Note that whenever the input is 1 and $ACC = 1$, the virtual power node is pulled up to V_{dd} . However, during the inactivity period ($ACC = 0$), the virtual power node leaks some charge to the virtual ground node settling at a voltage equaling approximately $-V_{TdepN}$, which is about .65V in our design. A complete energy estimate needs to take this switching of the V_{PWR} node between V_{dd} and $-V_{TdepN}$ into consideration. The V_{PWR} node has capacitance $C_{eq} = 2 * C_{diff}$. Consider a 0.3V swing between V_{dd} and $-V_{TdepN}$. The extra energy required then will be:

$$Extra\ Switching\ Energy = \xi = 0.3 * C_{diff} J$$

If the circuit is idle for Δ_t time, on average, after each active period, the switching will pay-off only if,

$$\Delta_t \geq C_{diff} * 7.9 * 10^7$$

This is because from Table I, the 70nm leakage current is $99.47nA/\mu m$. With $L_{eff} = 38nm$, this gives the leakage current as $(.038\mu m) * (99.47nA/\mu m) = 3.78nA$. The leakage energy over time Δ_t will equal $3.78 * 10^{-9} * \Delta_t$. For the leakage energy to exceed the extra switching energy of V_{PWR} ,

$3.78 * 10^{-9} * \Delta_t > 0.3 * C_{diff}$, which leads to $\Delta_t \geq C_{diff} * 7.9 * 10^7$. With 70 nm technology and minimum width transistors, C_{diff} typically equals 0.1 fF hence $\Delta_t \geq 7.9ns$. For typical high performance circuits, with clock cycle time as 0.2 ns, we get around 40 cycles as the idle time. Hence the dep-warmup CMOS circuit technique can be applied if the average idle time between active periods is far greater than this *break-even* window. The immediate application will be in SRAM arrays, which occupy more than 50% of the typical microprocessor's area and consumes more than 50% of the energy.

TABLE III

PERFORMANCE IMPACT OF CIRCUIT IN FIGURE 1

	$t_{pLH}(ps)$	$t_{pHL}(ps)$	$t_r(ps)$	$t_f(ps)$
Base	16.8	10.54	33.63	17.31
New	25.9	16.32	40.72	30.89
%Inc	54.2	54.80	21.10	78.50

The circuit in Figure-1 is not truly *regenerative* in idle mode, i.e., as the devices are cascaded, the high output V_H^i of i^{th} stage is less than V_H^{i-1} (of $(i-1)^{st}$ stage). Similarly the low output V_L increases. But the V_H doesn't go lower than $|V_{TdepN}|$ and V_L doesn't increase beyond $V_{dd} - |V_{TdepP}|$. We can study a cross-coupled inverter configuration to determine the limiting leakage energy. This will be the minimum leakage reduction possible through this method as V_{PWR} will be equal to V_H and V_{GND} will equal V_L . In this scenario, one of the factors in leakage reduction, reverse biased V_{gs} , doesn't exist. Leakage current for this cross-coupled inverter estimated using HSPICE [5] is 515 pA. And $V_{PWR} = HIGH = 742mV$, $V_{GND} = LOW = 225mV$. Even though the leakage current increases by 25 times (per inverter) when compared to circuit in Figure-1, we still have achieved a reduction of 12-15 times when compared to the original inverter. Moreover, the high and low levels are only 500mV apart, and hence require less switching energy. Further reduction in leakage current is possible if multiple cross-coupled inverters share the same depletion device pair.

III. REDUCING STATIC ENERGY IN ON-CHIP CACHES

As we saw in Section II, the immediate application of the *dep-warmup-CMOS* is in SRAM arrays. Hence, in this section, we study its application in caches. The latest processors have two or three levels of caches, namely L0, L1, and L2. L0 is closest to the CPU and L2 is closest to the main memory. To improve the performance, L2 caches are now made on-chip and sized upto 1MB. In contrast, L1 cache sizes are typically 32KB to 64KB. L1 miss rates are on the average less than 2%, hence L2 will be very infrequently accessed when compared to L1. Given this, we validate our circuit design technique with L1 cache, which is in the critical path. If the scheme works with L1, we can infer that it will work with L2 as well (L2 has longer idle periods and is less critical for processor performance).

