Laser Programmable Redundancy and Yield Improvement in a 64K DRAM

ROBERT T. SMITH, JAMES D. CHLIPALA, JOHN F. M. BINDELS, ROY G. NELSON, FREDERICK H. FISCHER, AND THOMAS F. MANTZ

***Abstract*—Yield improvement obtained with laser programmed redundancy in a 64K DRAM has ranged from 3000 percent during early model making to 500–800 percent after two years of volume production. The electrical design constraints on 64K redundancy organization are reviewed. The explosion and wicking phenomenon of polysilicon links by ~50 ns, 1.064-μm wavelength laser pulses is discussed in relation to the target geometry, laser spot size and targeting accuracy. The system hardware and main software modules are detailed. In particular, the algorithms for testing, repair diagnosis, and target coordinate calculation are explained. Elemental time analysis of the main operational steps is reviewed with emphasis on strategy for improved throughput. Evolution of the laser programming technology to the next generation of VLSI devices involves smaller spot sizes and submicrometer positioning accuracy.**

I. Introduction

During the past two years since first disclosure of fault-tolerant memory designs [1]-[3], technology using a laser to replace defective elements by redundant (spare) rows or columns has evolved greatly [4]. What was originally conceived as a yield improvement aid for the early stages of memory development has matured into an extremely cost-effective wafer fabrication tool for volume production. No longer is there any question of whether to use redundancy in VLSI memories, rather, the current debate is over how much redundancy is appropriate, and whether to use laser programming or electrically fusible links [4]-[10]. An interesting variation on the laser programming approach, wherein the laser is used to connect, rather than to disconnect, circuit elements by rapid thermal diffusion from doped to intrinsic polysilicon has also been reported [11]. Earlier work on laser coding of ROM's [12], [13], and LSI circuit personalization [14], [15] was not aimed at volume production. The work reported herein is the first detailed account of a practical cost-effective laser programming process applied successfully to yield improvement of fault-tolerant VLSI memory production. The original yield incentive [2] has been tested and amply verified in two years of large scale manufacture.

II. Design Constraints on Redundancy Organization

The electrical design details of the Bell System fault-tolerant 64K dynamic RAM have been described elsewhere [1], [2]. Only those details necessary to an understanding of the implementation and use of redundant memory elements are reviewed here. The fault-tolerant design constraints are: 1) a fault-free memory requires no programming action; 2) electrical performance, especially access time, is not degraded by the use of spare elements; and 3) defective spare elements can be replaced by other spares.

The redundancy organization is shown, highly simplified, in Fig. 1. Two spare rows, complete with decoder and driver circuitry, are associated with each 16K quadrant, organized as 64 rows by 256 columns. Either one of these spare rows may replace any one of the 64 main rows in the adjacent quadrant, or may replace each other, if necessary. Four spare columns, including decoder and sense amplifier circuits, are associated with each pair of 16K memory quadrants. Any one of the spare columns may be used to replace any of the 256 columns in the adjacent quadrant pair, or any other previously encoded spare column in the same group. The issue of fault coverage, appropriate type and number of spares, and spare element organization is intimately linked to the nature of the most prevalent defects and their density. This will be discussed further in Section V.

Replacement of a defective memory element, whether row or column, may be understood by referring to Fig. 2, which shows a standard and a spare row decoder schematic. The actual decoder circuitry is somewhat more complex, especially in the column direction. Since the complete circuits and operating principles have been adequately described in [2], we will concentrate here on the decoder interchange by laser disconnection of the faulty memory element, and programming or encoding of a spare decoder. The essential difference between the standard and spare decoder is that the former has half the number of decode transistors of the latter. The identity of the standard decoders is defined by unique connections of address and address complement to the appropriate decode gates. By contrast, the spare decoder has both address and the complement address tied to the gates of decode transistor pairs. Hence, regardless of the applied address, the spare decoder is heavily deselected, satisfying the first design constraint.

Disconnection of a faulty memory row is accomplished by exploding the programmable link between the standard row driver and the row line, using a single laser pulse as described in Section III-B. One of the spare decoders is then encoded or programmed to take on the identity of the faulty element by selectively disconnecting either an address or its complement from the six transistor pairs in the decode node of the spare. This is done by exploding links in the drain connections of

Manuscript received May 8, 1981; revised May 13, 1981.
The authors are with Bell Laboratories, Allentown, PA 18103.

