♦ Real-Time Traffic over the IEEE 802.11 Medium Access Control Layer

João L. Sobrinho and A. S. Krishnakumar

This paper proposes multiple access procedures to transport real-time traffic over IEEE 802.11 wireless local area networks (LANs). As currently defined, the IEEE 802.11 standard supports real-time traffic by switching from its normal, distributed access mode—carrier sense multiple access/collision avoidance (CSMA/CA)—to a centralized one. The centralized mode severely constrains the operation of wireless LANs and provides inadequate performance. Our proposed procedures, on the other hand, are totally distributed and can be overlaid on CSMA/CA. They use the carrier sense capabilities of the network interfaces and require only the ability to jam the channel with pulses of energy of specified duration. The resulting scheme guarantees priority to real-time traffic and provides round-robin service and bounded access delays to real-time stations. This paper examines the behavior of the new access techniques to derive conditions under which they can be considered stable. In addition, it presents simulation results that assess the impact of these access procedures on the average delay of data packets. The simulations are also used to provide estimates of the number of real-time stations that can be supported under various network operating conditions.

Introduction

Wireless local area networks (LANs) have been used for the past several years to untether data applications in the local area. In addition, they are now being viewed with interest as a means of supporting real-time voice and video applications, even though the characteristics and performance requirements of data and real-time traffic differ substantially. Real-time traffic requires bounded end-to-end delays beyond which information loses its value and may be discarded, whereas the delay requirements of data traffic are less stringent.

The IEEE has formed study group IEEE 802.11 to develop a wireless LAN standard, which is nearing completion.[1] The medium access control (MAC) layer of this standard recommends using carrier sense multiple access/collision avoidance (CSMA/CA) as the basic access mechanism. This scheme is a variant of the well-known CSMA with collision detection (CSMA/CD) used in hard-wired LANs and adapted to the constraints of the wireless environment. The probabilistic nature of CSMA/CD has long been recognized as unsuitable for supporting real-time applications.[2-7] Supporting real-time traffic is likely to be even more problematic in wireless LANs:

- Current wireless LANs operate at 2 Mb/s, whereas hard-wired LANs that use CSMA/CD operate at 10 Mb/s. Therefore, for the same relative loads, wireless LANs have higher queuing delays.
- For comparable system parameters, the CSMA/CA protocol has worse throughput-delay characteristics than CSMA/CD.
- While different segments of a hard-wired LAN are electromagnetically isolated, neighboring cells of a wireless LAN interfere with each

other, thereby decreasing their individual throughputs.

Moreover, the most effective solutions to give priority to real-time traffic in CSMA/CD-based LANs rely on the collision detection ability of hard-wired transceivers, a feature which is absent in wireless communications.

The MAC layer of the IEEE 802.11 standard also defines an optional centralized access mode based on polling, which was claimed to be more suitable to support real-time traffic than CSMA/CA. However, the use of a centralized scheme imposes heavy constraints on the operation of wireless LANs. First, the centralized mode cannot be operated simultaneously in neighboring cells. Second, it requires the existence of an access point with specialized access functions. In addition, Visser and El Zarki[8] conclude that the centralized scheme of the IEEE 802.11 standard performs poorly.

This paper describes multiple access procedures that allow real-time stations to access a CSMA/CA wireless LAN with priority. The proposed access procedures are totally distributed, guarantee bounded access delays to real-time traffic, and can be overlaid on an IEEE 802.11 implementation. They use the carrier sense capabilities of the IEEE 802.11 network interfaces and require only the ability to jam the channel with pulses of energy of specified duration. The real-time stations appear to be accessing a time division multiplexed (TDM) transmission structure, which is constantly being perturbed and repositioned by data packet transmissions. In contrast to the scheme proposed by Maxemchuk and Netravali[5,9] for real-time transmission over CSMA/CD hard-wired LANs, our scheme does not rely on CD capabilities and does not require data packets to be shorter than real-time packets.

The section "Medium Access Procedures" describes the proposed multiple access procedures, and "The Dynamics of Real-Time Stations" presents a formal analysis of its properties. The proofs of the propositions in that section can be skipped on a first reading without loss of continuity. "Results and Discussion" assesses the performance of the access mechanisms for real-time traffic, along with their impact on the average data packet delays.

Panel 1. Abbreviations, Acronyms, and Terms.

CA—collision avoidance
CD—collision detection
CSMA—carrier-sense multiple access
LAN—local area network
MAC—medium access control
TDM—time division multiplexed

Medium Access Procedures

This paper considers a wireless LAN constituted by data stations and real-time stations sharing a common radio channel in time. Unless stated otherwise, we assume that all stations can sense each other's transmissions. The data stations regulate their access to the channel according to the CSMA/CA protocol specified by the IEEE 802.11 standard, which is briefly reviewed in the next section, "CSMA/CA as the Access Procedures of Data Stations." The real-time stations follow the access procedures in the subsequent section, "Access Procedures of Real-Time Stations," which give them access priority over the data stations.