We use CACTI 3.0 [13] as the base model to evaluate the performance impact. Cacti uses sub-arrays to reduce the bit-line and word-line delays. For a 32KB 4-way cache with 32B

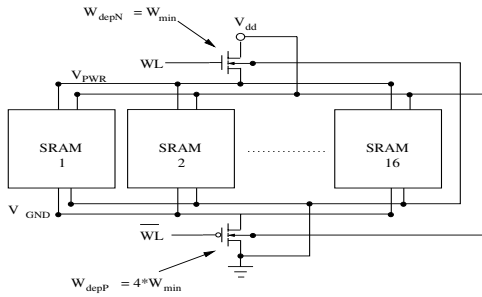


Fig. 2. Warm SRAM Wordline

block size, Cacti finds the optimal configuration as 8 sub-arrays with each having 128 rows of 256 bits. The timing parameters for various components in a 32KB, 32B, 4-way, 1 RW port, 70 nm cache are given in Table-IV. The cache access time for this configuration is 556.1 ps, which includes the data output driver delay of 76.2 ps. From Table-IV it is clear that data array delay does not form the critical path in cache access timing. A 32 bit processor with this cache configuration requires a tag array of 19KBits, which is just 7.4% of the data array size. Therefore, we apply our method of static leakage reduction only to data array, which is not in the critical path (increase in data array delay will not impact the cache access time), and consumes more than 92% of the static power in the caches.

TABLE IV

CACHE ACCESS TIMING FOR A 32KB, 4-WAY, 32B, 1 RW PORT, 1 SUB-BANK CACHE AS GIVEN BY CACTI

	Data Array Delay (ps)	Tag Array Delay (ps)
Decoder	208.572	099.410
Wordline	115.975	044.415
Bitline	011.765	011.898
Senseamp	072.625	044.625
Compare	-	112.912
Mux Driver	-	150.077
Sel Inverter	-	016.612
Total	408.936	479.949

We can use a depletion device pair per SRAM cell to reduce the leakage power as described in Section II. However, this will increase the SRAM cell size, and hence the area increase will out-weigh the leakage power reduction. An alternative is to share the depletion mode transistors with multiple SRAM cells. The wordline access signal can be used to control the depletion devices since it already encodes the active periods of a SRAM cell. Hence, no additional control signals need be generated. There are two ways to share the depletion or the *voltage clamp* devices. As we saw in Table-II the fall time t_f is approximately 4 times the rise time t_r . We decided to increase the depletion PMOS width by a factor of 4. This leads to approximately equal t_r and t_f . For an SRAM cell which is a cross-coupled inverter, write time is not the bottleneck as the cross-coupling effect aids in state transition. The read time is impacted only by the discharge path. Hence increasing depletion PMOS width alone is justified. From Table-IV data, the data array delay is 85.2% of the tag array delay and data bitline delay is 2.73% of the data array delay. Hence upto 6 times increase in bitline delay will still not increase the overall cache access time. We try to exploit this fact

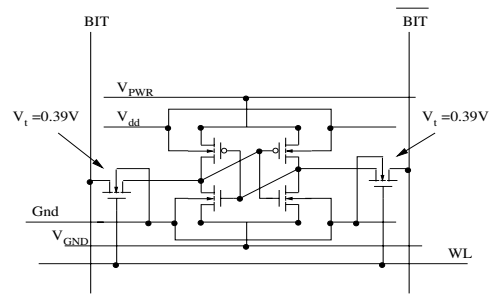


Fig. 3. Basic SRAM Cell used in Simulation

in deciding the number of *voltage clamp* devices and their sizes. From Table-II, we can share one depletion PMOS with 4 SRAM cells resulting in 4 to 5 times increase in the fall time t_f . Based on our earlier calculation of t_r and t_f , we can share one depletion NMOS with 16 SRAM cells. In our configuration, each wordline has 256 SRAM cells. For every group of 16 SRAM cells, one *voltage clamping* depletion device pair is used as shown in Figure-2. This dep-warmup CMOS configuration is compared with the base wordline configuration of 256 SRAM cells. The basic SRAM cell we use in HSPICE simulation is shown in Figure-3. We use low V_t devices in the cross-coupled inverters and high V_t devices for the access transistors, in order to reduce the leakage due to the access transistors. Except for the access transistors, all the other transistors have $|V_t| = 0.2V$.

