## TP 16.2 An 833MHz 1.5W 18Mb CMOS SRAM with 1.67Gb/s/pin

Harold Pilo, Archie Allen, Jim Covino, Patrick Hansen, Steve Lamphier, Chris Murphy, Terry Traver, Pui Yee

IBM Microelectronics, Essex Junction, VT

An 18Mb CMOS SRAM operates at 833MHz with 1.67Gb/s/pin. The 114.4mm$^2$ die consumes 1.5W and is fabricated in a 0.18µm CMOS process with four levels of copper interconnect [1]. The SRAM operates in two user-selectable double-data-rate modes (DDR and DDR2). Figure 16.2.1 summarizes chip features.

High-frequency operation is achieved by solving three frequency-limiting issues identified in previous SRAM designs [2]: managing data timing constraints associated with high-frequency operation in a high density SRAM core; maintaining coherency between SRAM output data timings and echo clock timings; and delivering symmetric data windows for 1s and 0s across a wide range of output driver supply levels.

Figure 16.2.2 shows a chip micrograph of the SRAM and Figure 16.2.3 shows a block diagram of the architecture. The core is divided into eight octants, each containing 32 subarrays that share a common data bus. For each operation in DDR x36 mode, one subarray in each octant is selected to provide a total of 72 bits for all eight octants. The 72 bits are prioritized by the data mux, stored in two sets of 36 latches, and clocked out of the SRAM in two subsequent clock cycles. Three stages of dynamic-precharged complementary data lines form the data bus. One of the two complementary data lines is shown in Figure 16.2.3 (precharge circuits omitted for clarity). For each selected subarray, one of two sense amps is decoded per data line. The decoded sense-amp output pulses an nFET device, such as 7L in Figure 16.2.3, that drives the first data-bus stage across eight subarrays. The second stage consists of four pFET drivers. A decoded subarray activates one of the four pFET drivers at any given time. Finally, an nFET-driver stage provides the data to the data mux. This last stage may also be accessed by data from either the redundant array or the write buffer [2]. A 512-wordline by 144-bitline subarray structure achieves high-frequency operation and supports a high level of array efficiency. This arrangement provides sufficient write and read bitline recovery margins at the minimum required cycle time.

The architecture addresses two problems that limit the minimum SRAM cycle time. The first problem is the potential for read data-bus contention between the fastest and slowest addressed subarrays within an octant; Subarray A (fastest) and Subarray B (slowest) are shown in Figure 16.2.2. Consecutive read cycles cause read data-bus contention when the slowest and then the fastest subarrays are addressed. This contention results in current cross-overs and potential corruption of data. The second cycle-limiting problem occurs when read data from one octant is multiplexed with data from other octants. Read data windows are not equivalent when the fastest to slowest subarrays are addressed. To prevent data glitches, mux inputs must switch while the read bus is not active. Non-equivalent data may eliminate the mux-input switch window as cycle time is reduced. To alleviate these cycle-limiting issues, access times from all memory locations are made equal. This is accomplished in each subarray by timing the sense-amp data strobe from a global dummy wordline [3]. Access times are made equal across multiple subarrays, in part, by adding delays to the decoding block that addresses the lower half of all octants (see the subarray decode blocks L and U in Figure 16.2.3). The difference in access times between the fastest and slowest subarrays is minimized to allow 833MHz operation.