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

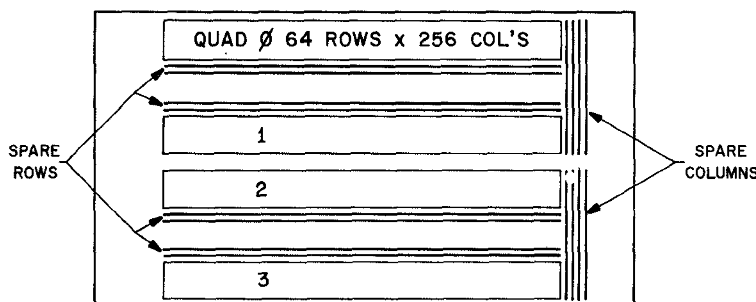


Fig. 1. Schematic layout of 64K DRAM including redundancy organization.

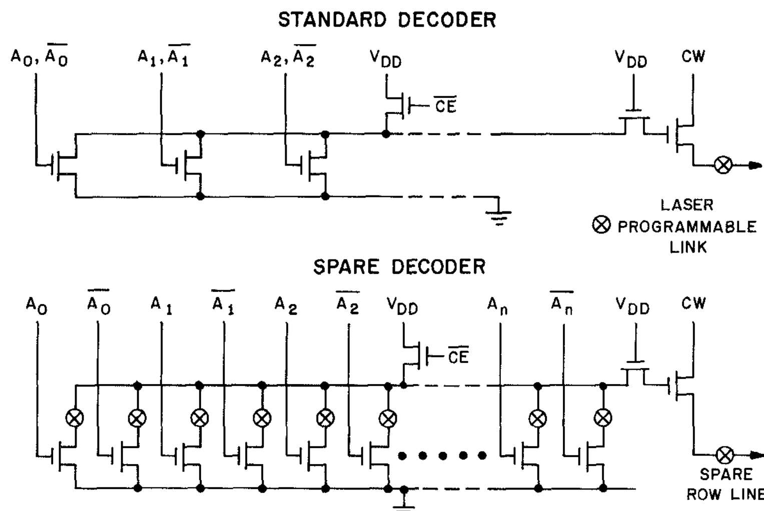


Fig. 2. Simplified schematic of standard and spare row decoders showing location of laser programmable links.

these transistors, so that the capacitive loading of an encoded spare is similar to that of a standard decoder. The electrical deselection of an encoded spare is virtually indistinguishable from that of a normal decoder, satisfying the second design constraint.

Replacement of a defective row requires the explosion of seven links, one to disconnect the faulty element and six more to encode the spare. Fig. 3 illustrates the decoder interchange phenomenon for a row. Replacement of a defective column is more complicated, requiring fourteen link explosions. Two links are removed to disconnect the faulty column from the associated I/O, $\overline{\text{I/O}}$ line pair, six more to encode the spare column decoder, and six more to disconnect the encoded spare column from three each of the four spare I/O and $\overline{\text{I/O}}$ line pairs. For more detailed circuit description and operation the reader is again referred to [2].

The third design constraint is satisfied by providing additional disconnect links in the spare circuitry so that a defective spare may itself be replaced by yet another spare.

III. Laser Programming System

A. *Hardware*

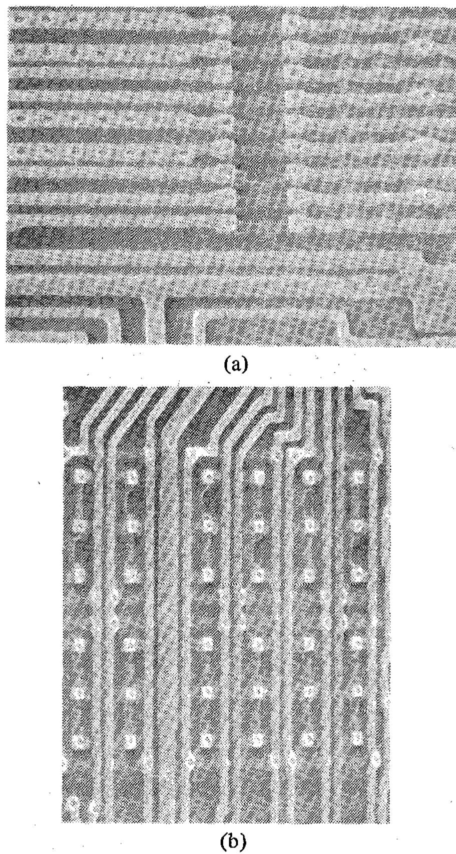
The basic hardware elements of an automatic laser programming system for fault-tolerant VLSI memories are a laser and

