Real-Time Diagnosis of Semiconductor Manufacturing Equipment Using a Hybrid Neural Network Expert System

Byungwhan Kim, *Member, IEEE*, and Gary S. May, *Senior Member, IEEE*

***Abstract*— This paper presents a tool for the real-time diagnosis of integrated circuit fabrication equipment. The approach focuses on integrating neural networks into an expert system. The system employs evidential reasoning to identify malfunctions by combining evidence originating from equipment maintenance history, on-line sensor data, and in-line post-process measurements. Neural networks are used in the maintenance phase of diagnosis to approximate the functional form of the failure history distribution of each component. Predicted failure rates are then converted to belief levels. For on-line diagnosis in the case of previously unencountered faults, a CUSUM control chart is implemented on real sensor data to detect very small process shifts and their trends. For the known fault case, continuous hypothesis testing on the statistical mean and variance of the sensor data is performed to search for similar data patterns and assign belief levels. Finally, neural process models of process figures of merit (such as etch uniformity) derived from prior experimentation are used to analyze the in-line measurements, and identify the most suitable candidate among faulty input parameters (such as gas flow) to explain process shifts. A working prototype for this hybrid diagnostic system has been implemented on the Plasma Therm 700 series reactive ion etcher located in the Georgia Tech Microelectronics Research Center.**

***Index Terms*—Diagnosis, expert systems, neural networks, reactive ion etching.**

I. Introduction

AS THE semiconductor industry moves toward submicron fabrication technology, tight control of process variability is an essential requirement. A certain amount of variability is inherent in sophisticated semiconductor equipment, and significant performance shifts may occur when this variability becomes large compared to random process noise (i.e., fluctuations resulting from small and essentially uncontrollable causes). Such shifts are often indicative of equipment malfunctions. When unreliable equipment performance causes operating conditions to vary beyond an acceptable level, overall product quality is jeopardized. Thus, timely and accurate equipment malfunction diagnosis can be a key to the success of the semiconductor manufacturing process. Diagnosis involves identifying the assignable causes for the equipment malfunctions and correcting them quickly to prevent the subsequent occurrence of expensive misprocessing. With the advent of highly proficient sensors capable of monitoring process conditions *in-situ*, it is now desirable to perform diagnosis on a real-time basis.

Algorithmic diagnostic systems such as *HIPPOCRATES* [1] have been developed to identify process faults from statistical inference procedures and electrical measurements performed on finished IC wafers. Although this system makes good use of quantitative models of process behavior, it can only arrive at useful diagnostic conclusions in the limited regions of operation over which these models are valid. Furthermore, in critical process steps such as reactive ion etching (RIE), the theoretical basis for determining causal relationships is not well understood, thereby limiting the usefulness of physical models [2]. Expert systems such as *PIES* [3] have been designed to draw upon experiential knowledge to develop qualitative models of process behavior. This approach has attained limited success in attempting to diagnose unstructured problems which lack a solid conceptual foundation for reasoning. However, a purely knowledge-based technique often lacks the precision inherent in deep-level physical models, and is thus incapable of deriving solutions for unanticipated situations from the underlying principles surrounding the process.

Neural networks have recently emerged as an effective tool for process modeling [4], [5] as well as fault diagnosis [6], [7]. Diagnostic problem solving using neural networks requires the association of input patterns representing quantitative and qualitative process behavior to fault identification. Robustness to noisy sensor data and high speed parallel computation make neural networks an attractive alternative for real-time diagnosis. However, the pattern recognition-based neural network approach has limitations. First, a complete set of fault signatures is hard to obtain, and the representational inadequacy of a limited number of data sets can induce network overtraining, thus increasing the misclassification or "false alarm" rate. Also, pattern matching approaches in which diagnostic actions take place following a sequence of several processing steps are sub-optimal since evidence pertaining to potential equipment malfunctions accumulates at irregular intervals throughout the process sequence. At the end of a sequence, significant misprocessing and yield loss may have already taken place, making post-process diagnosis alone economically undesirable.

Manuscript received January 31, 1996; revised March 1997. This work was supported by the National Science Foundation Grant DDM-9 358163 and the IEEE/CPMT Motorola Fellowship.