In DDR SRAMs, cycle time is also limited whenever latency exceeds cycle time. Any further decreases of cycle time can result in two failure mechanisms: 1) output data glitches resulting in system failures, and 2) loss of tracking between the output data and echo clocks resulting in setup and hold violations at the receiving device. Adjustable hard-coded timings are used in previous SRAMs to address these issues [2]. As a more flexible solution, a cycle-independent data to echo-clock tracking system automatically adjusts data and echo latch timings. The system adjusts four independent data and echo latch clocks by exact amounts which are determined by the relationship between the latency of the SRAM and the cycle time at which the SRAM is operating. The latch clocks are adjusted to set the latches after data from the array has arrived. This adjustment prevents the latch access from depending on the array data timings. A small separate array with circuits and wiring tracks identical to those of the slowest access path provides a timing reference. The small array is selected every cycle and tracks with the slowest memory access. The timing of the current cycle output of the small array is compared with the timing of the next cycle SRAM clock. A circuit that makes the timing comparison with 90ps resolution is shown in Figure 16.2.4. A series-connected latch chain is reset before the comparison. The arrival of data from the small array (MDATA) disables delay lines DELAY8-DELAY0 serially, starting with DELAY8. The arrival of the next cycle clock generates the STOP signal, which halts any further disabling of delay lines and registers the results. The longer the delay from MDATA to CLOCK, the more delay lines are disabled, decreasing the amount of delay added to the data-latch and echo-latch clocks. At faster cycle times, fewer delay lines are disabled, corresponding to a greater amount of delay that is added. The data-latch and echo-latch clock driver with variable delays is shown in Figure 16.2.5. Each variable delay path is tuned so that its delay corresponds to a latch-element delay (Figure 16.2.4). nFET T0 provides the fastest DOCLK path when delays are not necessary. nFETs T1 and T2 provide the path for the variable delays.

The output-buffer topology shown in Figure 16.2.6 guarantees by symmetry maximum data windows independent of data polarity or hi-z operation. Balancing the output data latch (when driven by DOCLK) provides DATAT and DATAC signals crossing at the 50% point. The formatting circuitry offsets the controls to the output stage to prevent cross-over currents while also translating the levels applied to the output stage using cross-coupled pFET structures. These pFET structures generate both the PDRIVE and NDRIVE signals to ensure matched timings. Pass devices in the output stage stagger the turn-on and turn-off switching of the output devices to minimize switching noise. Symmetric data pulses and output edge rates are possible due to symmetry in the turn-on and turn-off staggering. One of six pairs of default and impedance-programmable pFET output devices and one of five nFET pairs are shown in Figure 16.2.6.

Figure 16.2.7 shows hardware waveforms at 833MHz core frequency and 1.67Gb/s/pin I/O frequency. The waveforms also show the exact placement of echo clocks (CQ and $\overline{CQ}$) with respect to the data (DQ0 and DQ1).

*References*:

[1] Crowder, S., et al., "A 0.18µm High-Performance Logic Technology," Symposium on VLSI Technology Digest, pp. 105-106, June 1999.
[2] Braceras, G., et al., "A 940MHz Data-Rate 8Mb CMOS SRAM,"ISSCC Digest of Technical Papers, pp. 198-199, Feb. 1999.
[3] Pilo, H., et al., "A 300MHz, 3.3V 1Mb SRAM Fabricated in a 0.5µm CMOS Process," ISSCC Digest of Technical Papers, pp. 148-149, Feb.1996.

 

| | |
|---|---|
| **Organization** | **512kb x 36, 1Mb x 18** |
| **Performance (1.8V / 25°C)** | |
| **•Cycle Time** | **833MHz** |
| **•Data Rate** | **1.67Gb/s/pin** |
| **•Flow-Through Access** | **2.73ns** |
| **Average Core Power** | **1.5W at 833MHz, 1.8V** |
| **Power Supply** | **2.5V or 1.8V** |
| **I/O Interface** | **1.1V - 1.9V HSTL** |
| **Technology** | **0.18µm CMOS w/ 4 Cu Metal Levels** |
| **•$L_{effn}$ / $L_{effp}$** | **0.97µm / 0.117µm** |
| **•$T_{ox}$ (dual)** | **2.8nm / 4.0nm (physical)** |
| **•Operating Voltage** | **1.5V (regulated)** |
| **Cell Size** | **2.52 x 1.68 = 4.23µm$^2$** |
| **Chip Size** | **8.50 x 13.456 = 114.4mm$^2$** |
| **Array Efficiency** | **69.9%** |
| **Package** | **7x17 / 9x17 PBGA (C4 flip-chip)** |