The maximum propagation delay between pairs of stations is denoted by τ, and it includes the sensing delay as well as the turnaround times of the wireless transceivers. The operation of the integrated access procedures described requires the definition of three interframe spacings, denoted t_{short}, t_{med}, and t_{long}, respectively, such that

$$t_{short} + 2\tau < t_{med},$$
$$t_{med} + 2\tau < t_{long}.$$

The use of these interframe spacings will become clear in the sequel. Loosely, t_{short} and t_{long} are related to the operation of CSMA/CA, and t_{med} will be used for the access procedures of the real-time stations. The IEEE 802.11 standard already specifies three interframe spacings with the above properties, thereby allowing the access procedures described here to be overlaid on a standard implementation.

CSMA/CA as the Access Procedures of Data Stations

Consider a data station with a new packet ready for transmission at time t. If the channel was idle in the interval $(t - t_{long}, t\,]$, then the station transmits the packet without further ado. Otherwise, it waits until the channel has been idle for t_{long} consecutive seconds and enters the backoff mode. Similarly, a station whose packet has experienced c collisions waits until the channel is perceived idle for t_{long} consecutive seconds and enters the backoff mode. In this mode the station first initializes a timer with a value of $t_{wslot} \times$ rand $[f_{data}(c)]$ seconds, where t_{wslot}, $t_{wslot} > 2\tau$, is the duration of a *white slot*, the function rand $[b]$ returns a random number between 0 and $(b - 1)$, and $f_{data}(c)$ is the backoff function given by

$$f_{data}(c) = f_{data}(0) \times 2^{c}, \quad c = 0, 1, \ldots.$$

The timer counts down while the channel is perceived idle for more than t_{long} seconds, and the packet is transmitted as soon as the timer reaches 0.

Collisions can occur due to nonzero propagation delays. Given that a transmission has been started, the stations will only sense the medium busy after a propagation delay has elapsed; if another transmission is initiated in the meantime, a collision will occur. Moreover, even if the propagation delays were arbitrarily small, collisions would still occur—as, for example, when two stations entering backoff mode while the medium is busy load their timers with the same number of backoff slots.

An immediate positive acknowledgment scheme is used in which the recipient of a correctly received packet returns an acknowledgment minipacket starting t_{short} seconds after the end of packet reception. A data station that does not receive an acknowledgment minipacket in response to its packet transmission assumes that a collision has occurred and attempts to retransmit the packet according to the rules of the protocol.

Access Procedures of Real-Time Stations

Real-time stations generate continuous streams of information bits, referred to as calls, for long periods of time. During a call, real-time stations expect to have undisputed access to the shared radio channel at regular time intervals. This is accomplished with the access mechanisms described in the next section, "Basic Operation." Those mechanisms ensure that once the first packet of a call is successfully transmitted, the ensuing packet transmissions do not collide with other real-time packets. The subsequent section, "Bandwidth Control," elaborates on the ways in which a real-time station can use the instants when it accesses the channel. The section "Negative Acknowledgment Scheme" discusses the implementation of a scheme in which only unsuccessful transmissions elicit a response from the receiver.

Basic operation. The *access instants* of a real-time station are defined here as the instants of time in which the station acquires access to the channel to transmit a packet. Whenever a real-time station has an access instant, it performs two actions:

1. It transmits a packet that must last for at least t_{pkt} seconds, and
2. It schedules the next access instant to occur t_{sch} seconds in the future.

Suppose, now, that a real-time station has an access instant scheduled to occur at time t. If the channel was perceived to be idle in the interval $(t - t_{med}, t]$ and remains idle in the ensuing t_{obs} seconds, with $t_{obs} > 2\tau$, the station has an access instant at this later time. Otherwise, the station waits until the channel is perceived idle for t_{med} consecutive seconds and enters into a *black burst contention period*. At this time, the station jams the channel with a number of *black slots* (pulses of energy of prespecified duration). This number is proportional to the time that the station has been waiting for the channel to become idle. Specifically, if the station has been waiting for d seconds to access the channel, it transmits a black burst of duration $t_{bslot} \times \lceil d / t_{unit} \rceil$ seconds, where t_{bslot}, $t_{bslot} > 2\tau$, is the length of a black slot and t_{unit} is a system parameter to be defined shortly.