B. Kim is with the Memory R&D Division, Department of Equipment Engineering, Hyundai Electronics Industries Co., Ltd., Korea.

G. S. May is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250 USA.

Publisher Item Identifier S 1083-4400(97)04320-9.

This paper presents a prototype tool for the automated malfunction diagnosis of integrated circuit fabrication equipment. The methodology described combines the best characteristics of quantitative algorithmic, qualitative experiential and pattern recognition-based neural network approaches. This system offers advantages in that it yields a stable and reliable ranked list of fault possibilities, even in the presence of measurement noise (in part due to the inherent noise resistance of neural networks). In addition, the varying degrees of belief in each stage of diagnosis aids in the early detection of suspicious trends, often prior to an actual failure occurrences. This working prototype is currently being developed and implemented on a Plasma Therm 700 series RIE located in the Georgia Tech Microelectronics Research Center.

II. Diagnostic Inference Method

As a diagnostic inference method, the Dempster–Shafer theory of evidential reasoning [8] has proven to be suitable for real-time malfunction diagnosis applications [9]. This technique allows the combination of various pieces of uncertain evidence obtained at irregular intervals, and its implementation results in time-varying, nonmonotonic belief functions which reflect the current status of diagnostic conclusions at any given point in time.

One of the basic concepts in Dempster–Shafer theory is the *frame of discernment* (symbolized by θ), defined as an exhaustive set of mutually exclusive propositions. In diagnosis, the frame of discernment is the union of all possible fault hypotheses. Each piece of collected evidence can be mapped to a fault or group of faults within θ. The likelihood of a fault proposition A is expressed as a bounded interval $[s(A), p(A)]$ which lies in [0, 1]. The parameter $s(A)$ represents the *support* for A, which measures the weight of evidence in support of A. The other parameter, $p(A)$, called the *plausibility* of A, is defined as the degree to which contradictory evidence is lacking. Plausibility measures the maximum amount of belief that can possibly be assigned to A. The quantity $u(A)$ is the *uncertainty* of A, which is the difference between the evidential plausibility and support. For example, an evidence interval of [0.3, 0.7] for proposition A indicates that the probability of A is between 0.3 and 0.7, with an uncertainty of 0.4.

For diagnosis, proposition A represents a given fault hypothesis. An evidence interval for fault A is determined from a basic probability mass distribution (BPMD). The BPM $m\langle A\rangle$ indicates the portion of the total belief in evidence assigned to a particular fault hypothesis set. Any residual belief in the frame of discernment that cannot be attributed to any subset of θ is assigned directly to θ itself, which in effect introduces uncertainty into the diagnosis. Using this framework, the support and plausibility of proposition A are given by

$$s(A) = \Sigma m\langle A_i\rangle \tag{1}$$

$$p(A) = 1 - \Sigma m\langle B_i\rangle \tag{2}$$

where $A_i \subseteq A$ and $B_i \subseteq \bar{A}$ and the summation is taken over all propositions in a given BPM. Thus the total belief in A is

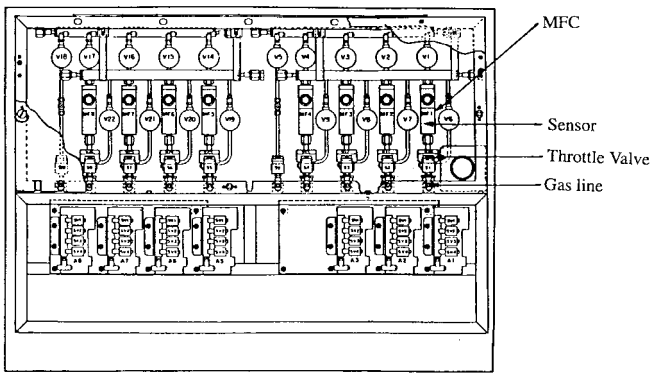

Fig. 1. Partial schematic of RIE gas delivery system.

