US007660248B1

(12) **United States Patent**
Duffield et al.

(10) **Patent No.: US 7,660,248 B1**
(45) **Date of Patent: Feb. 9, 2010**

(54) **STATISTICAL, SIGNATURE-BASED APPROACH TO IP TRAFFIC CLASSIFICATION**

(76) Inventors: **Nicholas G. Duffield**, 101 W. 12th St., Apt. 7S, New York, NY (US) 10011; **Matthew Roughan**, 15 Locust St., Morristown, NJ (US) 07960; **Subhabrata Sen**, 420 River Rd., Apt H6, Chatham, NJ (US) 07928; **Oliver Spatscheck**, 15 Lawrence Rd., Randolph, NJ (US) 07896

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 776 days.

(21) Appl. No.: **10/764,001**

(22) Filed: **Jan. 23, 2004**

(51) **Int. Cl.**
H04L 12/26 (2006.01)

(52) **U.S. Cl.** **370/230.1**; 370/229; 370/232; 370/235; 370/238; 370/252

(58) **Field of Classification Search** 370/229, 370/230, 230.1, 233, 234, 235, 235.1, 237, 370/238, 241, 242, 244, 245, 250, 252, 253, 370/231, 232
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

7,251,218 B2 * 7/2007 Jorgensen 370/235
7,302,682 B2 * 11/2007 Turkoglu 717/174
7,305,676 B1 * 12/2007 Boll et al. 718/107
7,359,320 B2 * 4/2008 Klaghofer et al. 370/230
7,433,943 B1 * 10/2008 Ford 709/223
7,441,267 B1 * 10/2008 Elliott 726/13

OTHER PUBLICATIONS

Dewes, et al, An Analysis of Internet Chat Traffic, Oct. 2003, Proceedings of ACM SIGCOMM Internet Measurement Conference.

* cited by examiner

Primary Examiner—Pankaj Kumar
Assistant Examiner—Mark Mais
(74) *Attorney, Agent, or Firm*—Henry Brendzel

(57) **ABSTRACT**

A signature-based traffic classification method maps traffic into preselected classes of service (CoS). By analyzing a known corpus of data that clearly belongs to identified ones of the preselected classes of service, in a training session the method develops statistics about a chosen set of traffic features. In an analysis session, relative to traffic of the network where QoS treatments are desired (target network), the method obtains statistical information relative to the same chosen set of features for values of one or more predetermined traffic attributes that are associated with connections that are analyzed in the analysis session, yielding a statistical features signature of each of the values of the one or more attributes. A classification process then establishes a mapping between values of the one or more predetermined traffic attributes and the preselected classes of service, leading to the establishment of QoS treatment rules.

