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

Mahadevan Gomathisankaran
Iowa State University
gmdev@iastate.edu

Akhilesh Tyagi
Iowa State University
tyagi@iastate.edu

① Introduction
② Proposed Circuit Technique
③ Reducing static energy in On-Chip Caches
④ Model Validity
⑤ Conclusion and Future Work

INTRODUCTION

Expected increase in the static leakage current

- ➜ Feature Size to reach 22*nm* in 2016
- ➜ Leakage current to increase by factor of 1K-10K in going from 180*nm* to 70*nm*

Leakage current will play a major role in circuit design

- ➜ Not only *arrays* but also high fan-out *logic* will be affected

New design methodologies have to be invented to avoid Red Brick Wall

- ➜ We propose *warmup-CMOS* which uses depletion mode transistors

Subthreshold Leakage in CMOS

Various leakage mechanisms

➜ PN Reverse Bias, Weak Inversion, DIBL, GIDL, Punchthrough

Leakage Current

$$I_{sub} = A * exp\langle \frac{q}{n'kT}(V_g - V_s - V_{th0} - \gamma' V_s + \eta V_{ds})\rangle * B \qquad (1)$$

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

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

Subthreshold Leakage in CMOS

Various leakage mechanisms

- ➜ PN Reverse Bias, Weak Inversion, DIBL, GIDL, Punchthrough

Leakage Current

$$I_{sub} = A * exp\langle \frac{q}{n'kT}(V_g - V_s - V_{th0} - \gamma' V_s + \eta V_{ds})\rangle * B \qquad (1)$$

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

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

EARLIER RESEARCH

Gated-V$_{dd}$

- + Interposes a high-V$_t$ transistor between the circuit and one of the power supply rails
- + Reduces the leakage current of a normal transistor to effectively the leakage current of the high-V$_t$ control transistor
- - Contents of the cell are lost
- - Control algorithm should be smart

ABB-MTCMOS

- + Dynamically raise V$_t$ by modulating the back-gate bias voltage, i.e., V_t = $V_{t0} + \gamma(\sqrt{\phi_{bi} + V_{sb}} - \sqrt{\phi_{bi}})$
- - Higher energy/delay per transition and higher V$_{dd+}$ offsets the leakage power savings

Earlier Research

Gated-V$_{dd}$

- + Interposes a high-V$_t$ transistor between the circuit and one of the power supply rails
- + Reduces the leakage current of a normal transistor to effectively the leakage current of the high-V$_t$ control transistor
- - Contents of the cell are lost
- - Control algorithm should be smart

ABB-MTCMOS

- + Dynamically raise V$_t$ by modulating the back-gate bias voltage, i.e., V_t = $V_{t0} + \gamma(\sqrt{\phi_{bi} + V_{sb}} - \sqrt{\phi_{bi}})$
- - Higher energy/delay per transition and higher V$_{dd+}$ offsets the leakage power savings

DVS

- + In sub-micron processes leakage current increases exponentially with supply voltage
- + Supply voltage is reduced to an optimum value (knee point of the curve, 1.5*V$_t$)
- + Two-fold reduction (both voltage and current) of the leakage power is achieved
- - Memory cell in standby (*drowsy*) mode cannot be read or written

What is Missing?

- ➜ A comprehensive solution which has low (much less) control overhead and still achieves the maximum possible leakage reduction
- ➜ Reduction is maximum if the circuit is in standby or low-leakage mode whenever it is not used

DVS

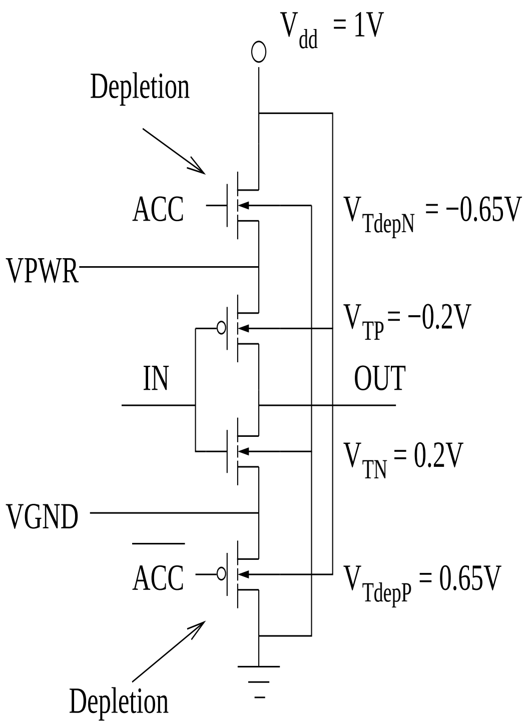
- + In sub-micron processes leakage current increases exponentially with supply voltage
- + Supply voltage is reduced to an optimum value (knee point of the curve, 1.5*V$_t$)
- + Two-fold reduction (both voltage and current) of the leakage power is achieved
- - Memory cell in standby (*drowsy*) mode cannot be read or written

What is Missing?