**Figure 16.2.1: 18Mb SRAM features.**

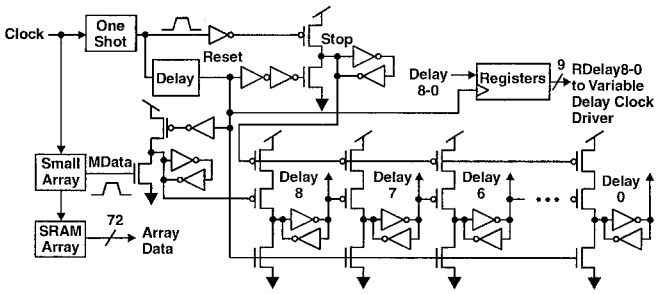


**Figure 16.2.4: Latency to cycle time compare circuit.**

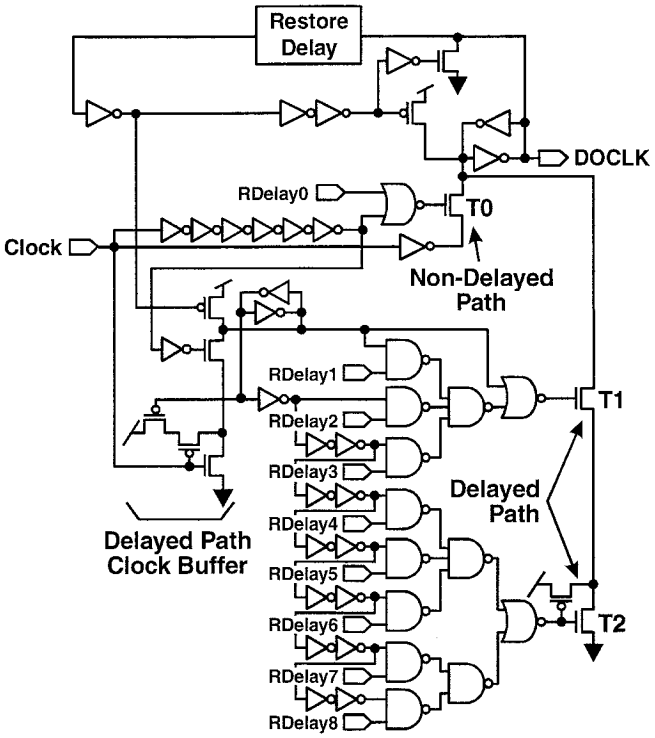


**Figure 16.2.5: Clock driver with variable delays.**

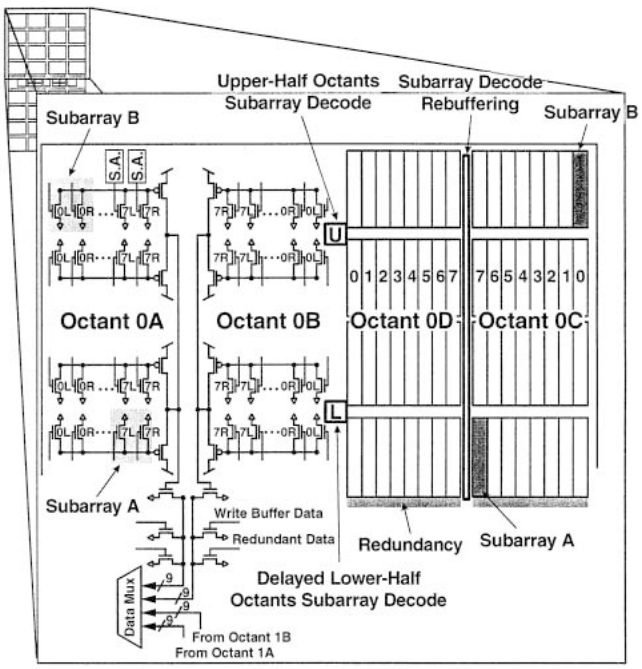


