A Pico-Joule Class, 1 GHz, 32 KByte x 64b DSP SRAM with Self Reverse Bias

Azeez J. Bhavnagarwala, Stephen V. Kosonocky, Michael Immediato, Dan Knebel & Anne-Marie Haen
IBM T J Watson Research Center, Yorktown Heights NY 10598, azeezb@watson.ibm.com

Abstract

New SRAM circuit techniques implemented in a standard 0.13 μm bulk Si CMOS process are reported in this work that (i) enable pico-joule energy dissipation per accessed bit at 1 GHz, (ii) lower total leakage power by over 80% from *all* unaccessed cells, during both active *and* standby modes, using a rigorous, self reverse biasing scheme that addresses leakage due to quantum tunneling *and* thermal excitation in *all* cell transistors, with an area, performance and noise margin penalty of less than 3% each and (iii) enable a programmable leakage reduction option that lowers leakage by over 90% when stored data is no longer desired.

Introduction

Single-cycle SRAM caches for fast, low cost wireless DSP applications such as 3G cell phones impose stringent design constraints on both active *and* standby power [1-2]. With increases in the spreads of Vt fluctuations seen in cell [3] and sense amplifier [4] transistors, delay variability of the bit path increases dramatically at low operating voltages, requiring larger margins for evaluate periods of the SRAM cell and sense amps, and as a result increases bitpath power, which dominates SRAM active power. Techniques to match the clock path with the bitpath using dummy BLs to minimize skew have been demonstrated [5] but are necessary only in large subarrays with long bit lines where the skew between pulses along the clock and bitpath become significant [6]. We propose an alternative scheme that employs very small subarrays with short WLs (16b) and short BLs (32b) where we tradeoff a rapidly improving bitpath delay *and* delay variability for a slowly deteriorating decode delay [7] while using proportionately smaller cell-pitch matched peripheral circuitry to reduce the area penalty that accompanies smaller subarrays. This scheme employs an Address Gated Pulse operation to synchronize and restrict switching activity spatially and temporally to only regions of the array that are being accessed for periods just long enough [8] to build sufficient signal for reads or writes. In tandem with mux-less operation along the entire bitpath, this scheme enables SRAM cell delays to track CMOS gate delays more closely at low voltages. Past attempts [9] to reduce standby power by gating the SRAM cell ground terminal with an NFET device to produce a stacking effect severely compromise the ability of the SRAM cell to retain data. With the cell storage node at '0', now 'floating', at high impedance and no longer firmly strapped to ground through the conducting pull-down cell NFET, fluctuations in cell transistor Vt make such a scheme very risky at low V_{dd} where the headroom for variation of the cell virtual ground due to random Vt variations becomes smaller with scaling. We propose an alternative where the virtual ground of the SRAM subarray is pinned to a predetermined, programmable voltage using a stack of diodes that function as regulators. The area, performance and noise margin penalty is shown to be less than 3%.

Address Gated Pulse Operation

A short (250-300ps) pulse generated off the rising edge of the clock (Fig. 1) samples the least significant address bit. This pulse propagates only through a decode path defined by other address bits. The decode path serves as a 'clock path' from which a sequence of pulses are generated to precharge and evaluate the bitpath in the selected global column (Fig.2). The R/W enable pulse widths (Fig. 1) are expanded by tying to V_{dd} or Gnd the gates of long channel NFETs or PFETs. These expanded R/W pulses, generated in a global column, not only gate the data in/out bus to the selected global column but also demonstrate ample margin to enable only local write buffers or local sense amplifiers (LSAs) when they intersect the short pulse off the clock path selecting a subarray within a global column. (Fig. 3).

Cell Pitch Matched Bitpaths

The local & global BL pairs directly drives a single cell pitch matched cross-coupled SA. Conventional SRAMs (Fig. 4) typically decode one of many columns of BL pairs using pass gates so that the LSA may fit within a larger pitch. At low V_{dd}, however, the voltage drop across the parasitic R of pass gates consumes an increasing fraction of the differential voltage developed by the cell (Fig. 5). Pass gates also add a significant additional parasitic C for the cell to discharge, increasing both BL power and cell delay. A Cell-Pitch-Matched bitpath thus lowers both bitpath power and cell delay, enabling much faster operation at lower V_{dd}. Parasitic I_{read} from cells deselected by the mux is also eliminated.