1 Claim, 1 Drawing Sheet

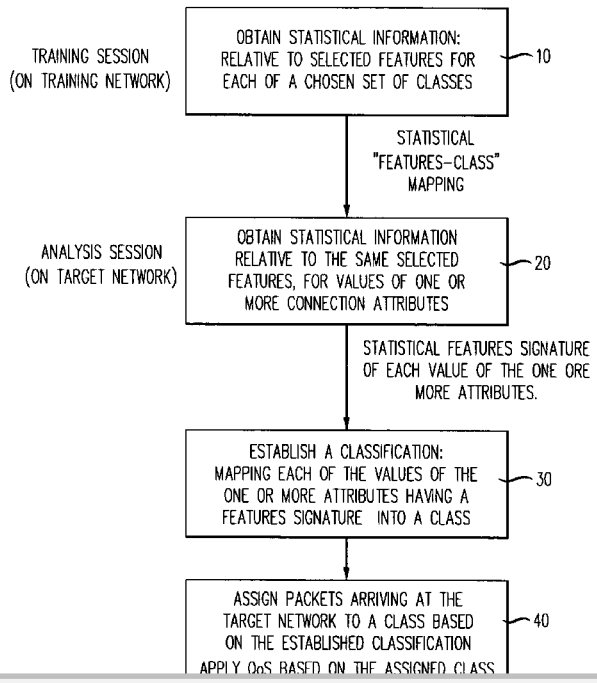
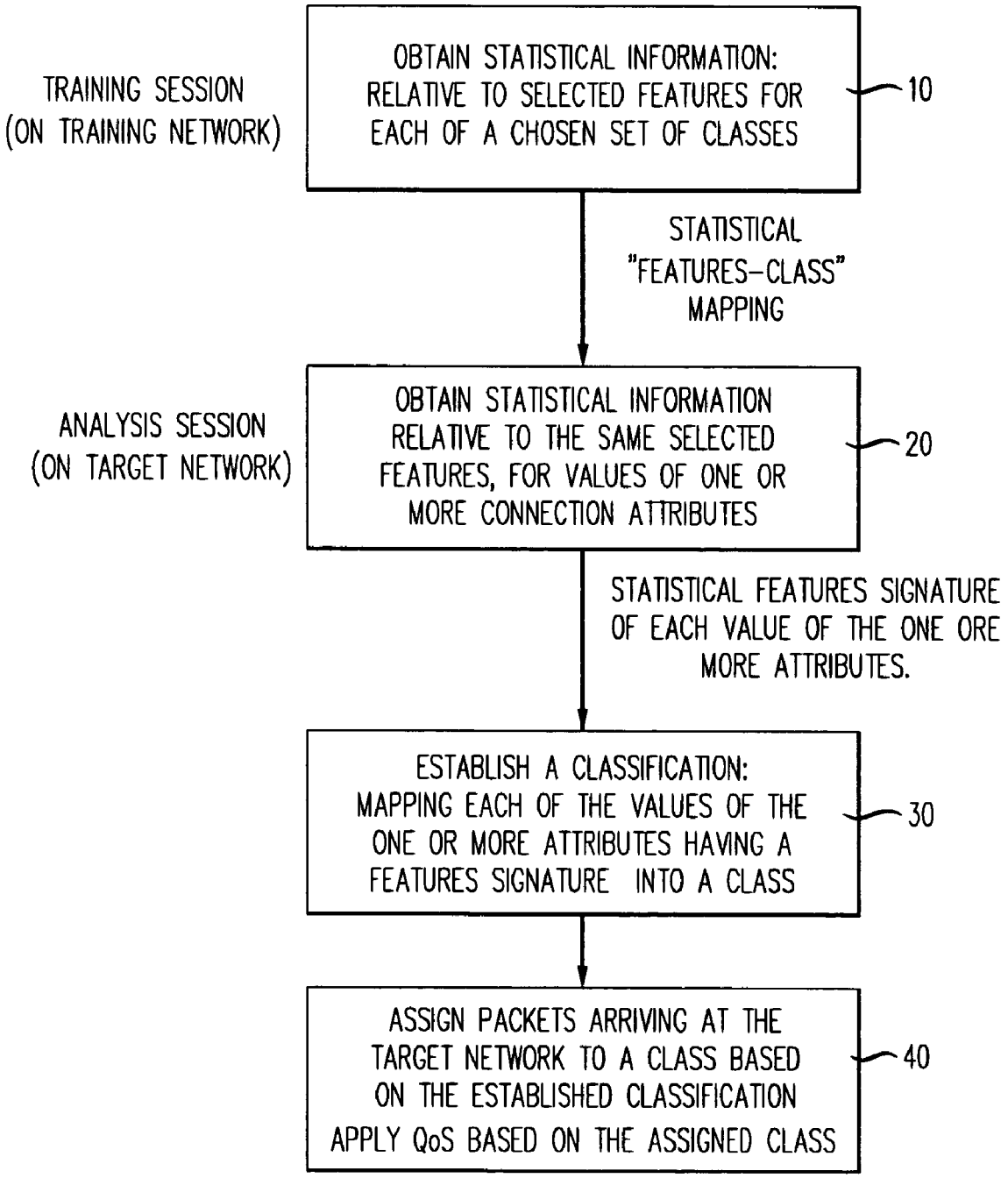


FIG. 1



STATISTICAL, SIGNATURE-BASED APPROACH TO IP TRAFFIC CLASSIFICATION

BACKGROUND OF THE INVENTION

This invention relates to traffic classification and, more particularly to statistical classification of IP traffic.

