Multiaccess Protocols in Packet Communication Systems

FOUAD A. TOBAGI, MEMBER, IEEE

(Invited Paper)

Abstract—**The need for multiaccess protocols arises whenever a resource is shared by many independent contending users. Two major factors contribute to such a situation: the need to share expensive resources in order to achieve their *efficient utilization*, or the need to provide a *high degree of connectivity* for communication among independent subscribers (or both). In data transmission systems, the communication bandwidth is often the prime resource, and it is with respect to this resource that we view multiaccess protocols here. We give in this paper a unified presentation of the various multiaccess techniques which we group into five categories: 1) fixed assignment techniques, 2) random access techniques, 3) centrally controlled demand assignment techniques, 4) demand assignment techniques with distributed control, and 5) mixed strategies. We discuss their applicability to different environments, namely, satellite channels, local area communication networks and multihop store-and-forward broadcast networks, and their applicability to different types of data traffic, namely stream traffic and bursty traffic. We also present the performance of many of the multiaccess protocols in terms of bandwidth utilization and message delay.**

I. INTRODUCTION

THE need for multiaccess protocols arises whenever a resource is shared (and thus accessed) by a number of independent users. One main reason contributing to such a situation is the need to *share scarce and expensive resources*. An excellent example is typified by time-sharing systems. Time-sharing was developed in the 1960's to make the powerful processing capability of a large computer system available to a large population of users, each of whom has relatively small or infrequent demands so that a dedicated system cannot be economically justified. Two advantages are gained: the smoothing effect of large populations on the demand resulting from the law of large numbers and a lower cost per unit of service resulting from the (almost always existing) economy of scale.

A second major reason contributing to the multiaccess of a common resource by many independent entities is the need for communication among the entities; we refer to this as the *connectivity requirement*. An excellent example today is the telephone system, the main purpose of which is to provide a high degree of connectivity among its subscribers. The multiaccess protocol used in the telephone system is conceptually simple; it merely consists of placing a request for connection to one or several parties, a request which gets honored by the system if all the required resources are available.

Packet Communication

Let us now consider data communication systems, the subject of interest in this paper. Communications engineers have long recognized the need to multiplex expensive transmission facilities and switching equipment. The earliest techniques for doing this were synchronous time-division multiplexing and frequency-division multiplexing. These methods assign a fixed subset of the time-bandwidth space to each of several subscribers and are very successful for stream-type traffic such as voice. With computer traffic however, usually characterized as *bursty*, fixed assignment techniques are not nearly so successful, and to solve this problem, *packet communication systems* have been developed over the past decade [1]-[7]. Packet communication is based on the idea that part or all of the available resources are allocated to one user at a time but for just a short period of time. Here each component of the system is itself a resource which is multiaccessed and shared by the many contending users. To achieve sharing at the component level, customers are required to divide their messages into small units called packets which carry information regarding the source and the intended recipient.

One type of packet communication network, known as the *point-to-point store-and-forward* network, is one where packet switches are interconnected by point-to-point data circuits according to some topological structure. Packets are transmitted independently and pass asynchronously from one switch to another until they reach their destination. The multiplexing of packets on a channel is done by queueing them at each switch until the outgoing channel is free. Typical examples are the ARPANET [7], the Cigale subnetwork [8], TELENET [9], and DATAPAC [10].

Another type of packet transmission network is the (single-hop) *multiaccess/broadcast* network typified by the ALOHA network [11], SATNET [12], and ETHERNET [5]. Here a *single* transmission medium is shared by all subscribers; the medium is allocated to each subscriber for the time required to transmit a single packet. The inherent single-hop broadcast nature of these systems achieves full connectivity at small additional cost. Each subscriber is connected to the common channel through a smart interface which listens to all transmissions and absorbs packets addressed to it.

Yet a third type of packet network can be identified. It is the (multihop) *store-and-forward multiaccess/broadcast* type which combines the features exhibited (and problems encountered) in the two types just mentioned. The best and perhaps only example of this type is the packet radio network (PRNET) sponsored by the Advanced Research Projects Agency [13], [14]. The PRNET is an extension of the

Manuscript received September 6, 1979; revised January 7, 1980. This work was supported by the Defense Advanced Research Projects Agency under Contract MDA 903-79-C-0201, Order A03717, monitored by the Office of Naval Research.