- ➜ A comprehensive solution which has low (much less) control overhead and still achieves the maximum possible leakage reduction
- ➜ Reduction is maximum if the circuit is in standby or low-leakage mode whenever it is not used

Our Proposed Solution

Warm Inverter



- ➜ Our solution uses Depletion mode devices
- ➜ The circuit is *warm*, i.e, when not accessed V$_{PWR}$ is less than V$_{dd}$ and V$_{GND}$ is greater than GND
- ➜ When compared to normal inverter in same technology, *warm inverter* achieves 377X leakage current reduction

Steady State Response

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

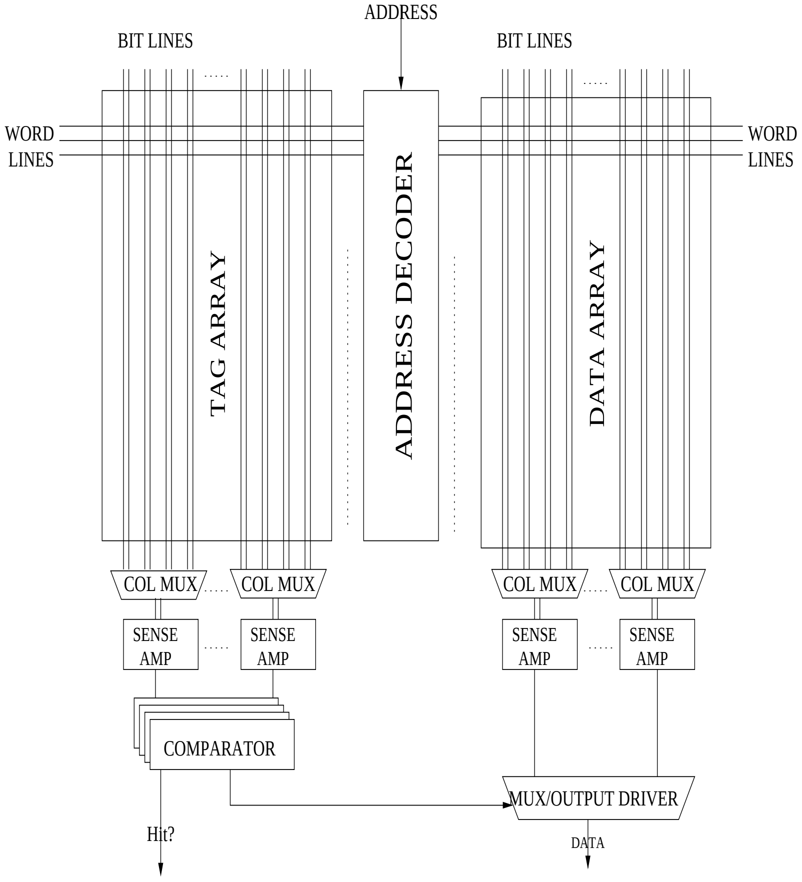
Limitations:

- ➜ Performance Penalty, as NMOS in the charging path and PMOS in the discharging path
- ➜ Energy Penalty, $Extra\ Switching\ Energy = \xi = 0.3 * C_{diff} J$
- ➜ Cascading Effect, for a cross coupled inverter we get High = 742*mV*, Low = 225*mV*, I_{off} = 515*pA* (compare with actual I_{off} 6.25*nA*)

Performance Impact

	$t_{pLH}\ (ps)$	$t_{pHL}\ (ps)$	$t_r\ (ps)$	$t_f\ (ps)$
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

Application to Caches



Cache architecture of a *n*-way Set-Associative Cache

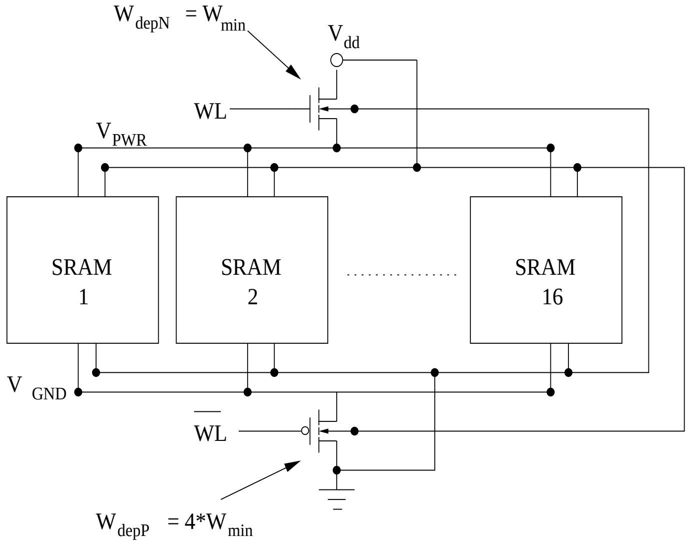
Cache Access Timing for a 32KB, 4-way, 1 RW Port, 1 Sub-bank Cache