time, no such system is commercially available, though the basic building blocks can be purchased and integrated with relative ease. Even if a complete hardware system were available, much would be left to the user since the major development effort consists of user dependent software. This comment applies equally well to an alternative approach to memory fault tolerance based on electrically fusible links [3], [5]-[10]. The extent of this software can be gauged better from Section IV.

Fig. 4 is a block diagram showing the interconnection between the basic hardware components of a laser programming system. Most features needed for the laser programmer are available in commercial laser trimming systems. Two criteria are paramount for laser programming fault-tolerant VLSI memories, effective laser spot size and laser positioning accuracy. Although some tradeoff between these two parameters is possible, as discussed in Section III-B, a nominal spot size of about 7-8 μm and a beam positioning accuracy relative to the target of ± 1 μm is considered minimal.

This combination has proven entirely adequate for the Bell System 64K DRAM designed three years ago with 3.5-μm design rules. Looking to the future, however, the laser programming system should be capable of providing a choice of laser wavelengths, a range of effective spot sizes, and targeting accuracy will need to be improved. This is discussed further

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



(a)

(b)

Fig. 3. SEM photographs showing (a) disconnected row at 1000X and (b) an encoded and a nonencoded spare row decoder at 664X.

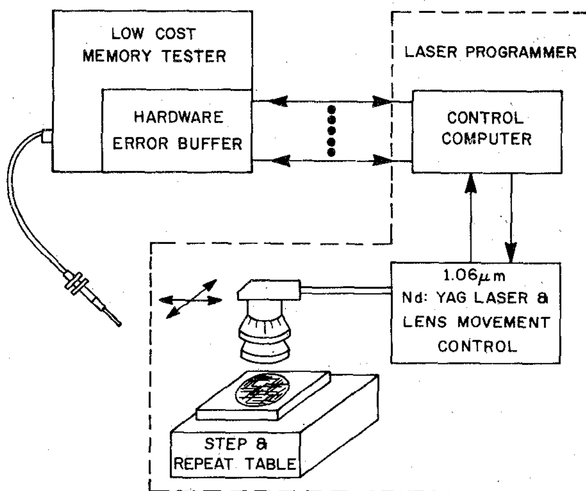


Fig. 4. Block diagram showing interconnection between major hardware components of laser programming system.

Q-switched Nd-doped YAG laser operated in fundamental mode at 1.064 μm wavelength. The final objective lens of the laser optics is also used as a viewing lens for a CCTV monitoring system. The lens is physically translated over the target wafer under control of a disk-based minicomputer, which also controls motion of a built-in step and repeat table. Physical translation of the laser optics presents a very useful advantage over galvanometric mirror scanning alternatives since the laser importance in minimizing classical lens aberrations, bearing in mind that the target features are of the order of 2-3 laser wavelengths. The CCTV camera has been specially interfaced to the control minicomputer to provide both automatic wafer and target die alignment.

The low-cost memory tester, which is linked to the control minicomputer by a bidirectional byte-wide databus, has several hardware features generally found only in larger general-purpose test systems. The principal such feature is a hardware error buffer, capable of capturing errors on the fly. The buffer may also be used to OR together the results of several device tests to allow detection and replacement of marginal defects. For example, it has proven highly worthwhile to perform multiple functional memory tests at low and high supply voltages and replace marginal defects regardless of which voltage extreme caused the device failure.