Self Reverse Biasing (SRB)

A novel SRB SRAM technique (Fig. 6) relieves the reverse bias on all SRAM cells within an accessed sub array by a pulse off the decode path that selects the sub array, so that the virtual ground of the accessed sub array is restored very close to ground before the WL selects a row of cells in that sub array. The SRB circuit PFET, with gate grounded during normal data-retaining operation, acts as a simple regulator raising the virtual ground of unaccessed sub arrays to enable leakage reduction (i) by reverse biasing the source-body junctions of all cell NFET devices, (ii) by a negative gate-source voltage on the leaking access NFET and (iii) by lowering the tunneling probability across the gate oxides of 4 of 5 affected cell devices (Figs. 7,8). Both cell static noise margin (Fig. 9) and area (Fig. 11) penalties are less than 3% and performance is unaffected (Fig. 12) with the SRB NFET sized to ensure that the subarray VGND stays within a few mV of GND. When data stored in the array is no longer desired, the SRB PFET gate terminal (SLP in Fig. 6) is driven to V_{dd}, into 'sleep' mode, raising the virtual ground to over ½ V_{dd} and lowering total leakage by over 90%.

Performance & Energy

Table I lists the access time and total power of the 32 KByte array. E_{bit} is calculated as the P_{total}-T_{access} product per data I/O.

References:

[1] T. Yamada et al, ISSCC Dig. Feb. 2002, pp. 370-371.
[2] S. Agarwala et al, ISSCC Dig. Feb. 2002, pp. 56-57.
[3] A. Bhavnagarwala, et al, JSSC, Vol 36, No. 4, pp. 658-665, Apr 2001.
[4] B. Amrutur et al, IEEE JSSC, Vol. 35, NO. 2, Feb. 2000, pp. 175-185
[5] K. Osada, et al, ISSCC Dig. Feb. 2001, pp. 168-169.
[6] K. Zhang, et al, Proc. Symp. VLSI Ckts., June 2000, pp. 226–227.
[7] A. Bhavnagarwala et al, ICCD, Sept. 2001, pp. 352-357
[8] T. Chappell, et al, ISSCC Dig. Feb. 1991, pp. 50-51.
[9] M. Powell, et al, ISLPED, Aug. 2000, pp. 90-95.

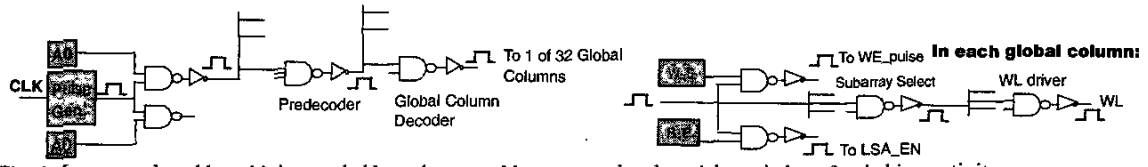


Fig. 1: Lowest order address bit is sampled by pulse to enable a temporal and spatial restriction of switching activity.

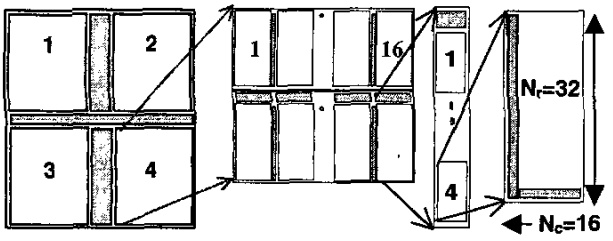


Fig. 2: Array architecture: Each of 4 macros yields 16b of data. A macro is divided into 32 global columns and each global column into 4 subarrays

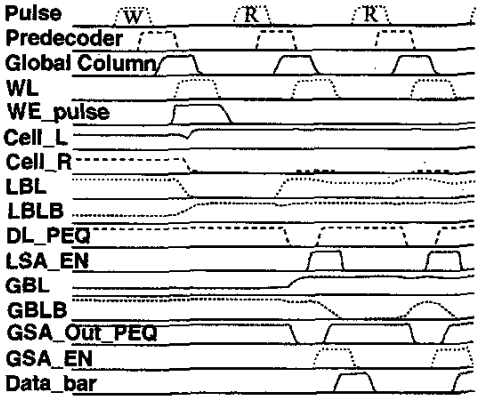
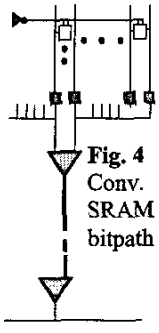


Fig. 3 (at right): The decode path functions as a 'clock' path with the pulse selecting a global column also synchronizing precharge and evaluate of the selected local and global BLs in the global column. Pulse widths decrease at higher V_{dd} permitting bitpath power to show sub-quadratic dependence on V_{dd}