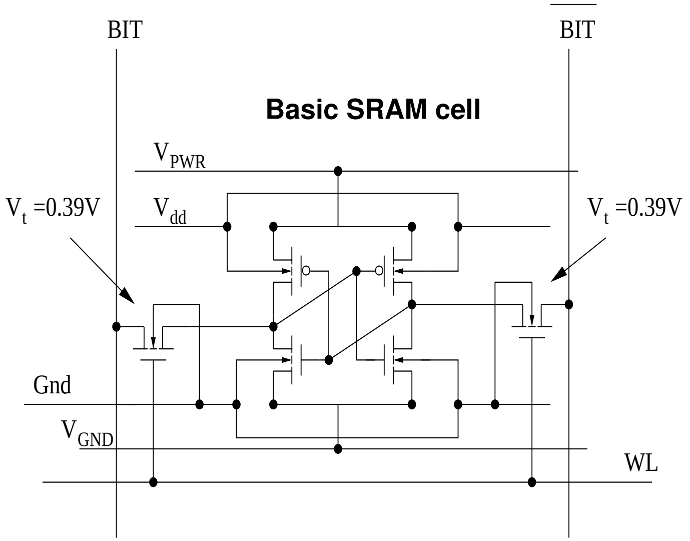
	Data Array Delay (ps)	Tag Array Delay (ps)
Decoder	208.572	099.410
Wordline	115.975	044.415
Bitline	011.765	011.898
Senseamp	072.625	044.625
Compare	-	112.912
Mux Driver	-	150.077
Sel Inverter	-	016.612
Total	408.936	479.949

- ➜ L1 cache sizes are typically 32KB - 64KB (Athlon has 128KB)
- ➜ L1 miss rates are on the average 2%
- ➜ On-Chip L2 caches are in the range of 256KB (Centrino has 1MB)
- ➜ We used CACTI 3.0 to find the cache access timing

Simulation Setup:

Warm SRAM configuration





- A depletion device pair per cell would increase the area hence offset the energy savings
- The wordline access signal is used to control the depletion devices
- PMOS$_{dep}$ is 4W$_{min}$, as cache read is in critical path this is justified
- Upto 6X increase in bitline delay (data array) will have no impact on cache access time
- Simulation is performed in HSPICE for a Subarray of size 128X256
- WL is not affected by addition of 16*C$_g$
- $\overline{WL}$ is generated from WL and since it is driving only 64*C$_g$ it delay can be made one tenth of WL

Leakage Reduction:

- ➜ Leakage power reduction - 23X
- ➜ V$_H$ has moved closer to $|V_{TdepN}|$, because one NMOS$_{dep}$ is shared with 16 SRAM cells
- ➜ V$_L$ has moved closer to $V_{dd} - |V_{TdepP}|$, but not as much as $|V_H|$, because width of PMOS$_{dep}$ has been increased

Steady State Response 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

Analysis of Write Operation:

- Transition delay values are as shown in the table
- Write operation is not getting affected by the presence of Depletion mode devices
- Two reasons,
  - Faster $\overline{WL}$ means V_{GND} transits to zero even before the access transistors are turned on
  - Since bits transit from non-zero initial value to V_H, the peak current requirement for the transition is smaller and could be supplied by the single NMOS$_{dep}$

Transient Analysis Parameters and Response

Param	Value	Param	Value
WL t_r and t_f	100 ps	Base t_r	47.0 ps
$\overline{WL}$ 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

Analysis of Write Operation (contd.):

- Irrespective of bit state changes, V$_{PWR}$ node and one of the output node (OUT$_{H}$) needs to be pulled up
- Considering the capacitance of V$_{PWR}$ node and OUT$_{H}$ node the extra energy would be 327.9*C$_{diff}$
- For 70*nm* device this would be 36*fJ* or 0.14*fJ*/bit which does not change state
- Warm SRAM uses more energy when 70 bits or less undergo state transition
- This extra energy (36*fJ*) is insignificant when compared to dynamic energy per access (0.3*nJ*), hence we ignored its impact

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

Analysis of Read Operation:

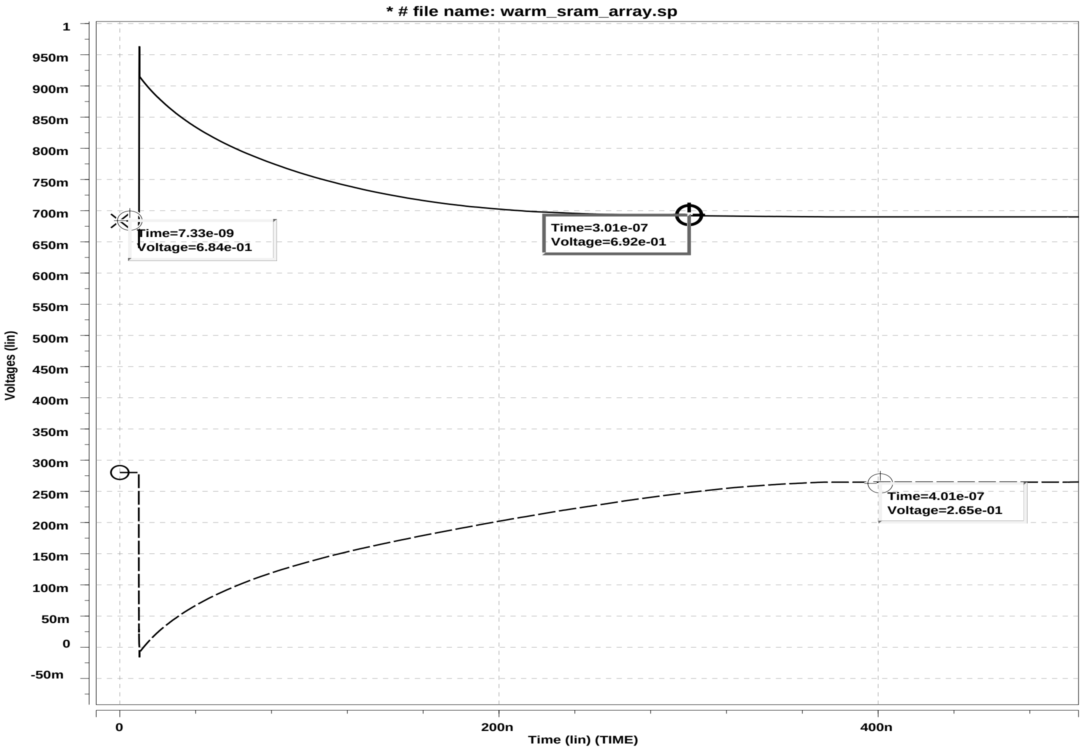
- ➜ Tag array access forms the critical path, hence Warm SRAM is used only in Data Array
- ➜ Since we use Hight-V$_t$ access transistors in SRAM cell, access time for precharge voltage of 0.5V closely matches with CACTI's estimated value
- ➜ Bitline delay increases by 4.5X for Warm SRAM, which doesn't increase both *cache access time* and *wave pipelined cycle time*
- ➜ The extra energy estimated in write operation also applies to read
- ➜ As V$_{PWR}$ node takes finite amount of time to discharge, extra energy depends on the inter-access time

Analysis of Read Operation (contd.):

Read Energy w.r.t Inter-Access time

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

Discharging of V$_{PWR}$ node



Architecture Level Estimation:

- SPEC2000 Integer benchmarks running on Simplescalar 3.0 is used to estimate the energy savings for a hypothetical 32KB,4-way L1 cache
- Two sources of extra energy
  - Energy to bring Warm SRAM to normal state (max 33.52*fJ* per access)
  - Generation of access control signals ($\approx$20*fJ* per access)
- Average net energy savings for 0.5ns cache access time (cycle time) is 94.11%

Access Percentage w.r.t Time

Benchmark	50 Cycles	100 Cycles	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 for 50 Cycles			74.31		
Avg for 100 Cycles			6.721		

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

Model Validity

- ➜ N_d (donor concentration) and d_I (implantation depth) could be varied to get the required device characteristics
- ➜ Two operating points need to be verified
  - $NMOS_{dep}$ should get cut-off when V_{sb} = 0.65V and V_g = 0V
  - When V_{gs} = 1V the gate should have gain comparable to what is predicted by the enhancement model
- ➜ The device should operate in Cut-Off or Surface Accumulation region
- ➜ We solved $V_T|_{V_{sb}=0.65}$ = -0.65V for various values of d_I and obtained viable values for N_d
- ➜ For all these values of N_d the requirement $V_{gs} > V_N$ is met

Process parameters for $NMOS_{dep}$

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

Conclusions and Future Work

- ➜ Static Leakage is one of the biggest challenges facing the semiconductor industry in the near future
- ➜ We have achieved more than 90% leakage energy reduction in On-Chip L1 caches without any performance loss
- ➜ Our technique is immediately applicable to any lower level caches (L2)
- ➜ On-Chip caches constitute a major fraction of processor's area, hence considerable leakage energy could be saved by using our methodology
- ➜ Currently investigating the usage of warmup CMOS design style in logic blocks
- ➜ Working on analytical model capturing the relationship between threshold of depletion devices and leakage reduction

THANK YOU!!

Questions?