TA 11.6 An 18Mb, 12.3GB/s CMOS Pipeline-Burst Cache SRAM with 1.54Gb/s/pin

Cangsang Zhao, Uddalak Bhattacharya, Martin Denham, Jim Kolousek, Yi Lu, Yong-Gee Ng, Novat Nintunze, Kamal Sarkez, Hemmige Varadarajan[1],

Portland Technology Development,
[1]Manufacturing Product Group, Intel Corp., Hillsboro, OR

This 18Mb pipeline-burst cache SRAM has 12.3GB/s data transfer rate. The 14.3x14.6mm^2 chip uses a 5.6μm^2 (2.22 x 2.52) 6-transistor cell and is fabricated on 0.18μm 6-metal-layer, 1.3-1.5V CMOS [1]. Figure 11.6.1 shows the chip floor plan and array architecture. The chip is divided into four quadrants, each composed of 19 global subarrays, including one for redundancy. A data transfer rate of 1.54Gb/s/pin on each I/O is achieved with a 770MHz 50% duty cycle clock at 1.5V, 25°C. Data transition is on both clock edges using dual-edge-triggered flip-flops. With an 8B I/O port (64-data and 8-parity pins), the chip has an aggregate bandwidth of 12.3GB/s. A cycle on this SRAM actually refers to one half period (or a phase) of the clock. Chip features are summarized in Table 11.6.1. Figure 11.6.6 is the die micrograzph.

Circuit techniques used in this SRAM for high bandwidth are: 1) an asynchronous core with synchronous I/O interface and adjustable latency; 2) a hierarchical sensing scheme with separated global read/write bitlines in different metal layers; 3) a data-capture technique allows high-speed operation; 4) an output buffer with both Vcc/2 and Vcc/4 swing options; and 5) a two-staged high-sensitivity, high-bandwidth input buffer.

An asynchronous array access using self-timings reduces circuit complexity, area, and power. Self-timed signals include wordlines, read mux enables (or write mux enables in write), sense-amp-timer triggers, and sense-amp enables. The part has an X-1-1-1 simple pipeline-burst operation, similar to a 4-1-1-1 pipeline-burst SRAM, but with "X" or latency being adjustable from 2 to 8 cycles [2]. The part can be wave-pipelined for latency settings above 4, which allows exploitation of the highest bandwidth that minimum array cycle time allows. Adjustable latency uses a bubble counter which delays the initial address-strobe pulse (ADS) by a set number of cycles. The delayed pulse LATENR captures the global sense-amp output data. Four chunks of data are then serialized and driven out by the output drivers. Figure 11.6.2a shows the simulated read-access timing waveforms. Figure 11.6.2b shows the measured output data waveform.

The asynchronous array design minimizes clock usage and therefore power in this SRAM. Clocks are distributed only in control, address input, and I/O interface areas. An on-chip phase-locked loop (PLL) provides a 50% duty-cycle clock with a 37ps peak-to-peak jitter at 800MHz with a low-power pattern. A balanced H-tree clock distribution network with 6 stages of inverter buffers gives measured worst-case clock skew and duty cycle of 100ps and 44%, respectively.

Hierarchical sensing, similar to a hierarchical sensing scheme with double global bitline pairs, is used for fast array access and small array-cycle time (Figure 11.6.3) [4]. The 0.18μm, 6-metal-layer technology is exploited to separate the global bitlines into read and write bitlines on Metal 4 and Metal 6, respectively, with local bitlines on Metal 2. This allows separate optimization of global read and write, and improves array efficiency. Since small-signal development is capacitance dominated, Metal 4 is used in global read bitlines for lower capacitance, compared to Metal 6 (Metal 6 is thicker than Metal 4). Metal 6 is used for global write bitlines for its minimum resistance. With a single pair of global read bitlines fitting in an 8-cell column space of 17.8μm, minimum-width wires with large spacing are used for minimum wire capacitance. In this line-to-line capacitance dominated metal system, 26% wire-capacitance reduction can be realized by increasing the Metal-4 wire spacing from minimum 0.52μm to 1.42μm [1]. Global write bitlines are wider at the near end of the write drivers (for smaller wire resistance) and narrower at the far end (for smaller wire capacitance). This tapered wire structure minimizes delays on the global write bitlines. Asynchronous array access time is 2.5ns at 1.5V and 25°C. With the part run at 770MHz and latency set to 4, measured chip access time is 3.0ns.