In order to provide extension of the fault-tolerant approach to other memory organizations, the buffer itself should be flexible in organization, and expandable. Byte-wide memories will, of course, require byte-wide drive and sensing capability. Flexibility of the buffer organization is important in one other respect, namely the addressing mode for reading the stored data. For example, if the redundancy scheme involves row replacements only, then rapid analysis of the error buffer is enhanced if the contents can be addressed and read out from the row direction in byte or multiple-byte data chunks. Similarly, if only column replacements need to be covered, then the buffer is better addressed and read in the column direction. Regardless of whether row only, column only, or both row and column redundancy schemes are chosen, the buffer organization should be optimized to match the diagnostic and spare allocation algorithm. Efficient analysis of the error buffer contents requires that they reflect a one-to-one physical mapping of the test memory. Topological address descrambling hardware simplifies both memory test program preparation and mapping to the error buffer.

B. Laser Target Explosion Phenomenon and Targeting Concerns

The laser output is spatially filtered by an aperture to produce a TEM_{00} beam. This beam is expanded and collimated via a Gallilean telescope beam expander, attenuated to suitable power level and focused to a small waist at the wafer surface. A single laser pulse of approximately 50-ns duration is used to sever a polysilicon target link. A short duration laser pulse is required to explode the target with no damage to adjacent and underlying structures. The energy per pulse is approximately 10 μJ. Clearly, laser targeting accuracy is a central concern in the programmable redundancy approach. The targeting stringency is defined by three factors: 1) effective laser spot diameter (ELSD); 2) target feature size; and 3) target feature nearest neighbor distance.

Fig. 5 illustrates the derivation of ELSD. A TEM_{00} mode, circularly symmetric, Gaussian power density distribution is assumed for the focused laser pulse at the wafer surface. A threshold power density may be defined such that explosion of a polysilicon link occurs if and only if the incident laser

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

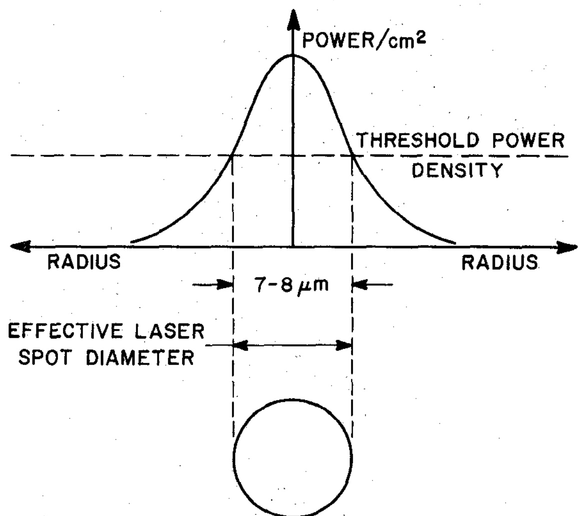


Fig. 5. Derivation of effective laser spot diameter (ELSD).

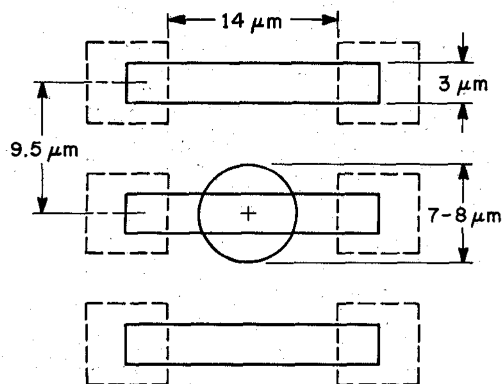


Fig. 6. Target link, nearest neighbor and laser spot geometries in scale representation.

the assumptions listed above, the region in which the incident laser power density exceeds the threshold level defines a circular area at the wafer surface in which any polysilicon will explode. An analysis of 64K DRAM polysilicon link explosion experience has resulted in an ELSD estimate of 7–8 μm as indicated in Fig. 5.

Fig. 6 is a schematic representation, to scale, of target link, nearest neighbor, and laser spot geometries. The target links are composed of heavily doped polysilicon, reside beneath a phosphorus-doped SiO_2 layer, and are 3 μm wide and 14 μm long. The most severe nearest neighbor target link distance is defined by the row line pitch where links are spaced 9.5 μm on center. The ELSD is included in Fig. 6 assuming perfect targeting. This illustration demonstrates both the targeting sensitivity involved (targeting errors on the order of microns causing possible failure in link disconnection) and the derivation of our design rule which states that the ELSD should just overfill the distance between polysilicon edges (6.5 μm).