A. Leakage Reduction

The most relevant parameters of the proposed dep-warmup CMOS cache is the leakage reduction achieved by this circuit and the steady state voltages of the nodes. The *current* measurements presented in Table-V are for a single SRAM cell. The steady state values were measured after a *bit* was written into the cell. When compared to the values presented in Section II for the cross-coupled inverter case, we see that V_H has decreased, i.e., moved closer to $|V_{TdepN}|$. This is because we have shared one depletion mode NMOS with 16 SRAM cells, hence the leakage current in the circuit has to decrease. All the 16 SRAM cells have to share the leakage current supplied by one depletion mode NMOS. V_L has also moved closer to $V_{dd} - |V_{TdepP}|$ but not as much as V_H . This is because we have increased the width of depletion mode PMOS by 3 times over the minimum. Hence it could supply more leakage current.

We can reduce the leakage current further by increasing the sharing but that will affect the transition delay and hence impact the performance. From Table-V it is clear that we have achieved more than 23 times static power reduction. However, we need to assess the performance impact as well. The factor which will contribute to performance reduction is the delay in charging up V_{PWR} node from 0.686V to 1.0V. In the following two sub-sections we will analyze the performance impact on *write* and *read* operations.

B. Performance Impact on Write Operation

We performed HSPICE simulations on a subarray of size 128 rows by 256 columns (as given by Cacti) to study the

TABLE V
STEADY STATE VALUES OF A WARM SRAM CELL

Param	Base	Warm SRAM
I_L (pA)	6250	262
$V(BIT)$ (V)	1.0	0.686
$V(\overline{BIT})$ (V)	0.0	0.252

impact on *read* and *write* times. Each bit and bitbar line was connected with a transmission gate to the bit value to be written, and with a PMOS transistor to precharge it to V_{bitpre} for read operation. The wordline enable signal was generated and characterized with the delay estimated by Cacti. The values of the different parameters are given in Table-VI. There are two assumptions involved in WL and \overline{WL} signal delay parameters. First, the delay of WL is not affected by the addition of $16 \cdot W_{min}$ depletion mode NMOS transistors. This assumption is valid as WL is already driving $512 \cdot C_g$ of memory cell access transistors, and an addition of $16 \cdot C_g$ does not increase the delay by more than 3% (or the driver could be sized to absorb this delay). Second, \overline{WL} signal is generated from WL and since it is driving only $64 \cdot C_g$, its delay can be made one tenth of the WL delay.

TABLE VI
TRANSIENT ANALYSIS PARAMETERS

Param	Value	Param	Value
WL t_r and t_f	100 ps	Base t_r	47.0 ps
WL t_r and t_f	10 ps	Base t_f	22.0 ps
WL Pulse Width	200 ps	Warm SRAM t_r	50.1 ps
V_{bitpre}	0.5 V	Warm SRAM t_f	00.0 ps

Table-VI gives the transition delay values for both the base configuration and the *warm SRAM* configuration. These delays are measured from the point WL signal is completely ON. It is clear from these values that the write operation is not getting affected by the presence of depletion mode devices. Firstly we are getting the advantage of depletion PMOS getting switched on ten times faster than the access transistors. Therefore V_{GND} transits to zero even before access transistors are completely opened allowing bit to become zero with zero delay. Secondly since the bit transits from a non-zero initial value to V_H , the peak current requirement for the transition is smaller. Thus the single depletion NMOS transistor is able to supply the required current for 32 inverter transitions. This fact also illustrates that the proposed circuit uses less energy for bit transitions.