The past few years have witnessed a dramatic increase in the number and variety of applications running over the Internet and over enterprise IP networks. The spectrum includes interactive (e.g., telnet, instant messaging, games, etc.), bulk data transfer (e.g., ftp, P2P file downloads), corporate; (e.g., Lotus Notes, database transactions), and real-time applications (voice, video streaming, etc.), to name just a few.

Network operators, particularly in enterprise networks, desire the ability to support different levels of Quality of Service (QoS) for different types of applications. This desire is driven by (i) the inherently different QoS requirements of different types of applications, e.g., low end-end delay for interactive applications, high throughput for file transfer applications etc.; (ii) the different relative importance of different applications to the enterprise—e.g., Oracle database transactions are considered critical and therefore high priority, while traffic associated with browsing external web sites is generally less important; and (iii) the desire to optimize the usage of their existing network infrastructures under finite capacity and cost constraints, while ensuring good performance for important applications.

Various approaches have been studied, and mechanisms developed for providing different QoS in a network. See, for example, S. Blake, et al., RFC 2475—an architecture for differentiated service, December 1998, http://ww.faqs.org/rfcs/rfc2475.html; and C. Gbaguidi, et al., A survey of differentiated services architectures for the Internet, March 1998, http://sscwww.epfl.ch/Pages/publications/ps_files/tr98__020.ps; and Y. Bernet, et al., A framework for differentiated services. Internet Draft (draft-ietf-diffserv-framework-02.txt), February 1999, http://search.ietf.org/internet-drafts/draft-ietf-diffserv-framework-02.txt.

Previous work also has examined the variation of flow characteristics according to applications. M. Allman, et al., TCP congestion control, IETF Network Working Group RFC 2581, 1999, investigated the joint distribution of flow duration and number of packets, and its variation with flow parameters such as inter-packet timeout. Differences were observed between the distributions of some application protocols, although overlap was clearly also present between some applications. Most notably, the distribution of DNS transactions had almost no overlap with that of other applications considered. However, the use of such distributions as a discriminator between different application types was not considered.

There also exists a wealth of research on characterizing and modeling workloads for particular applications, with A. Krishnamurth, et al., *Web Protocols and Practice*, Chapter 10, Web Workload Characterization, Addison-Wesley, 2001; and J. E. Pitkow, Summary of WWW characterizations, *WWJ*, 2:3-13, 1999 being but two examples of such research.

An early work in this space, reported in V. Paxson, "Empirically derived analytic models of wide-area TCP connections," *IEEE/ACM Transactions on Networking*, vol. 2, no. 4, pp. 316-336, 1994, examines the distributions of flow bytes and packets for a number of different applications.

Interflow and intraflow statistics are another possible

"Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226-244, June 1995, for example, found that user initiated events—such as telnet packets within flows or FTP-data connection arrivals—can be described well by a Poisson process, whereas other connection arrivals deviate considerably from Poisson.

Signature-based detection techniques have also been explored in the context of network security, attack and anomaly detection; e.g. P. Barford et al., Characteristics of Network Traffic Flow Anomalies, *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, October 2001; and P. Barford, et al., A Signal Analysis of Network Traffic Anomalies, *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, November 2002, where one typically seeks to find a signature for an attack.

Actually, realization of a service differentiation capability requires (i) association of the traffic with the different applications, (ii) determination of the QoS to be provided to each, and finally, (iii) mechanisms in the underlying network for providing the QoS; i.e., for controlling the traffic to achieve a particular quality of service.

