A 14-ns 1-Mbit CMOS SRAM with Variable Bit Organization

YOSHIO KOHNO, TOMOHISA WADA, KENJI ANAMI, MEMBER, IEEE, YUJI KAWAI, KOJIRO YUZURIHA, TAKAYUKI MATSUKAWA, AND SHIMPEI KAYANO

***Abstract*—This paper will describe a 14-ns 1-Mbit CMOS SRAM with both 1M word × 1-bit and 256K word × 4-bit organizations. The desired organization is selected by forcing the state of an external pin. The fast access time is achieved by use of a shorter divided-word-line (DWL) structure, a highly sensitive sense amplifier, a gate-controlled data-bus driver, and a dual-level precharging technique. The 0.7-μm double-aluminum and triple-polysilicon process technology with trench isolation offers a memory cell size of 41.6 μm^2 and a chip size of 86.6 mm^2.**

I. Introduction

With recent advances in process technologies and new circuit techniques, high-speed, high-density SRAM's have been developed. These RAM's have been used mainly as main memory in supercomputers, cache/buffer memory in minicomputers and workstations, and as test pattern memory in VLSI test equipment. Recently, 256K CMOS SRAM's were reported which achieved access times of faster than 25 ns [1]–[3] by utilizing the address transition detection (ATD) techniques, configurations with shortened bit lines and word lines, polycide gates, and double-aluminum processes. Several medium-speed 1-Mbit CMOS SRAM's with byte-wide organization have already been reported [4]–[7], however, their access times are above 25 ns. This paper will describe a variable bit-organization 1M word × 1-bit or 256K word × 4-bit CMOS SRAM [8] with a typical access time of 14 ns. This RAM was fabricated employing double-aluminum and triple-polysilicon CMOS technology with a 0.7-μm minimum design rule. In order to obtain the fast access time, a 32-block architecture with a divided-word-line (DWL) structure [9], a highly sensitive sense amplifier, a gate-controlled data-bus driver, and a dual-level precharging technique were combined with an ATD scheme.

In Section II, circuit technologies, which include the chip architecture, the modified DWL structure, the sense amplifier, the gate-controlled data-bus driver, the dual-level data-bus precharge, and the variable bit-organization function will be described. In Section III, the process technology will be explained. Characteristics of the RAM and conclusions are given in Sections IV and V, respectively.

Manuscript received March 17, 1988; revised May 23, 1988.
The authors are with the LSI Research and Development Laboratory, Mitsubishi Electric Corporation, 4-1 Mizuhara, Itami, Hyogo 664, Japan.
IEEE Log Number 8822500.

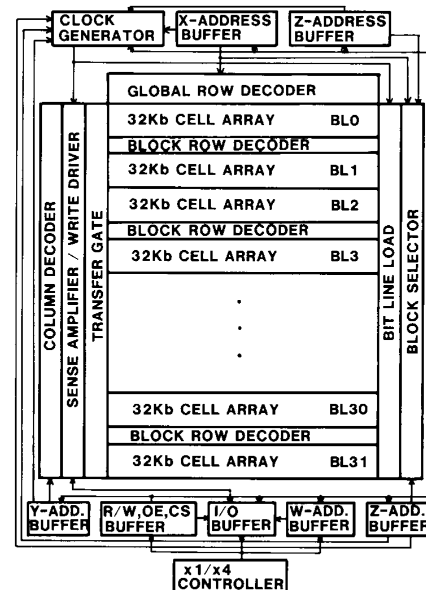

Fig. 1. Block diagram of RAM.

II. Circuit Design

A. Chip Architecture

The block diagram of the RAM is illustrated in Fig. 1. The RAM is organized as either 1M word × 1 bit or 256K word × 4 bit. These organizations are controlled by the × 1/ × 4 control pin. The address signals are split into four groups—*X*, *Y*, *Z*, and *W*. The *X*-, *Y*-, *Z*-, and *W*-address signals are used for row selection, column selection, block selection, and sense-amplifier selection, respectively. The *W* address is only used in the 1M word × 1-bit organization. Each address input buffer has a local ATD pulse generator. A detection signal from any of the local ATD pulse generators activates the internal clocks which control the bit-line loads and the sense amplifiers in order to accelerate a readout operation.