However, the *warm SRAM* has one disadvantage in terms of energy, i.e., every time a bit is written, whether the bit is transitioning or not, it has to raise the V_{PWR} node from V_H ($\approx 700mV$) to V_{dd} . This also applies to the output node of an inverter which is in V_H state. Since V_{PWR} node has capacitance $518 \cdot C_{diff}$ (per wordline) and the output node of an inverter has capacitance $768 \cdot C_{diff}$ (per wordline), this energy will approximately equal $327.9 \cdot C_{diff}$. For $70nm$ technology, C_{diff} is in the range of $0.11 fF$. Hence the extra energy spent will be approximately $36.07 fJ$. Note that this extra energy is paid only when the memory cell bit does not change state due to this write. This is because the normal write scheme will not consume any energy in such a case, however *warm SRAM* will still need $.14 fJ$ per bit (where state does

not change). Hence, the extra write energy is proportional to the number of bits that do not change state. We calculated the write energy by integrating the input current. This energy is parameterized by the number of bit-transitions (bits changing state). The results are shown in Table-VII. The first column of the table specifies the number of cells changing state. This table shows that *warm SRAM* consumes less energy than the conventional circuit even with 128 bits changing state (out of 256 in a single row/block). The break-even point is somewhere closer to the point where 70 bits change state. For a write with less than 70 bits changing state, the constant energy to raise V_{PWR} and V_{out} nodes for the bits not changing state dominates. This results in higher energy consumption for *warm SRAM* than for the conventional SRAM. Whenever a cache block is replaced, if we assume a uniform distribution of bit values, we will get on the average 128 bit transitions. For cache write access, this depends on the access width. The *warm SRAM* will be spending at most $36.07 fJ$ extra energy for any access causing less than 64 bits to transit. When compared to the dynamic energy per cache access, which is estimated by Cacti as $0.3nJ$, this extra energy is very insignificant. Hence we can safely assume that it has little effect on the overall dynamic energy.

TABLE VII
WRITE ENERGY COMPARISON

No of Bits	Energy (fJ)		Peak Current (mA)	
	Base	Warm SRAM	Base	Warm SRAM
256	320	144	5.53	0.997
192	240	132	4.14	0.930
128	160	118	2.75	0.840
64	80	99	1.36	0.735

C. Performance Impact on Read Operation

The most critical operation in cache is read operation which occurs twice as often as writes. It is critical because load latency cannot be hidden. The instructions waiting on read results often stall. Whereas, the store (write) latency can be easily hidden with write buffers. No instructions in the immediate vicinity depend on the outcome of the write operation. The tag array access is, however, in the critical path in the cache read, as we observed from Table-IV. Hence we can exploit this slack to make the data array path slower without impacting the same cache access time. This is the basis for the chosen depletion mode transistor widths. If we have to reduce the delay any further, we can reduce the degree of sharing, trading it with increase in leakage current. As per Cacti, bitline delay is defined as the delay between the time at which wordline enable is ON and voltage difference between bit and bitbar becomes $100mV$. Cacti uses a precharge voltage of 0.7 V for the $70nm$ technology and estimates the bitline delay to be $11.7 ps$. We varied the precharge voltage from 0.7V to 0.5V in $50mV$ steps to study its influence on bitline delay for both the base circuit and the *warm SRAM* circuit. The results are shown in Table-VIII. The difference in Cacti's estimation and our results can be attributed to the high- V_t access transistors. Since we are interested in estimating the leakage savings in *warm SRAM*, we used high- V_t ($\approx 0.39V$)

access transistors in all the other reported simulation results. We used a pulse width of 200 ps for the wordline signal in our simulations. *Warm SRAM* bitline delay for 0.7V and 0.65V precharge voltages are greater than this 200 ps pulse width for the wordline. Hence, the bitlines did not achieve the 100mV difference. Furthermore, since 0.5V closely matches with Cacti's estimation of the bitline delay, we use 0.5V as our precharge voltage for all the further simulations. From Table-VIII, it is clear that bitline delay increase for precharge voltage of 0.5 V and 0.55 V does not increase both *cache access time* as well as the *wave pipelined cycle time*. Thus *warm SRAM* does not have any performance impact on cache access time.

TABLE VIII

BITLINE DELAY AS A FUNCTION OF PRECHARGE VOLTAGE

V_{bitpre} (V)	Base (ps)	Warm SRAM (ps)	%Inc
0.70	39.8	>200	X
0.65	30.6	>200	X
0.60	24.1	181.4	652.7
0.55	21.2	129.3	509.9
0.50	20.5	113.6	454.1