While some of the above-mentioned studies assume that one can identify the application traffic unambiguously and then obtain statistics for that application, none of them have considered the dual problem of inferring the application from the traffic statistics. This type of approach has been suggested in very limited contexts such as identifying chat traffic in C. Dewes, et al., An analysis of Internet chat systems, *Proceedings of ACM SIGCOMM Internet Measurement Conference*, October 2003.

Still, in spite of a clear perceived need, and the prior art work reported above, widespread adoption of QoS control of traffic has not come to pass. It is believed that the primary reason for the slow spread of QoS-use is the absence of suitable mapping techniques that can aid operators in classifying the network traffic mix among the different QoS classes. We refer to this as the Class of Service (CoS) mapping problem, and perceive that solving this would go a long way in making the use of QoS more accessible to operators.

SUMMARY

An advance in the art of providing specified QoS in an IP network is achieved with a signature-based traffic classification method that maps traffic into preselected classes of service (CoS). By analyzing, in a training session, a known corpus of data that clearly belongs to identified ones of the preselected classes of service, the method develops statistics about a chosen set of traffic features. In an analysis session, relative to traffic of the network where QoS treatments are desired (target network), obtaining statistical information relative to the same chosen set of features for values of one or more predetermined traffic attributes that are associated with connections that are analyzed in the analysis session, yielding a statistical features signature of each of the values of the one or more attributes. A classification process then establishes a mapping between values of the one or more predetermined traffic attributes and the preselected classes of service, leading to the establishment of rules. Once the rules are established, traffic that is associated with particular values of the predetermined traffic attributes are mapped to classes of service, which leads to a designation of QoS.

Illustratively, the preselected classes of service may be interactive traffic, bulk data transfer traffic, streaming traffic and transactional traffic. The chosen set of traffic features

features. The predetermined traffic attributes may be the server port, and the server IP address. An illustrative rule might state that "a connection that specifies port x belongs to the class of interactive traffic." An administrator of the target network may choose to give the highest QoS level to such traffic.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** presents a flow chart of the IP traffic classification method disclosed herein.

DETAILED DESCRIPTION

In accord with the principles disclosed herein QoS implementations are based on mapping of traffic into classes of service. In principle the division of traffic into CoS could be done by end-points of the network, where traffic actually originates—for instance by end-user applications. However, for reasons of trust and scalability of administration and management, it is typically more practical to perform the CoS mapping within the network; for instance, at the router that connects the Local Area Network (LAN) to the Wide Area Network (WAN). Alternatively, there might be appliances connected near the LAN to WAN transition point that can perform packet marking for QoS.

CoS mapping inside the network is a non-trivial task. Ideally, a network system administrator would possess precise information on the applications running inside the administrator's network, along with simple and unambiguous mappings, which information is based on easily obtained traffic measurements (e.g., by port numbers, or source and destination IP addresses). This information is vital not just for the implementation of CoS, but also in planning the capacity required for each class, and balancing tradeoffs between cost and performance that might occur in choosing class allocations. For instance, one might have an application whose inclusion in a higher priority class is desirable but not cost effective (based on traffic volumes and pricing), and so some difficult choices must be made. Good data is required for these to be informed choices.

In general, however, the required information is rarely up-to-date, or complete, if it is available at all. The traditional ad-hoc growth of IP networks, the continuing rapid proliferation of new applications, the merger of companies with different networks, and the relative ease with which almost any user can add a new application to the traffic mix with no centralized registration are all factors that contribute to this "knowledge gap". Furthermore, over recent years it has become harder to identify network applications within IP traffic. Traditional techniques such as port-based classification of applications, for example, have become much less accurate.

One approach that is commonly used for identifying applications on an IP network is to associate the observed traffic (using flow level data, or a packet sniffer) with an application based on TCP or UDP port numbers. Alas, this method is inadequate.