**Figure 16.2.3: Block diagram of architecture.**

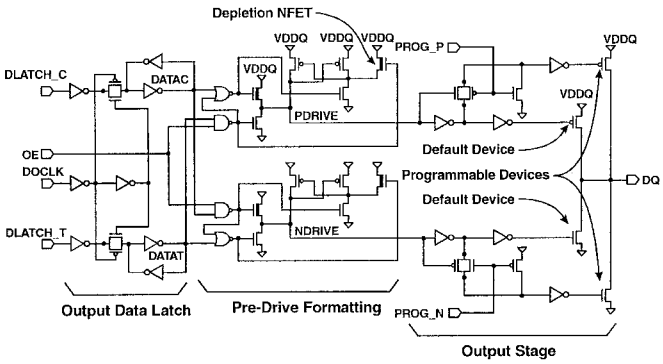


**Figure 16.2.6: Output buffer circuit.**

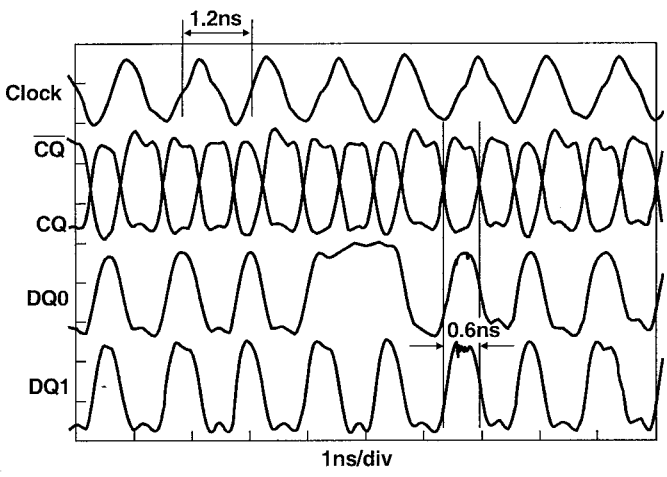


**Figure 16.2.7: Hardware results showing DDR1 operation at 833MHz with 1.67Gb/s/pin data rate.**

 

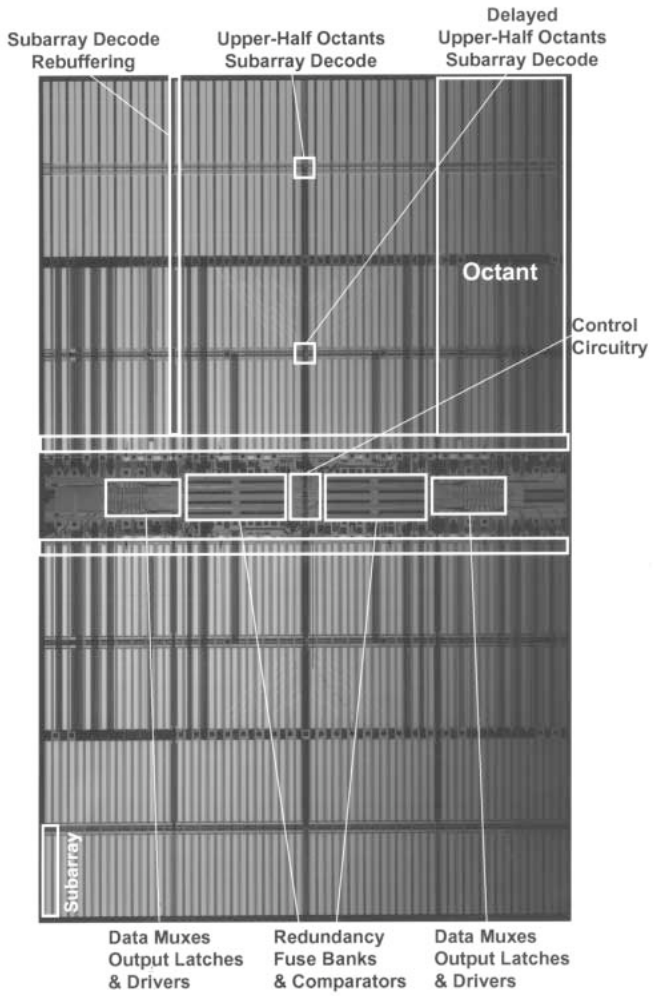


**Figure 16.2.2: Micrograph of 18Mb SRAM.**