As is the case with write, read operation also requires extra energy to charge the nodes to V_{dd} . This can be avoided if we don't switch on depletion NMOS for read operation. This requires additional control logic and a separate driver for depletion NMOS transistors. This may not pay off since the extra energy spent for charging V_{PWR} and output node is at most 36fJ (based on our prior estimation). This potential energy saving is insignificant when compared to per access dynamic energy of the cache. This energy estimate is worst case since the V_{PWR} node acts like a capacitor and is discharged by the leakage current in the circuit. The time the node takes to reach the steady state depends on the leakage current magnitude. It takes more than 200ns to leak down to the steady state value from V_{dd} at room temperature ($T=25^\circ C$) as shown in Figure-4. If the cache line was accessed again within this 200ns period, we will spend less energy than 36fJ on charging V_{PWR} and output node to V_{dd} . We calculated the energy used per word line access for different inter-access time intervals. The results are shown in Table-IX. These results are very close to our estimates. This also shows that if two accesses happen within 25 ns of each other, we don't spend any extra energy compared to the base SRAM cell design. The reason why the warm SRAM energy is less than the base circuit at 25 ns inter-access gap is as follows. During the read operation, the potential of the SRAM cell node which is pulling the bitline rises by a small amount over GND . The discharging NMOS transistor cannot do with V_{ds} at 0V. This causes a short circuit path from V_{dd} to ground. Moreover, this effect is higher in the case of base circuit as it is discharging more current.

D. Architecture/Program Level Energy Estimation

In the preceding sections we saw how *warm SRAM* can reduce the leakage energy without impacting the read/write access times of a cache. In this section, we evaluate the static leakage energy savings for a hypothetical 32KB, 4-way cache. The extra access energy for warm SRAM to charge

TABLE IX

READ ENERGY W.R.T TIME AFTER AN ACCESS

Base Read Energy: 25.92 fJ		
Time (ns)	Energy (fJ)	Extra Energy (fJ)
25	23.99	-1.93
50	33.86	7.94
75	41.56	15.64
100	47.22	21.30
125	51.38	25.46
150	55.27	29.35
175	57.45	31.53
200	59.44	33.52
300	59.44	33.52

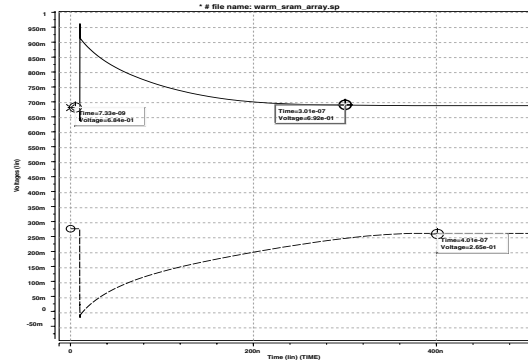


Fig. 4. Discharging of V_{PWR} Node and Charging of V_{GND} Node

V_{PWR} and output up to V_{dd} is accounted for in a cycle by cycle microarchitecture simulator, SimpleScalar version 3.0 [8]. The average number of cache accesses are captured by SimpleScalar. We executed SPEC2000 [6] integer benchmarks on this simulator. Suleyman Sair et. al [9], did analyze the memory behavior of SPEC2000 benchmarks and found the 500 million execution window which closely matches with the characteristics of the whole application. We use SimpleScalar 3.0 with the configuration shown in Table-X and the execution window specified in [9] to collect the statistics.

TABLE X

SIMPLESCALAR SIMULATION PARAMETERS

Parameter	Value
Processor	Alpha out of order 4 way issue
L1-DCache	32KB, 4 way, 32B line, 1 cycle access latency
L1-ICache	16KB, 1 way, 32B line, 1 cycle access latency
L2-Unified	256KB, 4 way, 64B line, 8 cycle access latency

In order to calculate the net energy savings, we need to account for the extra energy incurred by this method. We have so far seen one source of extra energy, i.e., energy spent in bringing *warm sram* to normal state. This energy penalty will be significant only if the accesses occur more than 50 ns apart. Table-XI lists the percentage of L1 cache accesses within 50 cycle window and 100 cycle window. From Table-XI it is evident that more than 80% of the accesses occur within 100 cycles of another access. Using this fact and the extra energy figures presented in Table-IX, we can estimate the extra energy per access. These numbers depend on the cycle time. We use both 0.2ns and 0.5ns as the cycle times to illustrate the energy savings. In both the cases, 100 cycles is less than 50ns. Hence we penalize only the accesses occurring after more than 100 cycles from another access with 33.52fJ. We chose 0.2ns because it is the wave pipelined cache cycle

time estimated by Cacti and $0.5ns$ because it approximates the cache access time.