Dempster's rules for evidence combination provide a deterministic and unambiguous method of combining BPMD's from separate and distinct sources of evidence contributing varying degrees of belief to several propositions under a common frame of discernment. The rule for combing the observed BPM's of two arbitrary and independent knowledge sources m_1 and m_2 into a third m_3 is as follows:

$$m_3\langle Z\rangle = \frac{\Sigma m_1\langle X_i\rangle * m_2\langle Y_j\rangle}{1-k} \tag{3}$$

where $Z = X_i \cap Y_j$ and

$$k = \Sigma m_1\langle X_i\rangle * m_2\langle Y_j\rangle \tag{4}$$

where $X_i \cap Y_j = \emptyset$. Here X_i and Y_j represent various propositions which consist of fault hypotheses and disjunctions thereof. Thus, the BPM of the intersection of X_i and Y_j is the product of the individual BPM's of X_i and Y_j. The factor $(1-k)$ is a normalization constant which prevents the total belief from exceeding unity due to attributing portions of belief to the empty set.

Consider the combination of m_1 and m_2 when each contains different evidence concerning the diagnosis of a malfunction in the RIE application. Such evidence could result from two different sensor readings for example. In particular, suppose that the sensors have observed that the flow of one of the etch gases into the process chamber is too low. Let the frame of discernment $\theta = \{A, B, C, D\}$, where A through D symbolically represent the following mutually exclusive equipment faults:

- A mass flow controller miscalibration;
- B gas line leak;
- C throttle valve malfunction;
- D incorrect sensor signal.

These components are illustrated graphically in the partial schematic of the etcher gas flow system shown in Fig. 1.

Suppose that belief in this frame of discernment is distributed according to the BPMD's:

$$m_1\langle A\cup C, B\cup D, \theta\rangle = \langle 0.4, 0.3, 0.3\rangle$$

$$m_2\langle A\cup B, C, D, \theta\rangle = \langle 0.5, 0.1, 0.2, 0.2\rangle.$$

The calculation of the combined BPMD (m_3) is shown in

TABLE I
ILLUSTRATION OF BPMD COMBINATION

m_1		m_2							
$A \cup C$	0.4	A	0.20	C	0.04	$\varnothing$	0.08	$A \cup C$	0.08
$B \cup D$	0.3	B	0.15	$\varnothing$	0.03	D	0.06	$B \cup D$	0.06
θ	0.3	$A \cup B$	0.15	C	0.03	D	0.06	θ	0.06
		$A \cup B$	0.50	C	0.1	D	0.2	θ	0.2

corresponding propositions from m_1 and m_2, along with the product of their individual beliefs. Note that the intersection of any proposition with θ is the original proposition. The BPM attributed to the empty set, k, which originates from the presence of various propositions in m_1 and m_2 whose intersection is empty, is 0.11. By applying (3), BPM's for the remaining propositions result in

$$m_3\langle A, A\cup C, A\cup B, B, B\cup D, C, D, \theta\rangle = \langle 0.225, 0.089, 0.169, 0.169, 0.067, 0.079, 0.135, 0.067\rangle.$$

The plausibilities for propositions in the combined BPM are calculated by applying (2). The individual evidential intervals implied by m_3 are $A[0.225, 0.550]$, $B[0.169, 0.472]$, $C[0.079, 0.235]$, and $D[0.135, 0.269]$. Combining the evidence available from knowledge sources m_1 and m_2 thus leads to the conclusion that the most likely cause of the insufficient gas flow malfunction is a miscalibration of the mass flow controller (proposition A).

III. NEURAL NETWORK-BASED RIE MODELING

Neural networks have the capability of learning complex relationships between groups of related parameters. They consist of parallel processing units (called *neurons*), which are interconnected in such a way that knowledge is stored in the weight of the connections between them. Each neuron contains the weighted sum of its inputs filtered by a sigmoidal activation function. The nonlinear mapping capabilities of neural networks have recently been applied by several other researchers in semiconductor process modeling [10]–[13]. To model the RIE process, the quantitative relationships which relate input parameters to output responses have been encoded in feed-forward neural networks via the error back-propagation (BP) algorithm [14]. The structure of a typical BP network appears in Fig. 2. The specific manner in which BP neural nets are used in RIE diagnosis is described below.