The author is with the Computer Systems Laboratory, Stanford, University, Stanford, CA 94305.

0090-6778/80/0400-0468$00.75 © 1980 IEEE

ALOHA network in that it includes many added features such as direct communication by a ground radio network between *mobile* users over *wide* geographical areas, coexistence with possibly different systems in the same frequency band, antijam protection, etc. The key requirement of direct communication over wide geographical areas renders store-and-forward switches, called repeaters, integral components of the system. Furthermore, for easy communication among mobile users and for rapid deployment in military applications, all devices employ omnidirectional antennas and share a high-speed radio channel; hence the multiaccess/broadcast nature of the system.

The main issue of concern here is how to control access to a common channel to efficiently allocate the available communication bandwidth to the many contending users. The solutions to this problem form the set of protocols known as *multiaccess protocols*. These protocols and their performance differ according to the environment in question and the system requirements to be satisfied. We devote the next few paragraphs to summarizing the basic relevant characteristics underlying these environments.

Consider first *satellite channels*. A satellite transponder in a geostationary orbit above the earth provides long-haul communication capabilities. It can receive signals from any earth station in its coverage pattern and can transmit signals to all such earth stations (unless the satellite uses spot beams). Full connectivity and multidestination addressing can both be readily accommodated. The many characteristics regarding data rates, error rates, satellite coverage, channelization, and design of earth stations have been fully discussed in a recent paper by Jacobs *et al.* [12]. Perhaps the most important characteristic relevant to this discussion is the inherent long propagation delay of approximately 0.25 s for a single hop. This delay which is usually long compared to the transmission time of a packet, has a major impact on the bandwidth allocation techniques and on the error and flow control protocols.

In *ground radio* environments, the propagation delay is relatively short compared to the transmission time of a packet, and as we shall see in the sequel, this can be of great advantage in controlling access to a common channel. It is important however to distinguish single-hop environments where direct full connectivity is assumed to prevail, and more complex user environments where, due to geographical distance and/or obstacles opaque to UHF signals, limited direct connectivity is achieved. Clearly, the latter situation is significantly more complex as it gives rise to a multihop system where global control of system operation and resource allocation (whether centralized or distributed) is much harder to accomplish. Another dimension of complexity results from the fact that, unlike satellite environments where earth stations are stationary, ground radio systems must also support mobile users. With mobile users, not only does demand on the system exhibit relatively fast dynamic changes, but the radio propagation characteristics are subject to important variations in received signal strength so that system connectivity is at all times difficult to predict; with these considerations it is important to devise access schemes and system control mechanisms that allow the system to adapt itself to these changes. Furthermore, multipath effects in urban environments can be so disastrous that special signaling schemes, such as spread spectrum, may be in order [14]. Finally, another point of growing concern today is RF spectrum utilization. This is becoming an increasingly predominant factor in determining the structure of radio systems, both in satellite and ground environments. A packet radio system which allows the dynamic allocation of the spectrum to a large population of bursty mobile user needs flexible high performance multiaccess schemes which can take advantage of the law of large numbers, and which permit coexistence of the system with other (possibly different) systems in the same frequency band.

Finally, we consider *local area communication* systems. These span short distances (ranging from a few meters up to a few kilometers) and usually involve high data rates. The transmission medium can be privately owned and inexpensive, such as twisted pair or coaxial cable. Local area environments are characterized by a large and often variable number of devices requiring interconnection, and these are often inexpensive. These situations call for communication networks with simple topologies and simple and inexpensive connection interfaces that can provide great flexibility in accommodating the variability in the environment and that achieve the desired level of reliability. With these constraints, we again face the situation in which a high bandwidth channel is to be shared by independent users. Short propagation delays and high data rates are the main characteristics that are exploited in devising multiaccess schemes appropriate to local area environments.

Multiaccess schemes are evaluated according to various criteria. The performance characteristics that are desirable are, first of all, high bandwidth utilization and low message delays. But a number of other attributes are just as important. The ability for an access protocol to simultaneously support traffic of different types, different priorities, with variable message lengths, and differing delay constraints is essential as higher bandwidth utilization is achieved by the multiplexing of all traffic types. Also, to guarantee proper operation of schemes with distributed control, robustness, defined here as the insensitivity to errors resulting in misinformation, is also most desirable.