The design rule described above benefits from a phenomenon termed the wicking effect, in which thermal energy appears to be preferentially drawn or wicked into the heavily doped polysilicon. For example, with the perfect targeting accuracy indicated in Fig. 6, the link in question would explode the

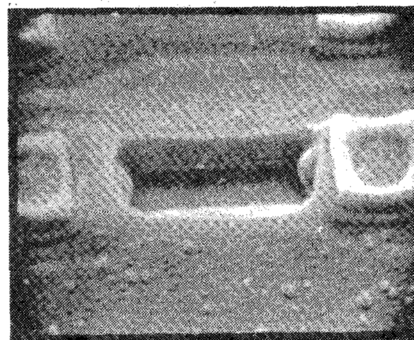


Fig. 7. SEM photograph of an exploded link with 5.5-μm ELSD at 5000×.

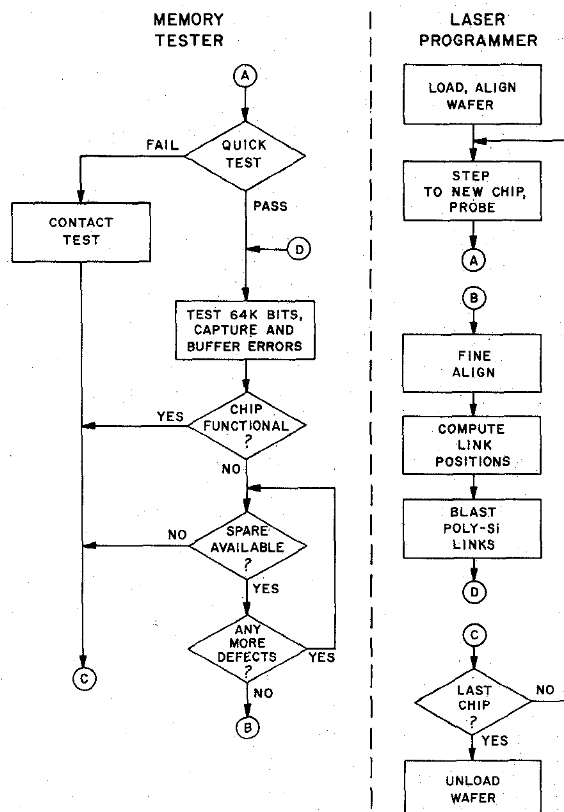


Fig. 8. Software system major flows.

entire 14-μm length even though only the middle 7–8 μm experienced an incident power density above threshold. We have observed targeting errors of several microns in the direction perpendicular to the link, in the situation of Fig. 6, with a clean severance of the link in question due to the wicking effect. If the ELSD is reduced to approximately 5.5 μm, the situation is as demonstrated in Fig. 7. Note that although the links are not blown header to header, nonetheless more of the link is gone than the 5.5 μm directly illuminated by the laser.

IV. System Software

A. Major Flows

The major logic flows of the laser programming process are indicated in Fig. 8. The operator loads a wafer on the step and repeat table chuck. The principal axes of the wafer are then

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

aligned with the system's axes. The bulk of the flowchart consists of a loop one circuit through which processes one chip. Each pass through the loop terminates in a test of the repeat criterion, i.e., was the last chip on the wafer just processed? If so, the wafer is unloaded. If not, another loop pass is initiated. Each loop pass begins by stepping to the next chip and positioning the test probes.

The first testing consists of a quick test designed to rapidly identify massive chip failures. If the device fails, a contact test is performed and the repeat criterion test is executed. If the chip passes quick test, a full test is then performed. If there are no errors, the chip is functional, and the repeat criterion test is executed. If there are errors, the program enters a loop in which one spare row or column is allocated for each error. This is called the "repair algorithm" below. There are two possible exits from the repair algorithm loop. 1) If there are more defects than spares, the chip is unrepairable and the repeat criterion test is executed. 2) If there are an equal or greater number of spares than defects, the chip is repairable and the program enters the target burn section.