The TCP/UDP port numbers are divided into three ranges: the Well Known Ports (0-1023), the Registered Ports (1024-49,151), and the Dynamic and/or Private ports (49,152-65,535). A typical TCP connection starts with a SYN/SYN-ACK/ACK handshake from a client to a server. The client addresses its initial SYN packet to the well-known server port of a particular application. The client typically chooses the

All future packets of a session, in either a TCP or UDP session, use the same pair of ports to identify the client and server side of the session. Therefore, in principle, the TCP or UDP server port number can be used to identify the higher layer application by simply identifying in an incoming packet the server port and mapping this port to an application using the IANA (Internet Assigned Numbers Authority) list of registered ports (http://www.iana.org/assignments/port-numbers). However, port-based application classification has limitations. First, the mapping from ports to applications is not always well defined. For instance.

- Many implementations of TCP use client ports in the registered port range. This might mistakenly classify the connection as belonging to the application associated with this port. Similarly, some applications (e.g., old bind versions), use port numbers from the well-known ports to identify the client site of a session.
- Ports are not defined with IANA for all applications, e.g., P2P applications such as Napster and Kazaa.
- An application may use ports other than its well-known ports to circumvent operating system access control restrictions. E.g., non-privileged users often run WWW servers on ports other than port 80, which is restricted to privileged users on most operating systems.
- There are some ambiguities in the port registrations, e.g., port 888 is used for CDDBP (CD Database Protocol) and access-builder.
- In some cases server ports are dynamically allocated as needed. For example, FTP allows the dynamic negotiation of the server port used for the data transfer. This server port is negotiated on an initial TCP connection, which is established using the well-known FTP control port.

The use of traffic control techniques like firewalls to block unauthorized, and/or unknown applications from using a network has spawned many work-arounds which make port based application authentication harder. For example, port 80 is being used by a variety of non-web applications to circumvent firewalls which do not filter port-80 traffic. In fact, available implementations of IP over HTTP allow the tunneling of all applications through TCP port 80.

- Trojans and other security attacks generate a large volume of bogus traffic which should not be associated with the applications of the port numbers those attacks use.

A second limitation of port-number based classification is that a port can be used by a single application to transmit traffic with different QoS requirements. For example, (i) Lotus Notes transmits both email and database transaction traffic over the same ports, (ii) sep (secure copy), a file transfer protocol, runs over ssh (secure shell), an interactive application using default TCP port 22. This use of the same port for traffic requiring different QoS requirements is quite legitimate, and yet a good classification must separate different use cases for the same application. A clean QoS implementation is still possible through augmenting the classification rules to include IP address-based disambiguation. Server lists exist in some networks but, again, in practice these lists are often incomplete, or a single server could be used to support a variety of different types of traffic, so we must combine port and IP address rules.

A possible alternative to port based classification is to use a painstaking process involving installation of packet sniffers and parsing packets for application-level information to identify the application class of each individual TCP connection or UDP session. However, this approach cannot be used with

bandwidth links. Also this approach requires precise prior knowledge of applications and their packet formats—something that may not always be possible. Furthermore, the introduction of payload encryption is increasingly limiting our ability to see inside packets for this type of information.

For the above reasons, a different approach is needed.

In accord with the principles disclosed herein CoS mapping is achieved using a statistical method. Advantageously, the disclosed method performs CoS mapping based on simply and easily determined attribute, or attributes of the traffic. Specifically, the disclosed method assigns traffic to classes based on selected attribute or attributes based on a mapping derived from a statistical analysis that forms a signature for traffic having particular values for those attributes.

Thus, in accord with the principles disclosed herein, a three-stage process is undertaken, as depicted in FIG. **1**; to wit,

1. statistics collection—blocks **10** and **20**,
2. classification and rule creation—block **30**, and
3. application of rules to active traffic—block **40**.

Block **10** obtains statistical information, in a training session, relative to selected features for each of a chosen set of classes by using training data that includes collections of traffic, where each collection clearly belongs to one of the chosen classes, and there is found a collection for each of the chosen set of classes. This may be termed statistical "features-class" mapping

Specifically, first the classes of traffic are selected/identified to which administrators of networks may wish to apply different QoS treatment, and traffic from a network having a well-established set of applications that belong to the identified classes (training network) is employed to obtain a set of statistics for a chosen set of features. The notion here is that if it is concluded, from the data of the training network, that feature A of class x applications is characterized by a narrow range in the neighborhood of value Y, then, at a later time, if one encounters traffic in a target network where feature A has the value Y one may be able conclude with a high level of confidence that the traffic belongs to class x.