Having so far discussed briefly the basic characteristics and system requirements underlying the various communication environments, we now proceed with a discussion of the multiaccess protocols appropriate to these environments.

II. MULTIACCESS PROTOCOLS

Multiaccess protocols differ by the static or dynamic nature of the bandwidth allocation algorithm, the centralized or distributed nature of the decision-making process, and the degree of adaptivity of the algorithm to changing needs. Accordingly, these protocols can be grouped into five classes. The first class, labeled *fixed assignment techniques*, consists of those techniques which allocate the channel bandwidth to the users in a static fashion, independently of their activity. The second class is that of *random access techniques*. In this class the entire bandwidth is provided to the users as a single channel to be accessed randomly; since collisions may result which degrade the performance of the channel, improved performance can be achieved by either synchronizing users so that their transmissions coincide with the boundaries of time slots, by sensing carrier prior to transmission, or both. The third and fourth classes correspond to *demand assignment* techniques. Demand assignment techniques require that explicit control information regarding the users' need for the communication resource be exchanged. A distinction is made between those techniques in which the decision-making is centralized (constituting the third class in question), and those techniques in which all users individually execute a distributed algorithm based on control information exchanged among them. The latter constitute the fourth class. The fifth class, labeled *adaptive strategies and mixed modes*, includes those techniques which consist of a mixture of several distinct modes, and those strategies in which the choice of an access scheme is itself adaptive to the varying need, in the hope that near-optimum performance will be achieved at all times.

We describe here the various protocols known today, either implemented or proposed, and discuss their performance and applicability to the different environments introduced in Section I. For this we consider the (conceptually) simplest situation consisting of M users wishing to communicate over a channel. This situation arises typically in a satellite communication environment or in a single-hop ground radio environment.

A. Fixed Assignment Techniques

Fixed assignment techniques consist of allocating the channel to the user, independently of their activity, by partitioning the time-bandwidth space into slots which are assigned in a static predetermined fashion. These techniques take two common forms: *orthogonal*, such as frequency division multiple access (FDMA) or synchronous time division multiple access (TDMA), and "*quasi-orthogonal*" such as code division multiple access (CDMA).

1) *FDMA and TDMA*: FDMA consists of assigning to each user a fraction of the bandwidth and confining its access to the allocated subband. Orthogonality is achieved in the frequency domain. FDMA is relatively simple to implement and requires no real time coordination among the users.

TDMA consists of assigning fixed predetermined channel time slots to each user; the user has access to the entire channel bandwith, but only during its allocated slots. Here, signaling waveforms are orthogonal in time.

In the author's opinion, a number of disadvantages exist for FDMA when compared to TDMA. FDMA wastes a fraction of the bandwidth to achieve adequate frequency separation. FDMA is also characterized by a lack of flexibility in performing changes in the allocation of the bandwidth and certainly the lack of broadcast operation. The major disadvantages in TDMA are the need to provide A/D converters for overlap traffic such as voice, and rapid burst synchronization and sufficient burst separation to avoid time overlap. However, it has been shown that guard bands of less than 200 ns are achievable (as in INTELSAT's MAT-1 TDMA system, for example) and many operational systems are moving towards the use of TDMA [16]. Timing at an earth

either explicitly by a reference station, or implicitly by measurement of the propagation delay from the earth station to the transponder. In order to allow the TDMA modems to acquire frequency, phase, bit timing and bit framing synchronization for each received burst, a preamble is included in front of each burst requiring typically from 100 to 200 bit times. Thus clearly, TDMA is more complex to implement than FDMA, but an important advantage is the connectivity which results from the fact that all receivers listen to the same channel while senders transmit on the same common channel at different times. Accordingly, many network realizations, both in ground and satellite environments, are easier to accomplish [12], [14].

From the performance standpoint it has also been established that TDMA is superior to FDMA in many cases of practical interest. I. Rubin has shown that the random variable representing packet delay is always larger in FDMA than in TDMA [17] for comparable systems. Lam derived the average message delay for a TDMA system with multipacket messages and a nonpreemptive priority queue discipline [18]. There, too, it was shown that TDMA is superior to FDMA.