After transmitting its black burst, the station waits for t_{obs} seconds to see if any other station is transmitting a longer black burst, implying that it would have been waiting longer for access to the channel. If the channel is perceived idle after the t_{obs} interval, then the station has an access instant and transmits a packet. Otherwise, it waits again for the next time that the channel is perceived idle for t_{med} consecutive seconds and repeats the algorithm.

The observation interval, t_{obs}, has to be less than t_{bslot}, so that a real-time station always recognizes when its black burst is shorter than that of another station. It also has to be less than t_{med}, so that the real-time stations do not access the channel to transmit

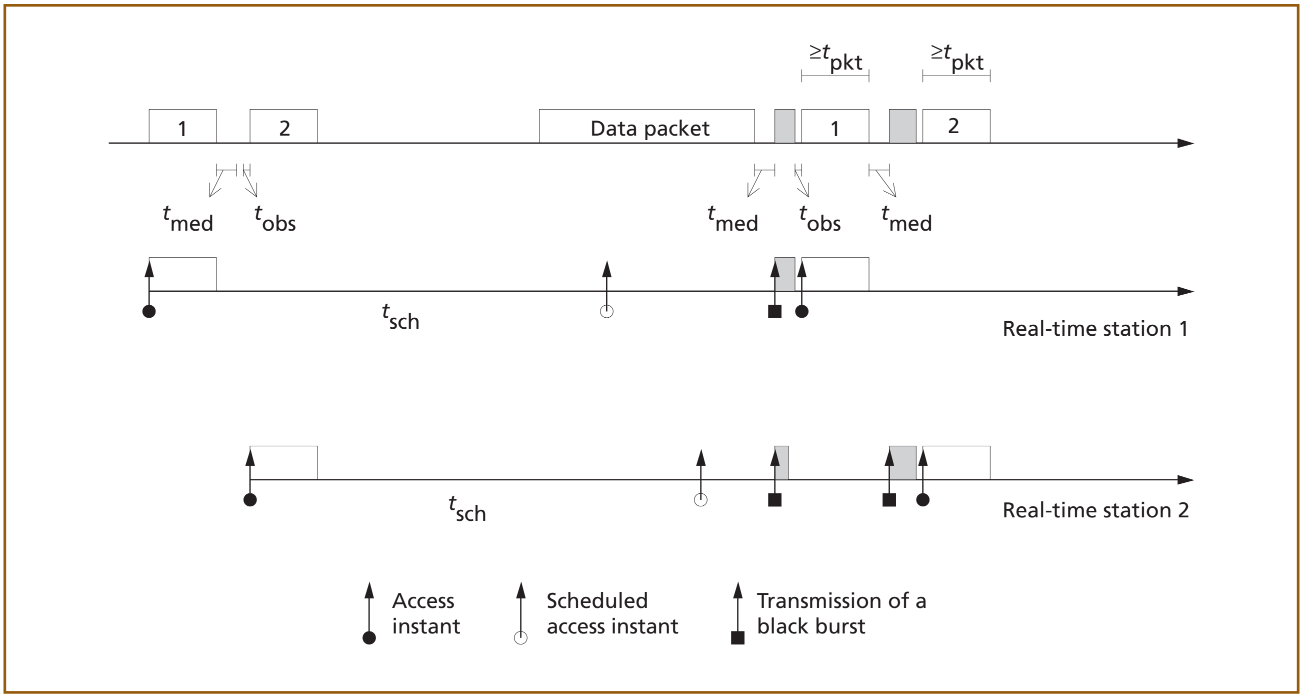

Figure 1.
A time diagram of the access procedures of real-time stations.

black bursts during the observation interval. In conclusion, $2\tau < t_{obs} < \min(t_{bslot}, t_{med})$. As to t_{unit}, the access instants of the real-time stations are shifted in time by at least $(t_{obs} + t_{pkt} + t_{med})$ seconds. This implies that the access delays of distinct real-time stations that find the channel busy will differ by at least that amount of time. Therefore, if $t_{unit} \leq (t_{obs} + t_{pkt} + t_{med})$, their black bursts will differ by at least one black slot, implying that every black burst contention period will result in a (unique) winner. The winner is the station that has been waiting the longest for access to the channel.

To ensure that a real-time station recognizes at the beginning of a call when it has acquired undisputed access to the channel, this station can use the access procedures of data stations to transmit its first packet. When an acknowledgment is received, the station knows that its previous access instant is shifted in time from those of the other real-time stations. From that point on, the access procedures of real-time stations ensure that the packets of this station do not collide with those of other real-time stations.