TABLE XI
ACCESS PERCENTAGE W.R.T TIME

Benchmark	50 Cycles	100 Cycles
crafty	59.73	9.15
eon	77.91	6.06
gcc	77.85	5.47
twolf	70.40	6.46
gzip	79.73	5.61
bzip	86.92	4.90
mcf	68.47	11.02
perlbmk	77.32	3.37
parser	75.18	7.36
vpr	69.59	7.81
Avg	74.31	6.721

Another source of extra energy is the access control signals generation for the depletion NMOS and PMOS. Since depletion NMOS is driven by the wordline WL signal, we will not incur any extra energy for that. However, we have to generate \overline{WL} and this will be approximately $(1/8)^{th}$ of the wordline driver energy (as this signal has to drive only $64 \cdot C_g$, $(1/8)^{th}$ the capacitance driven by WL). From Cacti, the wordline energy is $158.95fJ$. Hence, we assess $19.87fJ$ extra energy per access to generate \overline{WL} signal. We take this into account as well while calculating the energy savings. Table-XII lists various benchmarks and the number of execution cycles for 500 Million instructions and the corresponding energy savings.

The energy saving is calculated with the leakage currents given in Table-V. For a 32KB cache, this gives us the base leakage power of 1.638 mW while the *warm SRAM* leakage power as $68 \mu W$. Using these figures and taking the energy penalty incurred by our method into consideration, we achieve a 90% reduction in leakage energy on average in a 32KB data array (for both 0.2 and 0.5 ns cycle times) without impacting the cache access time. Table-XII lists the energy savings for various SPEC2000 integer benchmarks for both 0.2ns cycle time and 0.5ns cycle time.

IV. MODEL VALIDITY

In all our HSPICE simulations, we have assumed that *enhancement mode* model file with its threshold voltage modified to suit the *depletion device* will be able to model the depletion device's characteristics. We justify the validity of this assumption in this section. We discuss only the depletion NMOS case, however the discussion is equally applicable to depletion PMOS as well.

Depletion NMOS is formed by implanting donor atoms in the substrate. The threshold voltage of such a device is given by Equation-3 [[17],page 238].

$$V_T = V_{T0} + \gamma_I (\sqrt{\phi_{bi} + V_{sb}} - \sqrt{\phi_{bi}}) \quad (3)$$

Where,

$$V_{T0} = V_{FB} + \phi_{bi} - \frac{qN_d d_I}{C'_{ox}} \left(1 + \frac{d_I C'_{ox}}{2\epsilon_s}\right) + \gamma_I (\sqrt{\phi_{bi}})$$

$$\gamma_I = \left(1 + \frac{d_I C'_{ox}}{\epsilon_s}\right) \gamma$$

V_{FB} → Flat band voltage

ϕ_{bi} → Built in potential ($\phi_{Fp} - \phi_{Fn}$)

γ → Body effect coefficient of the unimplanted substrate
 d_I → Implantation depth
 N_d → Donor concentration

Note that a depletion mode transistor has higher body effect compared to enhancement mode devices. Hence N_d and d_I should be varied to get the required device characteristics. The first point of consideration is that the depletion NMOS should get cut-off when $V_{sb} = 0.65V$ and $V_{gs} = 0V$. Hence by equating (3) to $-0.65V$, we can solve for N_d and d_I . The second point of consideration is that when $V_{gs} = 1V$ the gate should have gain comparable to what is predicted by the enhancement mode model. If these two operating points can be verified then our circuit will yield a similar result if we use proper depletion model for the devices.

There are four regions of operation in a depletion device,

- 1 Cut-Off - device is completely depleted at the source end.
- 2 Surface depletion - surface is depleted but the buried channel exists conducting the current.
- 3 Surface Accumulation - as V_{gs} increases beyond threshold, inversion occurs and carriers are accumulated on the surface.
- 4 Surface Accumulation/Depletion - as V_{db} or V_{sd} increase beyond a certain value, it depletes the channel on one of the sides making the device behave like a saturated enhancement mode device.