For both FDMA and TDMA, the fixed preallocation of the frequency or time resource does not have to be equal for all users, but can be tailored to fit their needs (assumed constant). Kosovych studied two TDMA implementations [19]. In the first, called *contiguous assignment*, the users are cyclically ordered in the time sequence in which they have access to the channel. Each user is periodically assigned its *own* fixed time duration. In the second implementation, called *distributed allocation*, all access periods are of equal time duration, but the frequency of accesses can be different from one user to the other. It was shown that for situations in which the transmission overhead (defined as guard time and synchronization preamble time) is large, the contiguous fixed assignment implementation is better suited and provides substantially better performance than distributed fixed assignments, while when the transmission overhead is small, distributed fixed assignments provide slightly better performance.

Finally we note that, even though the allocation can be tailored to the relative need of each user, fixed allocation can be wasteful if the users' demand is highly bursty, as we shall explicitly see in the sequel. Given these limitations, one may increase the channel utilization beyond FDMA and TDMA by using asynchronous time division multiple access (ATDMA), also known as statistical multiplexing [70]. Basically the technique consists of switching the allocation of the channel from one user to another only when the former is idle and the latter is ready to transmit data. Thus the channel is *dynamically* allocated to the various users according to their need. The performance of ATDMA in packet communication systems corresponds to that of a work-conserving single server queueing system, and is the best we can achieve under unpredictable demand. Unfortunately, it is not always possible to accomplish the necessary coordination among the users. This mode of multiplexing is possible only when several colocated users (such as at the same earth station) are sharing a single point-to-point channel.

ple access allows overlap in transmission both in the frequency and time coordinates. It achieves orthogonality by the use of different signaling codes in conjunction with matched filters (or equivalently, correlation detection) at the intended receivers. Multiple orthogonal codes are obtained at the expense of increased bandwidth requirements (in order to spread the waveforms); this also results in a lack of flexibility in interconnecting all users (unless, of course, matched filters corresponding to all codes are provided at all receivers). However, CDMA has the advantage of allowing the coexistence of several systems in the same band, as long as different codes are used for different systems. Moreover, it is also possible to separate, by "capture," time overlapping signaling waveforms with the same code, thus achieving connectivity and efficient spectrum utilization. This interesting possibility falls into the class of random access techniques and is addressed in the following subsection.

B. Random Access Techniques

In computer communication, much data traffic is characterized as bursty e.g., interactive terminal traffic. Burstiness is a result of the high degree of randomness seen in the message generation time and size, and of the relatively low-delay constraint required by the user. If one were to observe the user's behavior over a period of time, one would see that the user requires the communications resources rather infrequently; but when he does, he requires a rapid response. That is, there is an inherently large peak-to-average ratio in the required data transmission rate. If fixed subchannel allocation schemes are used, then one must assign enough capacity to each subscriber to meet his peak transmission rates with the consequence that the resulting channel utilization is low. A more advantageous approach is to provide a single sharable high-speed channel to the large number of users. The strong law of large numbers then guarantees that, with a very high probability, the demand at any instant will be approximately equal to the sum of the average demands of that population. As stated in the introduction, packet communication is a natural means to achieve sharing of the common channel. When dealing with shared channels in a packet-switched mode, one must be prepared to resolve conflicts which arise when more than one demand is placed upon the channel. For example, in packet-switched radio channels, whenever a portion of one user's transmission overlaps with another user's transmission, the two collide and "destroy" each other (unless a code division multiple-access scheme is used). The existence of some positive acknowledgment scheme permits the transmitter to determine if his transmission is successful or not. The problem is how to control the access to the common channel in a fashion which produces, under the physical constraints of simplicity and hardware implementation, an acceptable level of performance. The difficulty in controlling a channel which must carry its own control information has given rise to the so-called random-access protocols, among others. We describe these here by considering again single-hop environments.