To realize a high data rate, the part uses a source synchronous I/O interface, in which a data-capture circuit captures four chunks of input data one-by-one in series while requiring an ADS-delayed pulse to arrive at the I/O port within one cycle [4]. To use this method at 1.5Gb/s/pin and above, it would be difficult to have an ADS-delayed pulse arrive at the I/O port within one cycle (<0.67ns), especially on an 18Mb SRAM. A data-capture technique does serial-to-parallel conversion first so data capture is at lower frequency, and only requires an ADS-delayed pulse to arrive at the I/O port within four cycles (2.7ns at 1.5Gb/s/pin) (Figure 11.6.4). Similar to the data sampling scheme in Reference 4, the combinational logic from a pair of circulating counter outputs (even and odd) and the data clocks STRB/STRB# is used to determine which incoming serial chunk is to be sampled. Once a data chunk is sampled, it is held in its latch for 3.5 cycles. All four data chunks are then alligned and parallelized (delayed further if necessary with extra hardware) so that data capturing takes place simultaneously at a later time through the ADS-delayed pulse: LATENW. The captured four data chunks are then written into the array in parallel.

The data output drivers are CMOS tri-state buffers as described in Reference 4 but with 7 binary weighted legs for higher resolution in the impedance control. When driving a 50Ω terminated load, the signal swing is expected to be Vcc/2. In such a case, the switching noise of 72 I/Os is excessive in simulation at data rates above 1.6Gb/s/pin. Therefore, an optional configuration shown in Figure 11.6.5a, is used to reduce the driver strength to 100Ω and the signal swing to Vcc/4. A DC pin signal determines the driver options by selectively turning on or off driver and terminator legs. Simulations based on a package and power supply network model show the output driver runs at 1.6Gb/s and 2.5Gb/s with <57ps and <99ps inter-symbol interference, respectively.

The reduced signal swings (Vcc/2 and Vcc/4) require a highly sensitive input buffer with good noise rejection characteristics. The two-stage CMOS input receiver is shown in Figure 11.6.5b. The first stage is a differential amplifier with differential outputs. This stage achieves bandwidth using a small resistance load. A separate bias, similar to a self bias, in this stage achieves stable biasing over process, supply, and temperature variations [5]. The second stage is a self-biased differential amplifier with medium gain [5]. This input buffer has a worst-case sensitivity of 50mV at a data rate of 1.7Gb/s in simulation. The reference voltage, V_{ref}, is sent to the die along with the input data to cancel out common-mode noise.

Acknowledgments:
The authors thank J. Greason, the former manager of this design team, for invaluable contribution, A. Bell, B. Benefiel, J. Chuang, M. Acken, and Q. Duong for layout, W. Holt, I. Young, W. Chen, G. Taylor, C. Webb, and T. Thomas for comments, STTD Sort group and PTD LYA group for test/debug support, and those who supported this project in Intel.

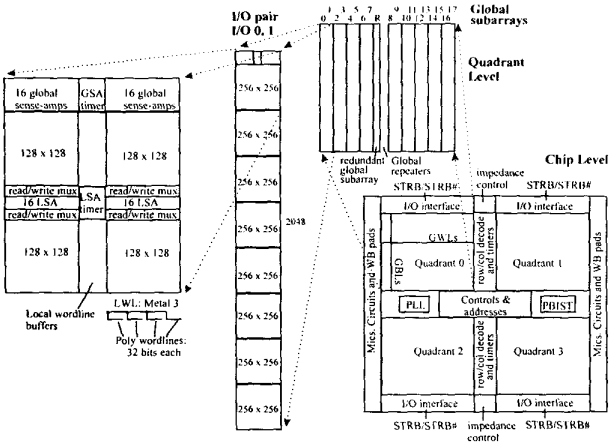


Figure 11.6.1: Chip floor plan and array architecture.