Figure 1 shows how the protocol operates. The access instants of stations 1 and 2 get perturbed by a data packet transmission. Station 2 contends twice with black bursts, and its second black burst is longer than its first because the delay from the scheduled access instant has increased. The protocol has two important characteristics. First, the real-time station that has been waiting the longest for access to the channel wins the next black burst contention period. This effectively ensures that real-time stations access the channel to transmit their packets in a round-robin order. Second, the real-time stations get priority over the data stations, because no data station will perceive the channel idle for t_{long} consecutive seconds until all contending real-time stations have obtained access to the channel.

Bandwidth control. The access procedures of the previous section are independent of the characteristics of the traffic to be transported. Whenever a station has an access instant, it is only required that it transmits for at least t_{pkt} seconds, where t_{pkt} is the same for all stations. The number of information bits actually transmitted at the access instants does not have to be fixed. Indeed, a real-time station generating information bits at a constant nominal rate will typically increase the number of information bits transmitted as a function of the delay it incurred before accessing the channel.[4,5] This compensates for the fact that the time between consecutive access instants of a real-time station may be longer than $(t_{sch} + t_{obs})$ seconds. By the same token, real-time stations with different bandwidth requirements can be supported by having those stations transmit different numbers of information bits when they access the channel. Deciding whether the network can support a new call with specified bandwidth and quality of service requirements is made when the call is set up.

Usually, a real-time application will produce blocks of bits, which are then enqueued in a buffer for transmission. The buffer may have finite capacity and only be able to enqueue the most recently generated blocks of information. When the station has an access instant, it will simply empty the contents of the buffer. The packets transmitted on the channel may be of fixed length, matched to the capacity of the buffer. If the buffer is not full, the fixed-length packets are filled with padding bits. Alternatively, the packets may just contain the blocks found in the buffer at the access instant, plus the required packet overhead.

Most wireless LANs operate with an access point that routes all packets to and from the terminals in its cell. The performance of the system improves if the access point groups real-time packets with different destinations into a frame and only contends for access to the channel with black bursts once for each frame, rather than once for each real-time packet.

Negative acknowledgment scheme. A positive acknowledgment scheme, as proposed in the IEEE 802.11 standard,[1] incurs a penalty in channel efficiency, because every correctly received packet has to be followed by an acknowledgment minipacket. On the other hand, the specific characteristics of the real-time access procedures support a negative acknowledgment scheme, in which a receiver only notifies the transmitter when it does not get a packet that was expected. This is only possible because under the proposed procedures a receiver expects a new real-time packet $(t_{sch} + t_{obs})$ seconds after the previous one, plus some prespecified delay tolerance. Not receiving a new packet within that time implies that the packet is lost or unduly delayed.

Whenever a destination station operating with a negative acknowledgment scheme receives a real-time packet, it schedules the transmission of an invitation minipacket to t_{neg} seconds in the future, with $t_{neg} > (t_{sch} + t_{obs})$. After that time has elapsed, if no new real-time packet has arrived, the destination station starts contending for access to the channel to invite the source station to (re)transmit a packet. The invitation minipacket can contend for access to the channel as if it were a data packet. If a real-time packet arrives in the interim, the destination aborts its attempts to send the invitation minipacket.

The use of invitation minipackets also offers robustness against the hidden stations problem.[10] This problem arises in CSMA wireless networks because a source station inhibits the other stations in its vicinity, rather than those in the vicinity of the destination. However, it is the latter stations that may interfere with the packet transmission from source to destination. If, instead, the destination invites the source to transmit a packet, then, by doing so, it is inhibiting the stations in its vicinity—exactly those that may interfere with the ensuing real-time packet transmission from source to destination. Further details of operation in the presence of hidden stations are still being studied.

The Dynamics of Real-Time Stations

When there is no data traffic, the real-time stations organize themselves in a TDM-like structure, without incurring access delays. Two consecutive access instants of a given real-time station are separated by exactly $t_{acc} \triangleq (t_{sch} + t_{obs})$ seconds. A data packet transmission may perturb this state of affairs, delaying the access instants of the real-time stations and forcing them to contend for access to the channel with black bursts. Eventually, the real-time stations will recover from the perturbation and reorganize themselves into a state in which they no longer incur access delays. This is the behavior that we want to formalize mathematically.