Our operating points should be in cut-off and surface accumulation regions. The device enters surface accumulation region if $V_{gs} > V_N$ where $V_N = V_{FB} + \phi_{bi}$. The gain of the device in surface depletion region is given by $\beta = \mu_b C'_{ox} / (1 + \sigma)$, where $\sigma = \frac{C'_{ox} d_I}{\epsilon_s} \left(\frac{C'_{ox} d_I}{2\epsilon_s} + 1 \right)$. Even though μ_b , bulk mobility, is typically larger than the surface mobility, the factor $(1 + \sigma)$ tends to reduce the gain in surface depletion region when compared to the enhancement mode device. Once surface accumulation occurs the variations of channel charge with V_{gs} occur at the surface, and, thus, the gain is determined by the surface mobility and oxide thickness only. In other words the gain is comparable to the enhancement mode device. One problem with depletion device could be that if the implantation depth is too high then controlling the buried channel just with gate voltage will not be possible as the surface inversion will occur before pinch-off condition could be reached. This can be overcome if the depth is not made very large or the buried channel could be depleted with high V_{sb} . In our circuit since both drain and source are always above 0.65 V this condition can never occur.

The process parameters for the 70 nm technology [4] we are considering are,

- N_{sub} → Substrate doping concentration = $6 \times 10^{16} \text{ cm}^{-3}$
 N_{ch} → Channel doping concentration = $1.2 \times 10^{18} \text{ cm}^{-3}$
 X_j → Junction depth = $299.99 \times 10^{-10} \text{ m}$
 t_{ox} → Oxide thickness = $16 \times 10^{-10} \text{ m}$
 γ → Body effect coefficient of bulk = 6.563×10^{-2}

We can solve Equation-3 for various values for γ_I i.e., d_I for the condition $V_T|_{V_{sb}=0.65} = -0.65V$. We can then choose

TABLE XII
NET ENERGY SAVINGS

Prog	Exec Cycles	Mem Access	Energy Penalty per access (μJ)	%Net Saving (0.2 ns/cyc)	%Net Saving (0.5 ns/cyc)
crafty	396782412	195828079	5.93	91.28	94.02
eon	350714953	240118536	6.06	90.57	93.74
gcc	393784461	223031723	5.68	91.45	94.09
twolf	444314516	172189507	4.76	92.58	94.54
gzip	277336702	169725136	4.21	91.22	94.00
bzip	269543836	185471790	4.19	91.10	93.95
mcf	487390086	195632037	5.23	92.57	94.54
perlbmk	346674071	216796572	5.71	90.82	93.84
parser	326925643	190878110	4.91	91.26	94.01
vpr	421717636	185474202	5.09	92.16	94.37
Avg	371518431.60	197514569.20	5.18	91.50	94.11

the solution closest to our assumptions. The various values of N_d and σ for various d_I values are listed in Table-XIII.

TABLE XIII

PROCESS PARAMETERS FOR DEPLETION NMOS

γ_I	d_I (10^{-10}m)	σ	N_d (10^{18}cm^{-3})	V_{T0} (V)	V_N (mV)
1.5 γ	24.21	0.625	28.2	-0.6786	-37.06
2.0 γ	48.41	1.5	14.23	-0.6881	-54.84
3.0 γ	100	5	5.667	-0.7084	-78.78

Note that there are various possible solutions but we should choose the one with the lowest σ . Depending on the process and the maximum N_d allowed, the device has to be scaled to take into account the decrease in the gain. In all these cases, note that $V_N < 0$ hence both V_{gs} and V_{gd} are $> V_N$. This places the device in surface accumulation region when switched on. Hence, our assumption of validating the *warm sram* circuit technique with enhancement device models is justified. With proper depletion mode models, we can verify other characteristics of the device such as subthreshold leakage characteristics. However, to within a first order approximation, our circuit can achieve the claimed static energy reduction.