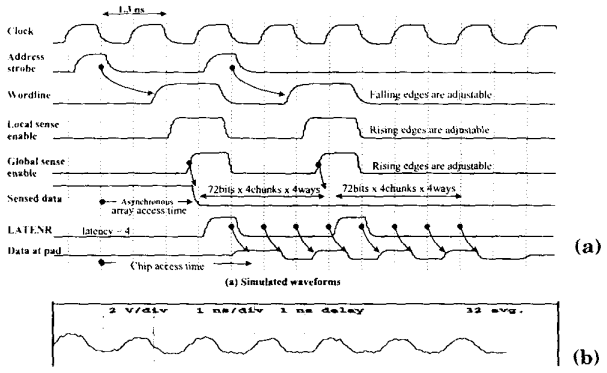


Figure 11.6.2: (a) Simulated read-timing waveforms. (b) Measured output data waveform.

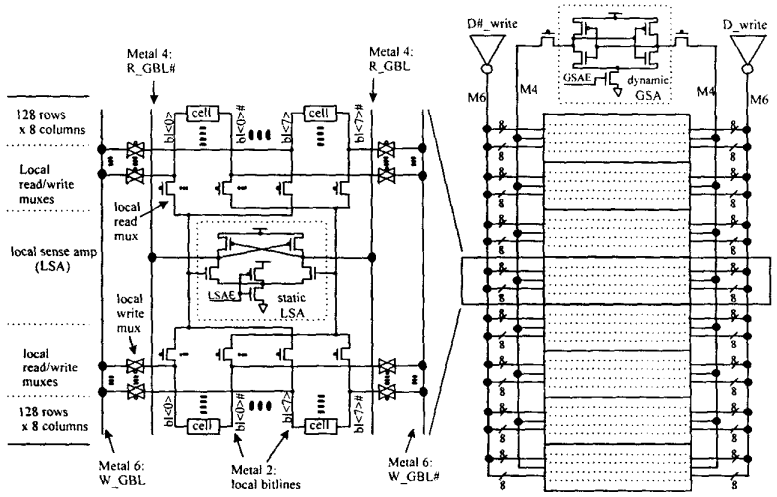


Figure 11.6.3: Hierarchical sensing with separated global read/write bitlines.

Organization	256kx72b, 4:1 burst
Supply	1.5V
I/O protocols	Source synchronous
I/O interface	CMOS
Data-transfer	12.3GB/s
Array access	2.5ns asynchronous
Chip access time	3.0ns (address latch to data out)
Cell size (6T)	5.6μm²
Chip size	14.3x14/6mm²
Active power	7.2mW at 770MHz
I/O buffer power	0.9mW (simulated at 770MHz)
CMOS technology	6-metal, 3.0nm Tox, 0.14μm L_{gate}

Table 11.6.1: SRAM features.