A. Time Series Modeling

For real-time diagnosis, it is critical to model the variation of in-situ sensor data and develop an efficient method for handling this voluminous and multidimensional data. Time-series modeling is a means to achieve each of these ends. Under malfunction conditions, sensor readings can serve as process "signatures" which assist in identifying the occurrence of a particular fault. Recently, neural networks have been proposed as a means to develop time series models of tool

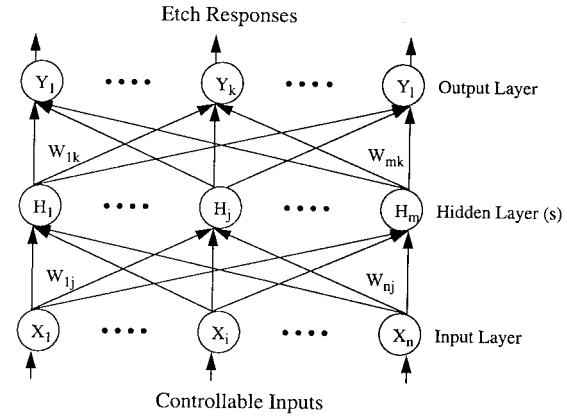

Fig. 2. Typical back-propagation neural network.

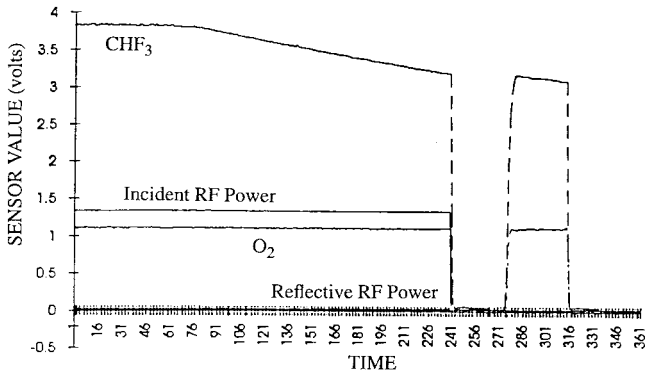

Fig. 3. Data signatures for a malfunctioning CHF_3 mass flow controller.

Neural networks used to generalize the behavior of a time series are referred to as *neural time series* (NTS) models. The NTS model is capable of simultaneously filtering both auto- and cross-correlated data. That is, the NTS model can account for correlation among several variables being monitored simultaneously. To illustrate, real-time tool data was collected via an equipment monitoring system designed to transfer data from an etcher to a remote workstation. Monitoring was accomplished using a Tektronix Model 2510 *TestLab* data acquisition system interfaced to the Plasma Therm RIE system via serial ports. In this example, an equipment alarm was signaled, and its cause was later identified to be an insufficient gas supply from the tri-fluoro methane mass flow controller (CHF_3). Fig. 3 depicts malfunctioning behavior of the CHF_3 gas flow.

An NTS network was trained to model the CHF_3 flow pattern in the RIE process using a simple sampling technique which involved training the network to forecast the next CHF_3 from the behavior of five past values. The training set for the NTS network consisted of one out of every ten data samples. As shown in Fig. 4, auto-correlation among consecutive CHF_3 measurements was accounted for by simultaneously training the network on the present value of CHF_3 and five past values. The cross-correlation among the CHF_3 was modeled by including as inputs to the network the present values of the temperature, incident and reflected RF power, oxygen and CHF_3. The accuracy of the trained network was measured by

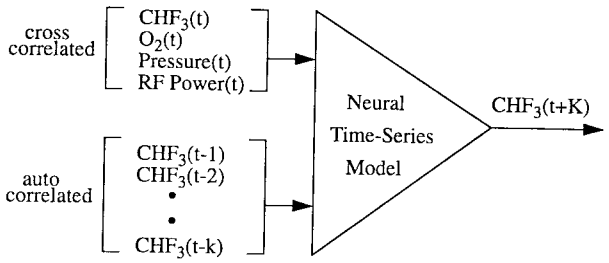

Fig. 4. Inputs (auto and cross-correlated data) and output of a neural time-series model.

trained, the NTS model provides a simple means to encode this fault signature for later use.