We assume zero propagation delays, real-time calls with infinite holding time, and real-time stations with the same bandwidth requirements. We further consider that, whenever a real-time station has an access instant, it transmits a packet of duration t_{pkt} seconds, independently of the delay it incurred before accessing the channel. The fixed-duration packets can be regarded as containers that accommodate the bits accumulated between two consecutive access instants, up to a prespecified maximum. Define $t_{inter} \triangleq (t_{obs} + t_{pkt} + t_{med})$. There are n real-time stations in the system, labeled from 0 to $(n-1)$, where n satisfies the equation

$$nt_{inter} + \varepsilon = t_{acc}, \text{ with } \varepsilon > 0. \qquad (1)$$

The duration of a black burst is proportional to the delay incurred by the real-time station in accessing the channel. The proportionality constant, α, $\alpha > 0$, is the ratio between the duration of a black slot and t_{unit}, $t_{unit} \le t_{inter}$:

$$\alpha = \frac{t_{bslot}}{t_{unit}} \qquad (2)$$

Station 0 had an access instant at time $t_{0,-1} = -t_{acc}$. Its next access instant was scheduled for time $t = -t_{obs}$, and a packet would be transmitted at $t_{00} = 0$ if the channel were idle at that time. Station i, $i = 1,\ldots,n-1$, had an access instant at time $t_{i,-1}$, such that

$$t_{i,-1} - t_{i-1,-1} = t_{inter} + \varepsilon_i, \quad i = 1,\ldots,n-1, \qquad (3)$$

where $\varepsilon_i \ge 0$, $i = 1,\ldots,n-1$, and $\sum_{i=1}^{n-1} \varepsilon_i \le \varepsilon$. The initial conditions of the system are described by the vector $\bar{\varepsilon} \triangleq (\varepsilon_1,\ldots,\varepsilon_{n-1})^t$, where $\bar{x}^t$ denotes the transpose of $\bar{x}$. There is an initial perturbation, due to a data packet transmission, that extends from $t = 0$ to $t = (T - t_{obs} - t_{med})$, $T \ge (t_{obs} + t_{med})$. Let d_{ik} denote the access delay of station i, $i = 0,\ldots,n-1$, in round k, $k = 0,1\ldots$, measured from the instant an access is scheduled until the instant when the station starts transmission of the black burst with which it wins the contention. The corresponding black burst has length αd_{ik} (see **Figures 2** and **3**). The vector of access delays in round k is defined by $\bar{d}_k \triangleq (d_{1k},\ldots,d_{n-1,k})^t$.

Definition 1 *Given the magnitude of the initial perturbation, T, the system is stable if and only if whatever the initial conditions,* $\bar{\varepsilon}$, *there is an* l, $l \ge 0$, *such that* $\bar{d}_k = \bar{0} \triangleq (0,\ldots,0)^t$, *for* $k \ge l$.

Definition 2 *The system is unconditionally stable if and only if it is stable no matter the magnitude of the initial perturbation, T.*

Proposition 1 *The initial perturbation causes the access delay of station 0 in round 0 to be* $d_{00} = T$ *and*

$$\bar{d}_0 = \left[T\bar{\xi} - \bar{\bar{B}}\bar{\varepsilon}\right]_+ ,$$

where $\bar{\xi} = (1+\alpha)\,(1,\ldots,(1+\alpha)^{n-2})^t$,

$$\bar{\bar{B}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1+\alpha & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ (1+\alpha)^{n-2} & (1+\alpha)^{n-3} & \cdots & 1 \end{bmatrix},$$

and, by definition,

$$\left[\bar{x}\right]_+ \triangleq (\max(x_1,0),\ldots,\max(x_{n-1},0)) .$$

Proof: From the initial setup, we immediately have $d_{00} = T$. From **Figure 2**, we obtain the following relation between the access delays of consecutive stations in round 0:

$$d_{i0} = \left[(1+\alpha)\,d_{i-1,0} - \varepsilon_i\right]_+$$

$$= \left[(1+\alpha)^i\, d_{00} - \sum_{j=1}^{i} (1+\alpha)^{i-j}\varepsilon_j\right]_+ , \qquad (4)$$

$i = 1,\ldots, n-1$.

The statement of the proposition follows by expressing the previous relations in matrix form. ■

Proposition 2 *The vectors of access delays in consecutive rounds are related by*

$$\bar{d}_k = \left[\bar{\bar{C}}\,\bar{d}_{k-1} - \varepsilon\bar{\xi}\right]_+ , \quad k = 1, 2,\ldots,$$

where the matrix $\bar{\bar{C}}$ *is positive and is given by*

$$\bar{\bar{C}} = \alpha\left(\bar{\xi}\,\bar{1}^t - \bar{\bar{B}}\right),$$

with $\bar{1} = (1, 1,\ldots, 1)^t$.