V. CONCLUSION

The static leakage energy is one of the biggest challenges facing the semiconductor industry in the near to intermediate term future (3-10 years). The static energy consumption grows exponentially with reduction in feature size. On-chip caches occupy a major fraction of the processor's area. This holds both for high performance and for low power embedded processors. Since static leakage energy constitutes a large fraction of cache energy, and hence of total processor energy, on-chip caches form a good target for static energy reduction techniques. We proposed, modelled, and verified a new CMOS design style, warmup CMOS. These devices operate best when warmed up like an engine, where some of the key nodes have attained a certain potential. The warming-up cost is paid only once for each activity period, but the savings more than offset the initial warmup cost. The specific version of warmup CMOS presented in this paper is based on depletion mode devices – dep-warmup CMOS. We presented a SRAM cell design in dep-warmup CMOS and its block level implementation. The detailed SPICE simulations estimate the static leakage energy savings for L1 caches at more than 90% without any affect on the performance. We are further investigating other warmup CMOS design styles and their application to broader logic blocks.

ACKNOWLEDGMENT

The authors would like to acknowledge support from NSF Grant CCR 0209078 and AFRL through contract number F33615-02-C-1238.

REFERENCES

- [1] International Technology Roadmap for Semiconductors, 2001 Edition, Executive Summary. URL: <http://public.itrs.net>
- [2] J. M. C. Stork, "Technology Leverage for Ultra-Low Power Information Systems". *Proceeding of the IEEE*, Vol.83, Issue.4, April 1995.
- [3] T. Bell, "Incredible Shrinking Computers". *IEEE Spectrum*, pp. 37-43, May 1991.
- [4] Berkeley Predictive Technology Model. URL: <http://www-device.eecs.berkeley.edu>.
- [5] Star-HSPICE 2001.4 Avant! Corporation.
- [6] URL: <http://www.specbench.org/osg/cpu2000/>
- [7] Taiwan Semiconductor Manufacturing Company Ltd. URL: <http://www.tsmc.com>.
- [8] Doug Burger and Todd M. Austin. "The SimpleScalar Tool Set, Version 2.0", *Computer Sciences Department Technical report #1342*, University of Wisconsin-Madison, June 1997.
- [9] Suleyman Sair and Mark Charney. "Memory Behavior of the SPEC2000 Benchmark Suite", *IBM Research Report*, October 2000.
- [10] K. Nii, H. Makino, Y. Tujihashi, C. Morishima, Y. Hayakawa, H. Nunogami, T. Arakawa, and H. Hamano, "A low power SRAM using auto-backgate-controlled MT-CMOS", *In International Symposium on Low Power Electronics and Design*, pages 293-298, 1998.
- [11] M. Powell, S. H. Yang, B. Falsaifi, K. Roy, and T. Vijaykumar, "Gated- V_{dd} : A circuit technique to reduce leakage in deep-submicron cache memories", *In International Symposium on Low Power Electronics and Design*, pages 90-95, 2000.
- [12] John L. Hennessy and David A. Patterson, "Computer Architecture: A Quantitative Approach", *Morgan Kaufman, CA*, 2003.
- [13] Steven J. E. Wilton and Norman P. Jouppi, "An Enhanced Access and Cycle Time Model for On-Chip Caches", *WRL Research Technical Report 93/5*, July 1994.
- [14] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy Caches: Simple Techniques for Reducing Leakage Power", *In Proceedings of 29th Annual International Symposium on Computer Architecture*, pages 148-157 IEEE CS Press, 2002.
- [15] J. A. Butts and G. S. Sohi, "A Static power model for architects", *In Proceedings of 33rd Annual International Symposium on Microarchitecture*, 2000.
- [16] Kaushik Roy and Sharat C. Prasad, "Low-Power CMOS VLSI Circuit Design", *John Wiley & Sons, INC.*, 2000.
- [17] Yannis Tsividis, "Operation and Modeling of The MOS Transistor", *WCB/McGraw-Hill*, 1999.
- [18] Kaouichi Kanda, Takayuki Miyazaki, Min Kyeong Sik, Hiroshi Kawaguchi, and Takayasu Sakurai, "Two Orders of Magnitude Leakage power reduction of Low Voltage SRAM's by Row-By-Row Dynamic V_{dd} Control (RRDV) Scheme", *In Proceedings of IEEE/ACM ICCAD*, 2002.
- [19] URL: <http://www.intel.com>
- [20] M. C. Johnson, K. Roy, and D. Somasekar, "A Model for Leakage Control by Transistor Stacking", *Technical Report TR-ECE 97-12*, Purdue University, Department of ECE.