1) *ALOHA [20]-[22]*: Historically, the *pure ALOHA* protocol was first used in the ALOHA system, a single-hop terminal access network developed in 1970 at the University of Hawaii, employing packet-switching on a radio channel [11], [20]. The simplest of its kind, pure ALOHA permits a user to transmit any time it desires. If they do so, and within some appropriate time-out period it receives an acknowledgment from the destination (the central computer), then it knows that no conflict occurred. Otherwise it assumes that a collision occurred and it must retransmit. To avoid continuously repeated conflicts, the retransmission delay is randomized across the transmitting devices, thus spreading the retry packets over time. A slotted version, referred to as *slotted ALOHA*, is obtained by dividing time into slots of duration equal to the transmission time of a single packet (assuming constant-length packets)[21], [22]. Each user is required to synchronize the start of transmission of its packets to coincide with the slot boundary. When two packets conflict, they will overlap completely rather than partially, providing an increase in channel efficiency over pure ALOHA. Due to conflicts and idle channel time, the maximum channel efficiency available using ALOHA is less than 100 percent, 18 percent for pure ALOHA and 36 percent for sloted ALOHA. Both schemes are theoretically applicable to satellite, ground radio and local bus environments. The slotted version has the advantage of efficiency, but in multihop ground radio, it has the disadvantage that synchronization may be hard to achieve.

Although the maximum achievable channel utilization is low, the ALOHA schemes are superior to fixed assignment schemes when there is a large population of bursty users. This point is illustrated in comparing the performance of FDMA with that of slotted ALOHA when M users, each of which generates packets at a rate of Λ packets per second, share a radio channel of W•Hz [23]. Figs. 1 and 2 display the constant delay contours in the (M, Λ) and (W, Λ) planes, respectively, showing the important improvement gained in terms of bandwidth required, population size supported, and delay achieved when the users are bursty.

2) *Carrier Sense Multiple Access (CSMA) [24], [25]*: In ground radio environments the channel can be characterized as wideband with a propagation delay between any source-destination pair that is small compared to the packet transmission time. In such an environment one may attempt to avoid collisions by listening to the carrier due to another user's transmission before transmitting, and inhibiting transmission if the channel is sensed busy. This feature gives rise to a random access scheme known as carrier sense multiple access (CSMA) [24], [25]. While in the ALOHA scheme only one action could be taken by the terminals, namely, to transmit, here many strategies are possible so that many CSMA protocols exist differing according to action that a terminal takes to transmit a packet after sensing the channel. In all cases, however, when a terminal learns that its transmission had incurred a collision, it reschedules the transmission of the packet according to the randomly distributed delay. At this new point in time, the transmitter senses the channel again and repeats the algorithm dictated by the protocol. There are two main CSMA protocols known as *nonpersistent* and *p-persistent* CSMA depending on whether the transmission by a station which finds the channel busy

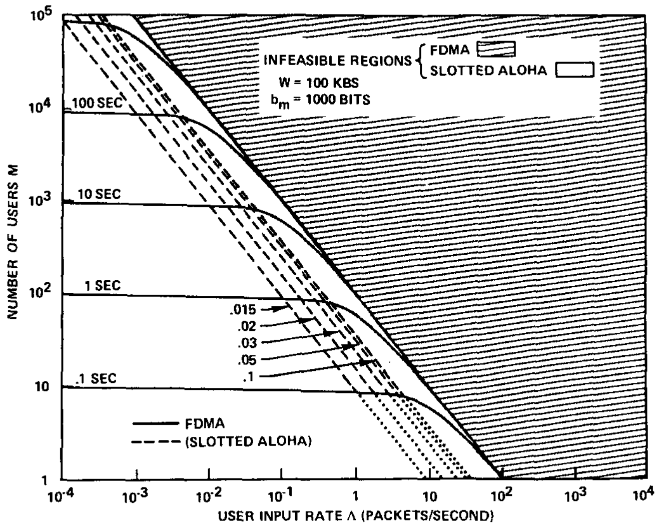

Fig. 1. FDMA and slotted ALOHA access: performance with 100 kbits/s bandwidth and 1000 bit packets [23].