A block diagram of the memory architecture is shown in Fig. 2. The 1-Mbit memory cell array is organized as 512 rows × 2048 columns and is divided into 32 blocks. Only one memory cell block is activated at a time for power

0018-9200/88/1000-1060$01.00 ©1988 IEEE

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

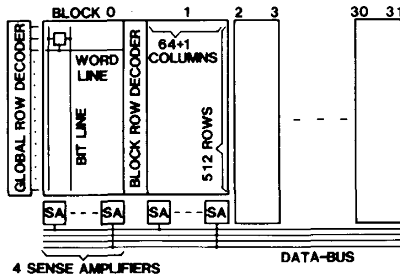

Fig. 2. Block architecture.

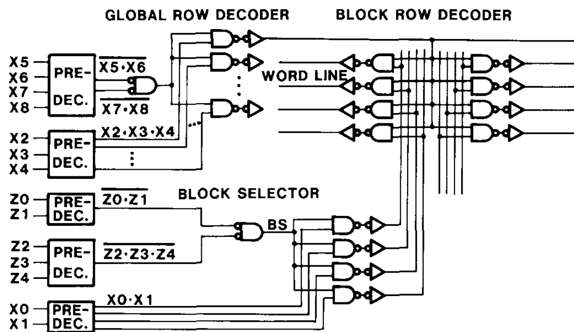

Fig. 3. Word-line selection circuit.

reduction. Each block contains 512×64 columns with one redundant column. Only 65 memory cells are connected to a word line (DWL). The word line has been formed from tungsten-silicide polycide with a sheet resistance of 5 Ω/sq. The short word line with low-resistive material reduces the word-line delay time to around 0.5 ns, which is one key factor in reducing access time. Each block is further divided into four submatrices of 512 rows×16 columns, compatible with the nibble-wide organization (256K word ×4 bit). Row address input signals are hierarchically decoded, using the DWL structure [9]. The block row decoders of the DWL structure are arranged between every two blocks, resulting in a total of 16 block row decoders. The global row decoder is placed at one side of the memory array.

B. Modified DWL Structure

Fig. 3 shows a word-line selection circuit in the modified DWL structure introduced for improvement of the conventional DWL. $X0-X8$ and $Z0-Z4$ are address signals for row selection and block selection, respectively. The upper X address group of $X2-X8$ is predecoded in the global row decoder which activates one of the row-group select lines. The row group consists of four rows. Each of the row-group select signals is input to eight NAND gates of each block row decoder. One of the four rows is selected by the predecoded signals of the lower addresses of $X0$ and $X1$. Fig. 4 shows the effect of the modified DWL structure on the capacitance of the bit line and row select

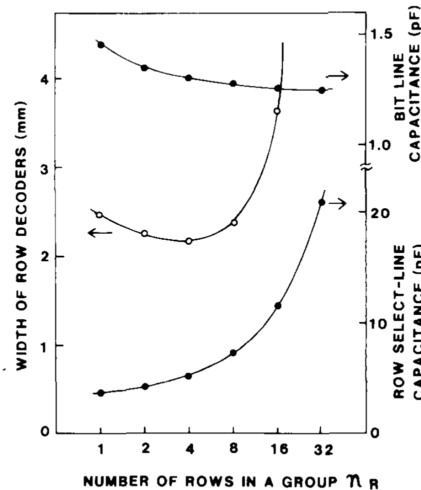

Fig. 4. Effect of modified DWL.

line, and the width of the row decoders (including the global row decoder and the block row decoders). The bit lines and the row select lines are fabricated by the first- and second-level aluminum, respectively. As the number of rows in a group (n_R) increases, the bit-line capacitance decreases because the crossover capacitance with the row select lines decreases. However, the slope of the curve is very broad, especially beyond 4 of n_R. In contrast, the capacitance of the row select lines increases rapidly as n_R increases. Moreover, the capacitance on the predecoded signals of $X0$ and $X1$ decreases. On the other hand, the width of the global row decoder decreases and the width of the block row decoder increases as n_R is larger. Therefore, the width of the row decoders is minimized at 4 of n_R. Consequently, this modified DWL structure has been applied to the 1-Mbit SRAM, setting n_R at 4. The modified DWL architecture with the small bit-line capacitance contributes to the fast access time by a 1.5-percent decrease of chip size compared to the conventional DWL.