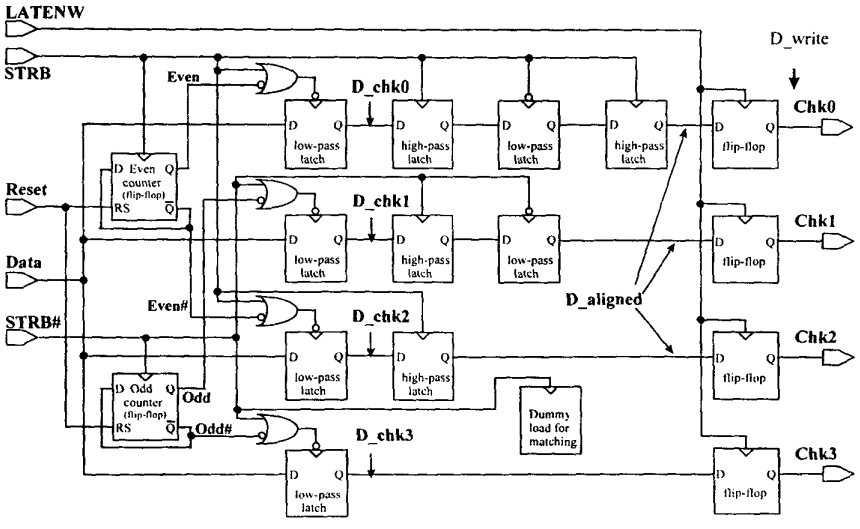


Figure 11.6.4: Logic diagram for data capturing.

References:

[1] Yang, S., et al., "A High Performance 180nm Generation Logic Technology," IEDM, Digest of Technical Papers, 1998.

[2] Nakamura, K., et al., "A 500Mhz 4Mb CMOS Pipeline-Burst Cache SRAM with Point-to-Point Noise Reduction Coding I/O," IEEE J. Solid-State Circuits, vol. 32, no. 11, pp. 1758-1765, Nov., 1997.

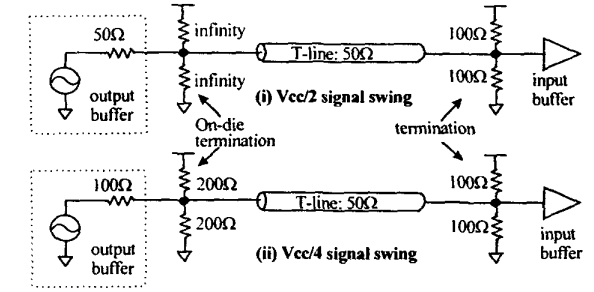
[3] Osada, K., et al., "A 2ns Access, 285MHz, Two-Port Cache Macro using Double Global Bit-line Pairs," ISSCC Digest of Technical Papers, pp. 402-403, Feb., 1997.

[4] Taylor, G., et al., "A 2MB 3.6GB/s Backside Bus Cache for IA32 450MHz Microprocessor," Symposium on VLSI Circuits, Digest of Technical Papers, pp. 184-185, June, 1998.

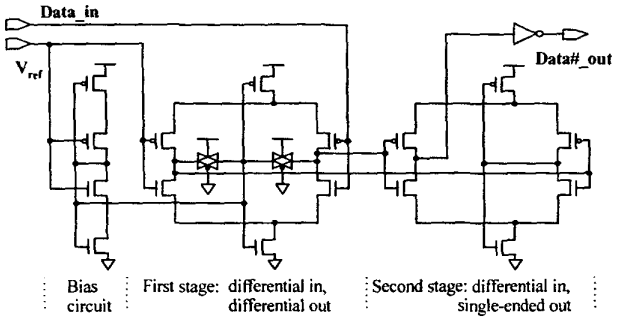
[5] Bazes, M., "Two Novel Fully Complementary Self-Biased CMOS Differential Amplifiers," IEEE J. Solid-State Circuits, vol. 26, No. 2, pp. 165-168, Feb 1991.

Figure 11.6.5: See page 461.

Figure 11.6.6: See page 461.



(a) Output driving options: (i) Vcc/2 signal swing; (ii) Vcc/4 signal swing



(b) High-gain and high-bandwidth CMOS input receiver

Figure 11.6.5: I/O buffers.