The first task in the target burn section involves alignment of the system to the chip to a resolution better than the 2.5-μm step size of the step and repeat table. Then the coordinates of the links are calculated and the laser explodes these targets. The program reenters the full test section remembering the spares that were just used. This test-laser-burn loop is repeated until either the chip is functional or proves unrepairable.

B. Software Modules

The laser-programmable-redundancy software resides in the two principal components of the system, laser programmer and memory tester. The code for the laser programmer is written in Pascal and assembly language; that for the memory tester is written in assembly and pseudo-test language. The program which executes on our development system does so with an overlay structure due to the great many engineering options and the restrictive memory limitations imposed by the operating system. The production code runs nonoverlay for high throughput by means of sacrificing all but those features essential for the memory-repair proper.

A very desirable aspect of this software is its modularity. Since the system should be capable of adaptation to different memory devices and/or memory testers, it is important to modularize those sections of code which are common, and those which are unique to a specific device or memory tester. The software modules are four in number: 1) main program, 2) testing, 3) repair algorithm, and 4) laser and stepper movement and control.

1) Main Program: This module executes in the control minicomputer of the laser programmer. The main program selects from a menu of possible options or execution modes in its first section, utilizes these modes in its second section to accomplish the desired tasks, and leaves the bulk of execution detail to a large number of procedure and function subprograms. There are four types of options: a) parameter setup (e.g., number of laser pulses per burn, length of various programmed delays, wafer map information, target coordinates data files, etc.); b) calibration (e.g., laser targeting, stepper and automatic-alignment hardware calibration, etc.); c) selection of execution options (e.g., manual or automatic alignment, extended repair information report, etc.); and d) selection of mode (e.g., laser targeting, test only, test and burn for memory repair, etc.).

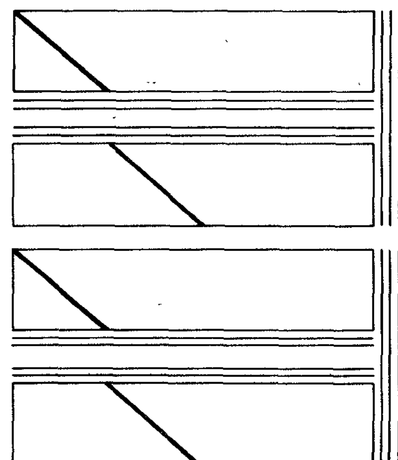


Fig. 9. Address path for diagonal quick test.

2) Testing: This module executes primarily in the memory tester. Testing software first performs a "quick diagonal" test on each memory device to rapidly reject massive, unrepairable failures. The diagonal lines, shown in Fig. 9, illustrate the bits tested in each half of the device. The test consists of a WRITE/READ function at each diagonal address followed by an increment of the row and column address. Each new error along this path would require a separate spare row or column. If the total for either half exceeds the number of spares available for that half (8 for the 64K DRAM), the memory is unrepairable. If the device fails this test, mechanical and electrical probe conditions are validated by performance of a contact test.

If diagonal quick test is passed, the tester executes a complete functional device test at two voltage extremes, logging any errors into the hardware buffer described in Section III-A.

3) Repair Algorithm: This module has, at one time or another, executed in either the memory tester or the laser control computer (Fig. 8 happens to illustrate the former), depending on available memory in either machine and tradeoffs between diagnostic time and data transmission overhead. Fig. 10 is a schematic representation of one half of a 64K DRAM with spare rows and columns and a repairable error pattern. For each half, the repair algorithm will assign spare elements one by one with the selection criterion being repair of maximum number of errors with each successive spare assignment. This process continues until there are no more errors (repairable half) or there are no more spares and some remaining errors (unrepairable half).

The repair algorithm would proceed with the situation indicated in Fig. 10 by first allocating a spare column to repair column 246 since it is entirely defective. As each spare is assigned, the errors in the buffer are deleted and the spare count is decremented. It deserves emphasis that there is no laser activity at this time since the decision of whether laser activity is required is being made. The repair algorithm will now replace row 11 since it has more errors than any other row or column.

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.