C. Sense Amplifier

Fig. 5 shows the readout circuitry, which includes the sense amplifier, data-bus driver, dual-level data-bus precharger, and data-output buffer. The sense amplifier is comprised of three stages. The first and second stages are symmetric types, which have dual inputs and dual outputs. This symmetric sense amplifier is suitable to amplify the I/O line's signals with a small voltage swing. Therefore, both ZERO READ and ONE READ access times coincide with each other. The third stage is a normal current mirror circuit, which has dual inputs and a single output. Each stage has a gain of around 3. In total, a voltage difference of less than 50 mV between I/O lines is successfully amplified. The pulses ($SEQ1$, $SEQ2$, $SEQ3$, DEQ, and $\overline{DEQ}$) are generated from an internal clock, which is caused by the ATD signal. The signals $SEQ1$ and $SEQ2$

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

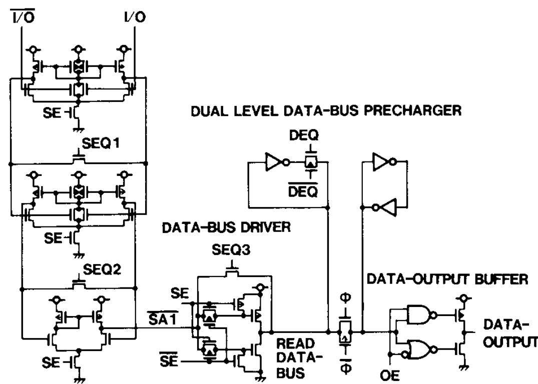

Fig. 5. Readout circuit.

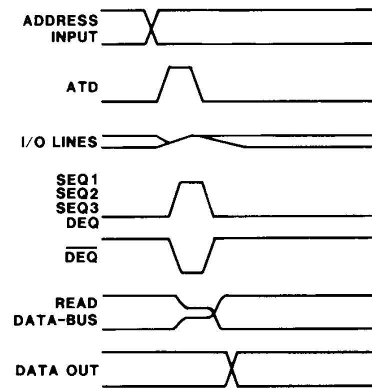

Fig. 7. Timing diagram.

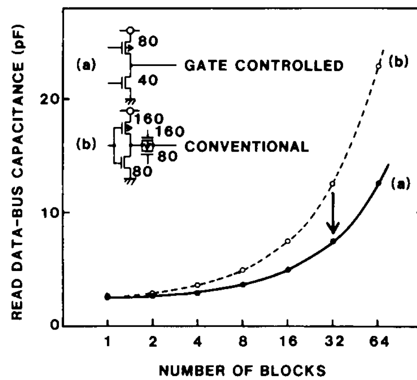

Fig. 6. Relation between data-bus capacitance and number of blocks.

equalize the complementary signals in the sense amplifier. Bit lines and I/O lines are also equalized by this internal clock. These equalizations accelerate the sensing speed.

D. Gate-Controlled Data-Bus Driver

The data-bus driver is also shown in Fig. 5. The data-bus driver is a tri-state CMOS inverter whose gates are controlled by the sense-enable signal SE. If it is selected, the SE signal is high and the $\overline{SE}$ signal is low. Accordingly, this data-bus driver works as a CMOS inverter with the same drivability. As shown in Fig. 2, when the memory cell array is divided into many blocks, the parasitic capacitance of the data bus becomes larger, because of the long wiring capacitance and the output capacitance of the data-bus driver. Fig. 6 shows the relation between data-bus capacitance and the number of memory cell blocks. The solid line (curve a) and the broken line (curve b) correspond to the gate-controlled tri-state data-bus driver and a conventional tri-state data-bus driver with a CMOS transmission gate [5], respectively. In order to maintain the drivability of the data-bus driver, the channel width of the conventional data-bus driver is doubled. The data-bus capacitance increases gradually as the number of blocks increases. In the 32-block architecture, which is applied to the 1-Mbit SRAM, the gate-controlled data-bus driver reduces the data-bus capacitance by around 40 percent. This reduction is effective in minimizing the delay time of the data bus.