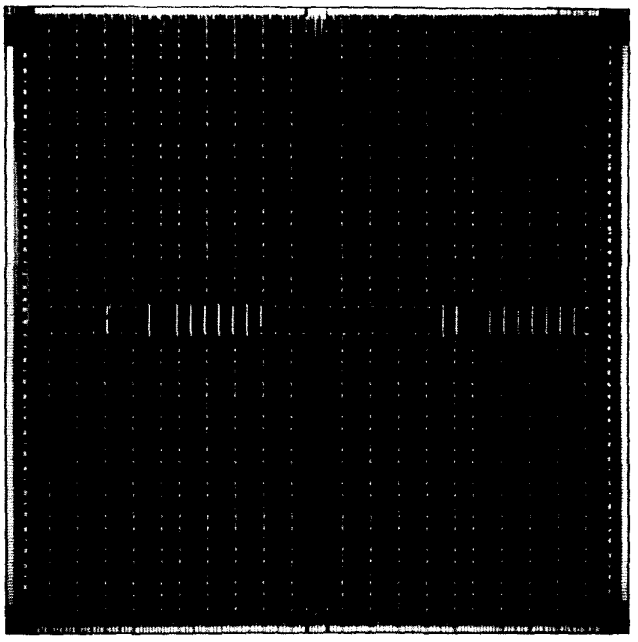


Figure 11.6.6: Die micrograph.

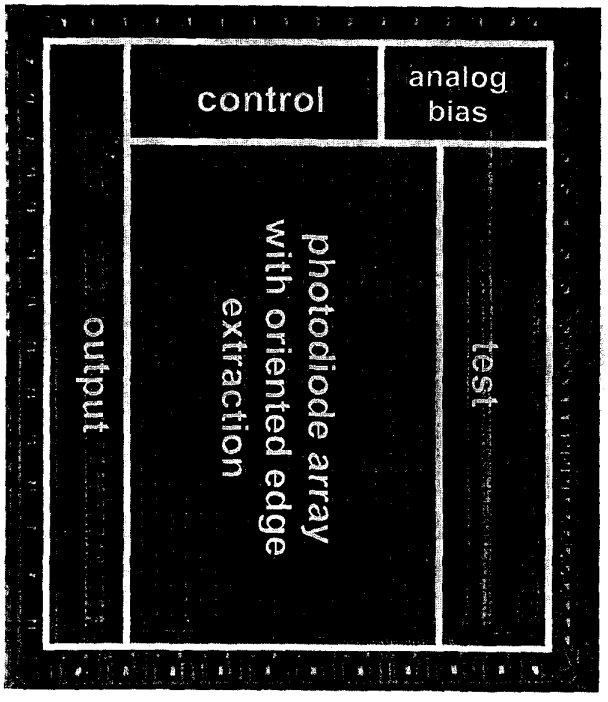


Figure 12.1.6: Micrograph of the retina chip.

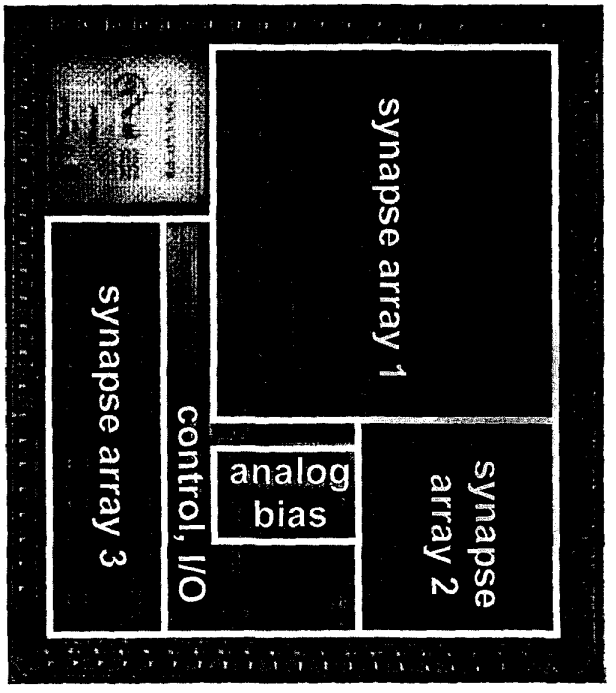


Figure 12.1.7: Micrograph of the classifier chip.