B. Process Modeling

Diagnostic information can also be extracted from in-line measurements of post-processed wafers. To achieve this, these measurements must be compared to values predicted by a process model. Differences between model predictions and measured responses are indicative of potential equipment malfunctions. In [5], neural network models of RIE responses were developed from a Box–Wilson *central composite circumscribed* design requiring 27 trials [16]. Etching was performed on a test structure designed to facilitate the simultaneous measurement of the etch rate, uniformity, anisotropy of SiO_2 in a CHF_3 and oxygen plasma, as well as the selectivity of the SiO_2 etch with respect to photoresist. This characterization experiment provided neural network training data.

A "forward" neural network-based process model defines a functional relationship between RIE process conditions (inputs) such as RF power or gas composition and responses (outputs) such as etch rate or uniformity. The forward process model also provides a mechanism for comparing measured RIE output responses to predicted values. Large differences, which may indicate potential equipment faults, must then be traced back to fluctuations in model input parameters. To calculate the shift of the process input settings from their nominal values, an inverse neural process model is employed. This inverse model is obtained by training the network "in reverse" (i.e., using output/input pairs, rather than input/output pairs). The inverse model provides a means to identify the input parameter which is most likely responsible for an output process shift. Process shifts required for generating evidential support and plausibility can then be computed by utilizing the inverse neural process models.

IV. Generation of Evidential Support

The three relevant time periods for evidence collection in semiconductor manufacturing are:

1) during equipment maintenance periods (before processing);
2) during on-line equipment operation (during processing);
3) during in-line, post-process physical and/or electrical

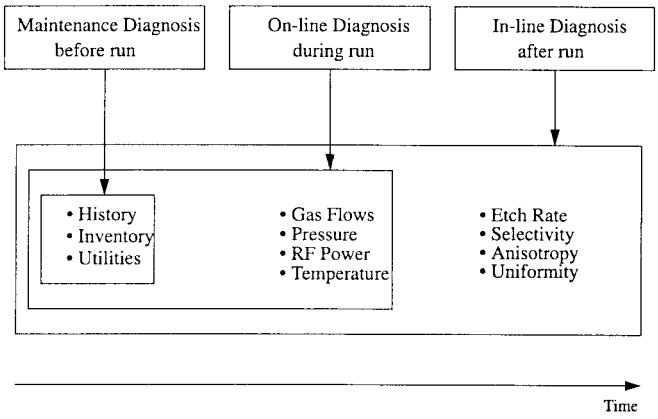

Fig. 5. Chronological evidence sources for equipment malfunction diagnosis.

Diagnosis based on this framework for evidence collection takes place in three chronological stages (Fig. 5). Maintenance diagnosis is performed by examining the relevant historical records of equipment performance and building reliability models of each equipment component. During on-line diagnosis, both neural time-series models and CUSUM control chart [17] techniques are employed to analyze fault patterns available from equipment monitoring system. For in-line diagnosis, measurements on processed wafers are used in conjunction with neural network process models. In each phase, evidential support and plausibility for various fault hypotheses are generated and mapped to particular equipment components. The methodology employed to do so is discussed below.

A. Maintenance Diagnosis

During the maintenance phase, the objective of the diagnostic system is to derive evidence of potential component failures based on the historical performance of equipment components. The data available from which evidential belief may be generated is limited, consisting of only the number of failures a given component has experienced and the component age. In order to derive evidential support for potential malfunctions from this information, a reliability modeling technique has been developed to investigate the aging behavior of components. The failure probability as a function of time and the instantaneous failure rate (or "hazard" rate) for each component may be estimated from a neural network trained on failure history. The neural reliability model may then be used to generate evidential support, plausibility and uncertainty for each fault hypotheses (i.e., each potentially faulty component) in the frame of discernment.

1) The Weibull Distribution: Consider reliability modeling based on the Weibull distribution. The Weibull distribution has been used extensively as a model of time to failure in electrical and mechanical components and systems. Examples of systems which lend themselves to the Weibull model include electrical components such as batteries and ceramic multilayer capacitors, mechanical systems such as gas turbine engines

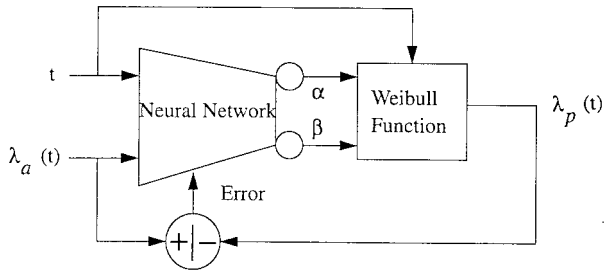

Fig. 6. Scheme to estimate Weibull function parameters.

mechanical parts such as bearings, and structural elements in aircraft and automobiles [18]. When a system is composed of a number of components and failure is due to the most serious of a large number of possible faults, the Weibull distribution seems to be a particularly accurate model [17], and this closely resembles the situation being addressed in semiconductor equipment malfunction diagnosis.