E. Dual-Level Data-Bus Precharge

In order to further reduce the delay time on the data bus, the data bus is precharged at a middle voltage level [1], [4]. However, a special receiver circuit of the middle-level data bus is required to suppress a wrong data output during the middle-level precharging period. For example, the Schmitt trigger latch STL circuit [1] and the dual-threshold data transfer circuit [4] have been reported. These receiver circuits are connected between the READ data bus and data-output buffer, so that they introduce an undesirable delay time. To remove this delay time in the receiver circuits, the dual-level data-bus precharging technique is newly introduced. This technique does not need any special receiver circuits. The timing diagram of the readout operation is shown in Fig. 7. The pulses ($SEQ3$, DEQ, and $\overline{DEQ}$) are used for the data-bus precharging. The precharged level of the data bus is split into two levels according to the previous READ data, which are stored in the small latch circuit connected to the input of the data-output buffer (Fig. 5). The low and high levels are around 2 and 3 V, respectively. As the logical input threshold level of the data-output buffer is set to 2.5 V, a wrong data output during the precharge period is suppressed. The gate-controlled data-bus driver and the dual-level precharge technique realize a high-speed data transfer from the sense amplifier to the data-output buffer.

F. Variable Bit-Organization Function

The testing time of RAM's consists of dc parametric test and ac (functional and operating margin) tests. The ac test time increases in proportion to the memory capacity when N-test patterns (linear addressing patterns) are used. Therefore, the test time of 1-Mbit SRAM's is quadrupled compared with that of 256-kbit SRAM's when the bit organizations are the same. To overcome this problem in DRAM's the test mode [10] is introduced. However, the

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

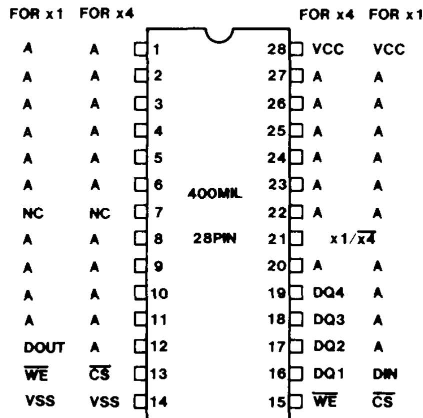

Fig. 8. Pin configuration.

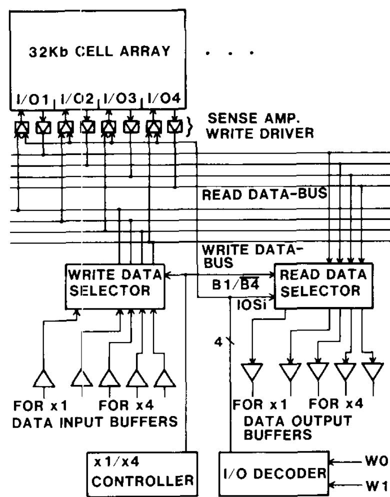

Fig. 9. Circuit schematic of the variable bit-organization function.