Proof: With the help of **Figure 3a**, we conclude that the access delay of station 0 in round k is given by

$$d_{0k} = \left[\alpha \sum_{j=1}^{n-1} d_{j,k-1} + nt_{inter} - t_{obs} - t_{sch}\right]_+$$

$$= \left[\sum_{j=1}^{n-1} d_{j,k-1} - \varepsilon\right]_+ , \qquad (5)$$

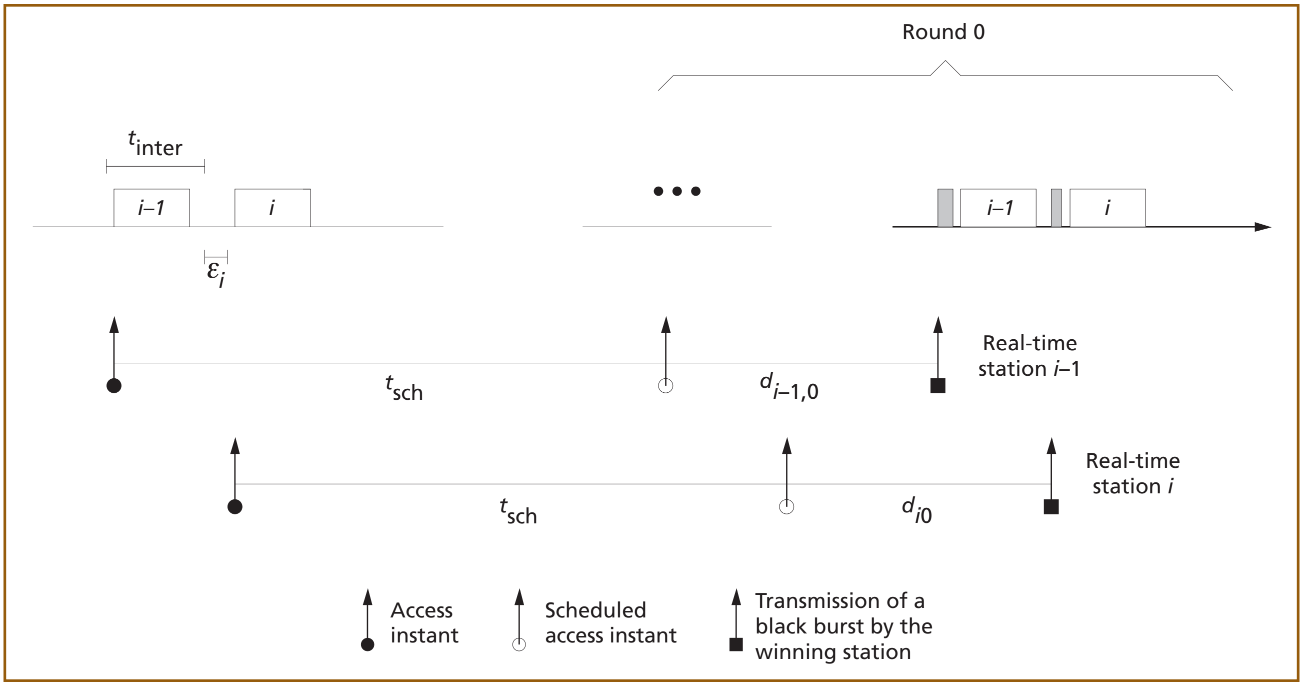

Figure 2.
Determination of $\bar{d}_0$.

$k = 1, 2, \ldots.$

Similarly, from **Figure 3b**, we deduce that the access delays of consecutive stations in round k are related by

$$d_{ik} = \left[(1+\alpha)\, d_{i-1,k} - \alpha d_{i,k-1}\right]_+$$

$$= \left[(1+\alpha)^i d_{0k} - \alpha \sum_{j=1}^{i} (1+\alpha)^{i-j} d_{j,k-1}\right]_+ , \qquad (6)$$

$i = 1, \ldots, n-1, \quad k = 1, 2, \ldots.$

Substituting (5) into (6) and arranging the result in matrix form yields the statement of the proposition. ■

Proposition 3 *The vector of access delays in round k,* $\bar{d}_k$ *, can be expressed as a function of the initial vector of access delays,* $\bar{d}_0$ *, as*

$$\bar{d}_k = \left[\bar{\bar{C}}^k \bar{d}_0 - \varepsilon \left(\sum_{l=0}^{k-1} \bar{\bar{C}}^l\right) \bar{\xi}\right]_+ , \quad k = 0, 1, \ldots.$$

Proof: The proof is by induction and is omitted. ■

It is then clear that the behavior of the system is governed by the matrix $\bar{\bar{C}}$. We proceed to derive its spectral properties.[11]