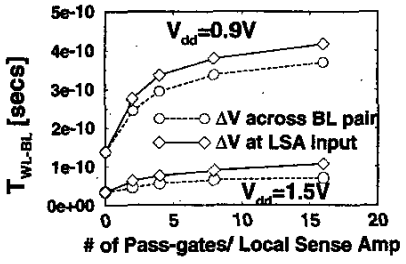


Fig. 5: Cell delay dependence on parasitic C (dots) and on parasitic R and C (solid) of bitline mux

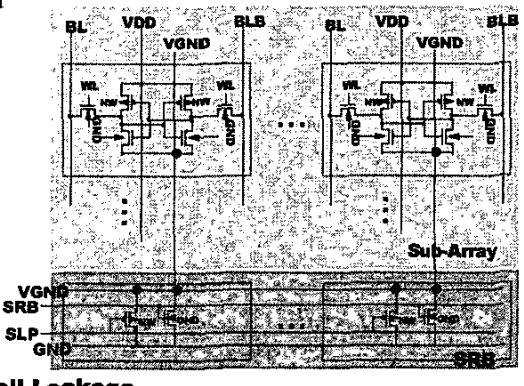


Fig. 6 (above): SRB circuit

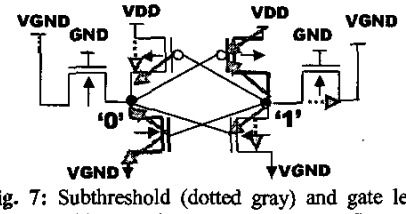


Fig. 7: Subthreshold (dotted gray) and gate leakage paths (solid gray) improved by SRB. BLs settle at VGND. The access Xtr on RHS develops negative Vgs

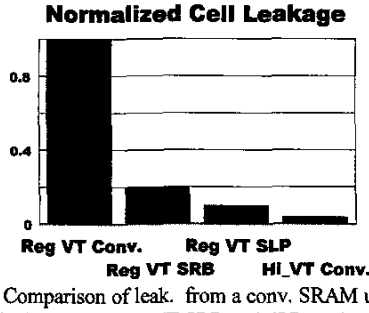


Fig. 8: Comparison of leak. from a conv. SRAM using Reg VT and Hi VT with Reg VT SRB and SLP mode operation

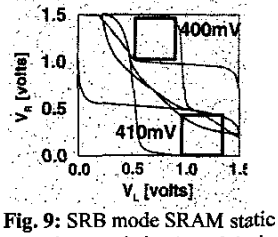


Fig. 9: SRB mode SRAM static characteristics (top right in figure) during standby degrade static noise margin by <3%

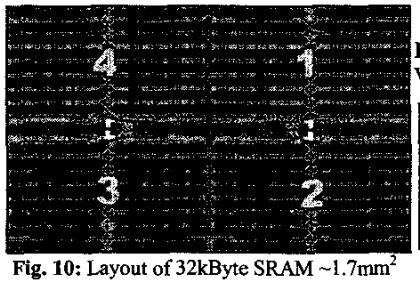


Fig. 10: Layout of 32kByte SRAM ~1.7mm^2

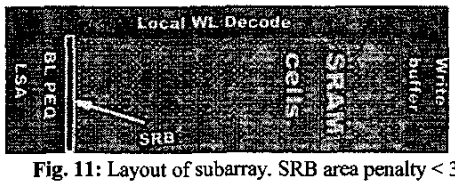


Fig. 11: Layout of subarray. SRB area penalty < 3%

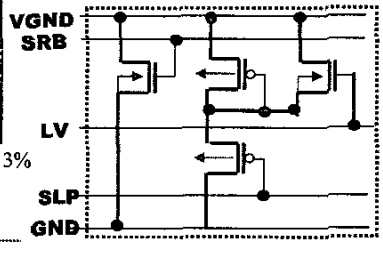


Fig 13: SRB Ckt. with a stack of PFETs programmable to enable SRB operation at nominal and at low voltages

V_{dd} (V)	T_{access} (ns)	Max CLK (GHz)	P_{total} @ Max CLK (mW)	E_{bit} (pJ)	I_{leak} (μA) w/SRB
0.9	1.91	0.50	< 14 mW	0.42	14.92
1.5	0.97	1.00	< 90 mW	1.36	46.55

Table I: 32kByte Single-V_{dd}, standard Vt, SRAM performance, power and energy/bit from simulations of extracted netlists with 33% & 25% activity on address and data respectively. I_{leak} @ 50°c

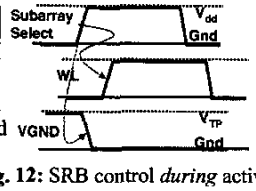


Fig. 12: SRB control *during* active mode