test mode in DRAM's is not suitable to measure precise access times, because plural memory cells are written with the same data and only the logical relation among the READ data of these memory cells is checked. This means that the route of the data flow in the test mode is different from that of the normal READ operation. In high-speed SRAM's, the precise evaluation of the access time is important as well as the reduction of test time. The variable bit-organization function is newly applied to reduce the test time while keeping the measurement accuracy of the access time. Fig. 8 shows the pin configurations of the 1-Mbit SRAM. This RAM can be used for 1M word × 1-bit organization or 256K word × 4-bit organization. The inner pin names are used for the 256K word × 4-bit RAM and the outer for the 1M word × 1-bit RAM. These organizations are controlled by pin #21. Pins 12, 13, and 15–19 serve different functions according to the × 1/$\overline{\times 4}$ control signal. All other pins have the same function for both organizations. Fig. 9 shows the circuit schematic of the variable bit-organization function. The memory cell block corresponds to four sense amplifiers/WRITE drivers. Each sense amplifier is connected to the READ data bus. In the case of 256K word × 4 bit, the READ data-bus signals are transferred to the four data-output buffers. In the case of 1M word × 1 bit, one of the READ data-bus signals is transferred to the one data-output buffer according to the I/O select signals (IOSi). The READ data selector circuit is composed of CMOS transmission gates. As the READ data-bus signals are used for both organizations, the route of the data flow is the same. Consequently, the variable bit-organization function preserves the same access time for both organizations. As the 1M word × 1-bit RAM is changed to 256K word × 4 bit, the testing time of the 1M word × 1-bit RAM is reduced while keeping the measurement accuracy of the access time. A short and precise test methodology has been realized.

TABLE I
PROCESS PARAMETERS

Parameter		Value
Process		Twin-well CMOS, N-Sub Double Level Aluminum Triple Level Polysilicon
Gate Length	(NMOS)	0.7µm (LDD)
	(PMOS)	0.9µm (LDD)
Gate Oxide Thickness		180Å
Junction Depth	(NMOS)	0.2µm
	(PMOS)	0.3µm
Trench N^+	(width/space)	1.0µm / 0.7µm
LOCOS N^+	(width/space)	2.0µm / 1.5µm
1st Poly Si	(width/space)	0.7µm / 0.8µm
2nd Poly Si	(width/space)	1.0µm / 1.0µm
3rd Poly Si	(width/space)	1.0µm / 1.0µm
1st Al	(width/space)	1.4µm / 1.0µm
2nd Al	(width/space)	2.0µm / 2.0µm
Contact Hole		0.8µm x 0.8µm
Direct Contact Hole		0.8µm x 0.8µm
Via Hole		1.0µm x 1.0µm

III. 0.7-µm CMOS PROCESS TECHNOLOGY

In order to achieve a high-density and high-speed SRAM, a 0.7-µm CMOS process technology was developed. The SRAM was fabricated with a double-level aluminum and triple-level polysilicon (including polycide) twin-well CMOS technology. The key process parameters are summarized in Table I. Fig. 10 describes the cross-sectional view of the memory cell.

In order to obtain the small-area memory cell, triple-level polysilicon and trench isolation technology is adopted as in the 128K × 8-bit 1-Mbit SRAM [5]. The first polysilicon layer (WSi_x/polysilicon) is used for the MOS transistor gate electrodes. The second polysilicon layer is used for V_{cc} lines and the third polysilicon is for the high-resistive loads. The first aluminum layer is utilized for the bit lines and the second layer is for the row-group select lines. Fig. 11 shows a photograph of the memory cells and block row decoder. The second-level aluminum lines of the row-group

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

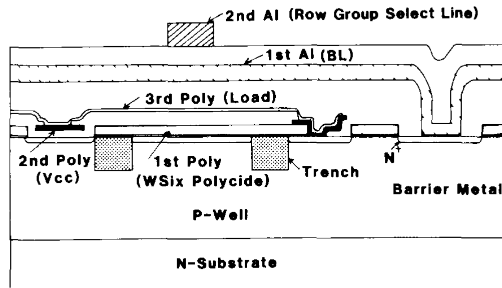

Fig. 10. Cross-sectional view of memory cell.