Proposition 4 *The characteristic polynomial of the matrix* $\bar{\bar{C}}$, $p(x)$ *, is given by*

$$p(x) = \frac{(x+\alpha)^n - (1+\alpha)^n x^{n-1}}{x-1} .$$

Furthermore, for each eigenvalue λ, there are associated right and left eigenvectors, $\bar{v}$ *and* $\bar{w}$ *, with coordinates*

$$v_i = \frac{\alpha}{\lambda}\left(\frac{1+\alpha}{1+\alpha/\lambda}\right)^i \quad \textit{and}$$

$$w_i = \frac{\alpha - v_{n-i}}{\lambda - 1}, \quad i = 1, \ldots, n-1,$$

respectively. The eigenvectors $\bar{v}$ *and* $\bar{w}$ *are such that* $\bar{1}^t \bar{v} = \bar{w}^t \bar{\xi} = 1.$

Proof: Let λ be an eigenvalue of $\bar{\bar{C}}$ and $\bar{v}$, $v = (v_1, \ldots, v_{n-1})^t$, a corresponding right eigenvector, that is, $\bar{\bar{C}}\bar{v} = \lambda \bar{v}$. Using the equation given in Proposition 2 for the matrix $\bar{\bar{C}}$, we can write

$$\left(\frac{\lambda}{\alpha}\bar{\bar{I}} + \bar{\bar{B}}\right)\bar{v} = \left(\sum_{j=1}^{n-1} v_j\right)\bar{\xi} , \qquad (7)$$

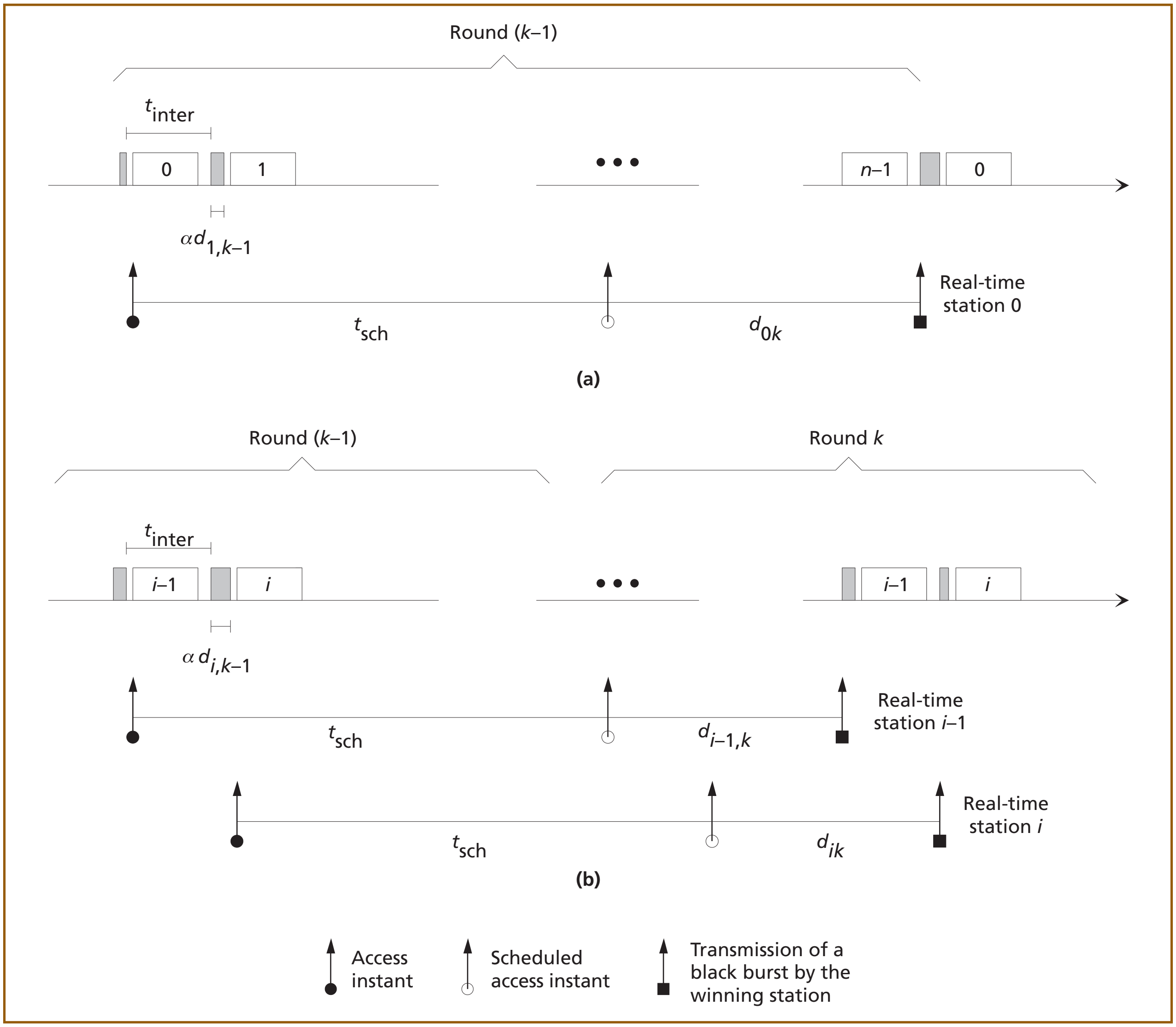

Figure 3.
Determination of $\bar{d}_k$ ***from*** $\bar{d}_{k-1}$.