The cumulative distribution function (which represents the failure probability of a component at time t) for the two-parameter Weibull distribution is given by

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \tag{5}$$

where α and β are called scale and shape parameters, respectively. If a device exhibits Weibull-like reliability behavior, the appropriate selection of α and β will allow this distribution functions to closely approximate the observed failure behavior throughout its lifetime. The Weibull hazard rate, $\lambda(t)$, is given by

$$\lambda(t) = \frac{\beta t^{\beta-1}}{\alpha^{\beta}}. \tag{6}$$

The hazard rate may be computed from the failure history of each component by plotting the number of failures versus time and finding the slope of this curve at each time point.

A scheme designed to extract the shape and scale parameters using neural networks has been developed and tested, and is outlined in [18]. This scheme is depicted schematically in Fig. 6. Here the network outputs represent the initially unknown scale and shape parameters. These outputs are iteratively adjusted to reach to their optimal values as the neural network learns. The outputs are fed into the failure hazard function in (6), a predicted hazard rate (λ_p) is computed, and the result is compared with the actual hazard rate (λ_a), which has been computed from the failure history data.

The standard back-propagation training algorithm for feed-forward neural networks begins with a random set of weights. An input vector which has been normalized so that all input data lies in the interval between -1 and 1 is then presented to the network, and the output is calculated using this initial weight matrix. Next, the calculated output vector is compared to the measured output vector, and the squared difference between the two is used to determine the system error. Error minimization is accomplished via the gradient descent approach, in which the weights are adjusted in the direction

The error signal for the modified back-propagation neural network in this case is

$$E = 0.5(\lambda_p - \lambda_a)^2. \tag{7}$$

Since the predicted hazard rate is differentiable, the error gradient with respect to the network weights may be computed using the chain rule as

$$\frac{\partial E}{\partial w_{ijk}} = \left(\frac{\partial E}{\partial \lambda_p}\right)\left(\frac{\partial \lambda_p}{\partial \mathrm{out}_{ik}}\right)\left(\frac{\partial \mathrm{out}_{ik}}{\partial w_{ijk}}\right) \tag{8}$$

where out_{ik} is the calculated output of the ith neuron in the kth layer. The first partial derivative in (8) is $(\lambda_p - \lambda_a)$, and the third is the same as in the standard implementation of the back-propagation algorithm [14]. As for the second factor, this partial derivative may be computed separately for each individual output neuron (or equivalently, for each unknown parameter to be estimated). Due to the initially random network weights, the first predicted values of the hazard rate are arbitrary. However, after several training iterations, the predicted hazard rate converges to the actual rate. At this point, the scale and shape parameters computed at the network output are the estimates which best fit the distribution indicated by the training data.

Following parameter estimation using this technique, the evidential support for each equipment component is then obtained from the Weibull distribution function in (5) with the estimated parameters. The corresponding plausibility is the *confidence level* (C) associated with this probability estimate, which is defined as [19]

$$C(t) = 1 - [1 - F(t)]^n \tag{9}$$

where n denotes the total number of component failures which have been observed at time t.

2) The Exponential Distribution: Although the Weibull distribution provides one approach to component reliability modeling, due to its simple functional form, the exponential distribution is widely used to describe the time elapsing between two failures by characterizing the period during which a failure rate is constant [17]. The cumulative distribution function of this distribution is

$$F(t) = 1 - e^{-\lambda t} \tag{10}$$

where λ is a constant equal to the reciprocal of the mean-time-to-failure. This parameter may be estimated as

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} \tag{11}$$

where t_i represents the elapsed time between ith and $(i-1)$th failure of a specific component. The evidential support is obtained by inserting (11) into (10) and subsequently computing the corresponding Dempster–Shafer plausibility by

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.