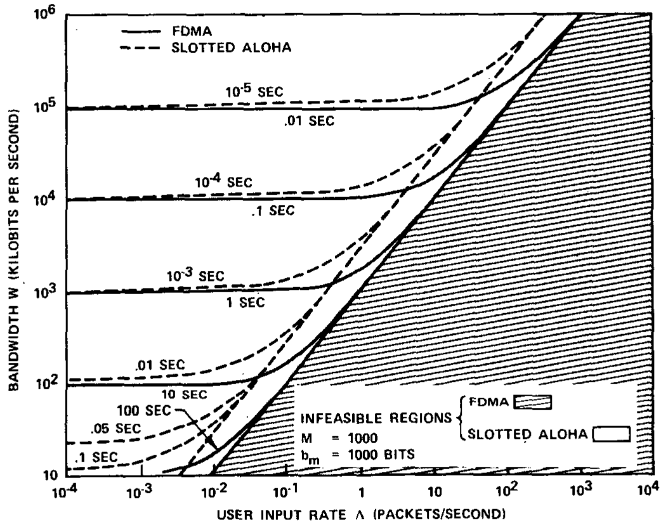

Fig. 2. FDMA and slotted ALOHA random access: bandwidth requirements for 1000 terminals. Contours are for constant delay [23].

is to occur later or immediately following the current one with probability p. Many variants and modifications of these two schemes have also been proposed. Thus, in nonpersistent CSMA, a ready terminal senses the channel and operates as follows:

1) If the channel is sensed idle, it transmits the packet.

2) If the channel is sensed busy, then the terminal schedules the retransmission of the packet to some later time according to the retransmission delay distribution. At this new point in time, it senses the channel and repeats the algorithm described.

The 1-persistent CSMA protocol, a special case of p-persistent CSMA, was devised in order to (presumably) achieve acceptable throughput by never letting the channel go idle if some ready terminal is available. More precisely, a ready terminal senses the channel and operates as follows:

1) If the channel is sensed idle, it transmits the packet with probability one.

2) If the channel is sensed busy, it waits until the channel goes idle and then immediately transmits the packet with probability one (i.e., persisting on transmitting with $p = 1$).

A slotted version of these CSMA protocols can also be considered in which the time axis is slotted and the slot size is τ s where τ is the maximum propagation delay among all pairs. Note that this definition of a slot is different from that used in the description of slotted ALOHA. Here a packet transmission time is equivalent to several slots. We make this distinction by referring to a slot of size τ s as a "minislot." All terminals are synchronized and are forced to start transmission only at the beginning of a minislot. When a packet's arrival occurs in a minislot, the terminal waits until the next minislot boundary and operates according to the protocols described above.

In the case of a 1-persistent CSMA, we note that whenever two or more terminals become ready during a packet transmission period, they wait for the channel to become idle (at the end of that transmission) and then they all transmit with probability one. A conflict will also occur with probability one. The idea of randomizing the starting time of transmission of packets accumulating at the end of a transmission period seems reasonable for interference reduction and throughput improvement. Thus we have the p-persistent scheme which involves including an additional parameter p, the probability that a ready packet persists ($1 - p$ being the probability of delaying transmission by τ seconds, the propagation delay). The parameter p is chosen to reduce the level of interference while keeping the idle periods between any two consecutive nonoverlapped transmission as small as possible.

More precisely, the p-persistent CSMA protocol consists of the following: the time axis is minislotted and the system is synchronized such that all terminals begin their transmission at the beginning of a minislot. If a ready terminal senses the channel idle, then with probability p, the terminal transmits the packet; and with probability $1 - p$, the terminal delays the transmission of the packet by τ seconds (i.e., one minislot). If at this new point in time, the channel is still detected idle, the same process is repeated. Otherwise some packet must have started transmission, and the terminal in question schedules the retransmission of the packet according to the retransmission delay distribution (i.e., acts as if it had conflicted and learned about the conflict). If the ready terminal senses the channel busy, it waits until it becomes idle (at the end of the current transmission) and then operates as above.

Packet broadcasting technology has also been shown to be very effective in satisfying many local area in-building communication requirements. A prominent example is ETHERNET, a local communication network which uses CSMA on a tapped coaxial cable to which all the communicating devices are connected [5]. The device connection interface is a passive cable tap so that failure of an interface does not prevent communication among the remaining devices. The use of a single coaxial cable achieves broadcast communication. The only difference between this and the single-hop radio is that, in addition to sensing carrier, it is possible for the transceivers, when they detect interference among several transmissions (including their own), to abort the transmission

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**