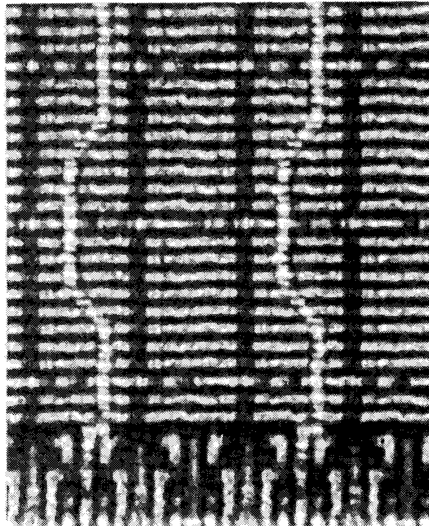

Fig. 11. Photograph of memory cell and block row decoder.

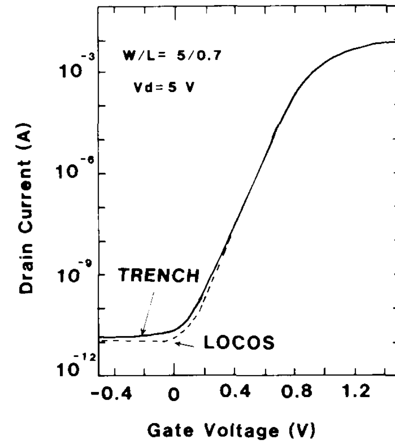

Fig. 12. Comparison of subthreshold current of NMOSFET's ($W/L = 5/0.7\ \mu m \times 10\,000$) isolated by LOCOS and trench.

ADDRESS INPUT

DATA OUTPUT

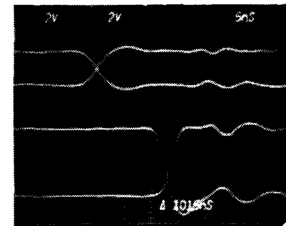

Fig. 13. Output waveforms of RAM.

select line cross the first-level aluminum bit lines every four memory cells. The intermediate insulator between the first and second aluminum is p-CVD oxide. In the peripheral circuits, the second-level aluminum electrode is utilized for signal lines, V_{cc}, and ground lines.

The triple-level polysilicon process enables a smaller memory cell in comparison with the double-level polysilicon process, and the higher resistive load can be obtained easily.

One of the key process technologies in realizing a small memory cell is the bird's beak free isolation. Instead of the conventional LOCOS isolation, a shallow trench isolation technique is adopted. The narrowest isolation width is 0.7 μm. After silicon etching and boron implantation for channel cut, the trench wall is slightly oxidized to reduce leakage current. Then a silicon dielectric oxide is buried in the trench, and a successive etching-back process for surface planarization is carried out. After the trench isolation process, the conventional LOCOS isolation process is employed for peripheral circuit formation. The most important electrical characteristic for the trench-isolated memory cell is the subthreshold-current curve of NMOSFET's. If the level of the leakage current through the driver transistor is the same as or much larger than the current through the high-resistive load connected to the V_{cc} line, the memory cell cannot hold data. Fig. 12 shows the subthreshold current of NMOSFET's ($W/L = 5/0.7\ \mu$m, 10 000 FET's are arranged in parallel) isolated by the conventional LOCOS and by the trench. An explicit difference is not observed for the subthreshold slope and the junction leakage current level.

In this RAM, 0.7-μm NMOSFET's and 0.9-μm PMOSFET's help to achieve the high-speed circuit operation. Both NMOSFET's and PMOSFET's are formed with a lightly doped drain (LDD) structure. This transistor structure assures reliability against hot-carrier-induced degradation for NMOSFET's and prevents degradation caused by the offset phenomenon for PMOSFET's.

IV. Characteristics

Fig. 13 shows oscillographs of the data-output waveforms under typical conditions of $V_{cc} = 5$ V and $T_a = 25°$C. The address access time and the $\overline{CS}$ access times are 14 ns with a load capacitance of 30 pF. An active current of 100 mA (40 MHz), a standby current of 20 mA (40 MHz), and a typical standby current at CMOS inputs of 2 μA have been obtained.

Fig. 14 shows the die photomicrograph. The memory array is divided into 32 blocks to shorten the word-line length. There are 16 block decoders and a global row decoder. Each block has one redundant column (512 cells).

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE

fastcase®
Smarter legal research.