With respect to class definitions, it makes sense to limit the set of selected classes to those for which corporate network administrators might wish to employ for service differentiation. It is noted that today's corporate networks carry four broad application classes, which are described below, but it should be understood that additional, or other, classes can be selected. The four application classes are:

Interactive: The interactive class contains traffic that is required by a user to perform multiple real-time interactions with a remote system. This class includes such applications as remote login sessions or an interactive: Web interface.

Bulk data transfer: The bulk data transfer class contains traffic that is required to transfer large data volumes over the network without any real-time constraints. This class includes applications such as FTP, software updates, and music or video downloads.

Streaming: The streaming class contains multimedia traffic with real-time constraints. This class includes such applications as streaming and video conferencing.

Transactional. The transactional class contains traffic that is used in a small number of request response pairs that can be combined to represent a transaction. DNS, and Oracle transactions belong to this class.

In order to characterize each application class, it is clear that a reference data set is needed for each class. The problem is that features that ought to be chosen should be ones that characterize and disambiguate the classes. To break this circular dependency, in accord with the principles disclosed herein one or more specific "reference" applications are selected for each class that, based on their typical use, have a low likelihood of being contaminated by traffic belonging to another class. To select those applications, it makes sense to select applications that:

are clearly within one class (to avoid mixing the statistics from two classes);

are widely used, so as to assure we get a good data-set;

have server ports in the well-known port range to reduce the chance of mis-usage of these ports.

In a representative embodiment of the disclosed method, the reference applications selected for each application class are:

Interactive. Telnet,

Bulk data. FTP-data, Kazaa,

Streaming: RealMedia streaming,

Transactional. DNS, HTTPS.

As indicated above, the statistical information that is gathered for each class pertains to the chosen set of features. As for the features that one might consider, it is realized the list of possible features is very large, that the actual selection is left to the practitioner. However, it is beneficial to note that one can broadly classify those features into categories:

1. Simple packet-level features such as packet size and various moments thereof, such as variance, RMS (root mean square) size etc., are simple to compute, and can be gleaned directly from packet-level information. One advantage of such features is that they offer a characterization of the application that is independent of the notion of flows, connections or other higher-level aggregations. Another advantage of such features is that packet-level sampling is widely used in network data collection and has little impact on these statistics.

Another set of statistics that can be derived from simple packet data are time series, from which one can derive a number of statistics; for instance, statistics relating to correlations over time (e.g., parameters of long-range dependence such as the Hurst parameter). An example of this type of classification can be seen in Z. Liu, et al., Profile-based traffic characterization of commercial web sites, *Proceedings of the* 18^{th} *International Teletraffic Congress* (*ITC*-18), volume 5a, pages 231-240, Berlin, Germany, 2003, where the authors use time-of-day traffic profiles to categorize web sites.

2. Flow-level statistics are summary statistics at the grain of network flows. A flow is defined to be a unidirectional sequence of packets that have some field values in common, typically, the 5-tuple (source IP, destination IP, source port, destination port, IP Protocol type). Example flow-level features include flow duration, data volume, number of packets, variance of these metrics etc. There are some more complex forms of information one can also glean from flows (or packet data) statistics; for instance, one may look at the proportion of internal versus external traffic within a category—external traffic (traffic to the Internet) may have a lower priority within a corporate setting. These statistics can be obtained using flow-level data collected at routers using, e.g., Cisco NetFlow, described in White paper—netflow services and applications, http://www.cisco.com/warp/public/cc/pd/iosw/ioft/neflct/tech/napps_wp.htm. These do not require the more resource-intensive process of finer grain packet-level traces. A limitation is, that flow-collection may sometimes aggregate

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.