where $\bar{\bar{I}}$ is the identity matrix. First, we show that $\lambda = -\alpha$ is not an eigenvalue of $\bar{\bar{C}}$. Indeed, if it were, then the first equation in (7) would imply $\sum_{j=1}^{n-1} v_j = 0$. The other equations would successively imply $v_i = 0$, for i = 2,..., n–2 and, therefore, also $v_{n-1} = 0$, contradicting the fact that $\bar{v}$ is an eigenvector. Next, we show that for every eigenvector $\bar{v}$, associated with eigenvalue λ, we have $\sum_{j=1}^{n-1} v_j \neq 0$. Indeed, if $\sum_{j=1}^{n-1} v_j = 0$, then the resolution of the lower triangular system in (7) yields $v_i = 0$, for i = 1,..., n–1, again contradicting the fact that $\bar{v}$ is an eigenvector. In short, given the eigenvalue λ, we can choose the eigenvector $\bar{v}$ to be such that $\sum_{j=1}^{n-1} v_j = 1$. Its coordinates are determined by solving the lower triangular system

$$\left(\frac{\lambda}{\alpha} \bar{\bar{I}} + \bar{\bar{B}} \right) \bar{v} = \bar{\xi}, \qquad (8)$$

the solution of which is

$$v_i = \frac{\alpha}{\lambda} \left(\frac{1+\alpha}{1+\alpha/\lambda} \right)^i, \quad i = 1,\ldots, n-1. \qquad (9)$$

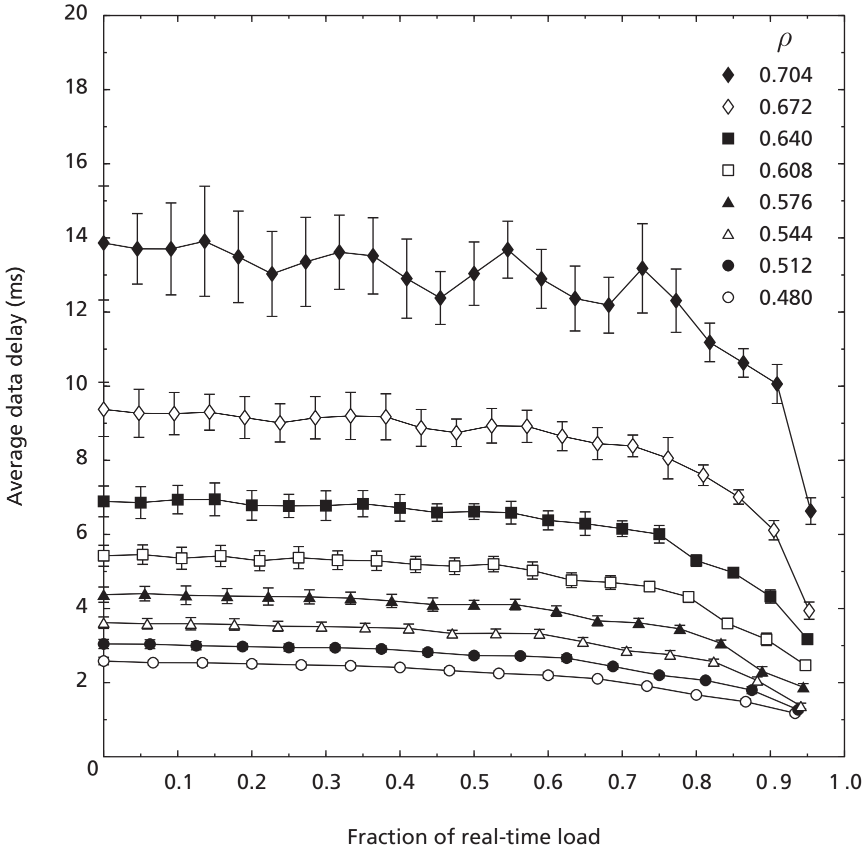

Figure 4.
Average data packet delay as a function of the fraction of real-time load for constant total load.

Now, if indeed λ is an eigenvalue, we have that $\sum_{j=1}^{n-1} v_j = 1$. Conversely, if we can normalize $\sum_{j=1}^{n-1} v_j$ to 1, then λ is an eigenvalue. Consequently, the eigenvalues of $\bar{\bar{C}}$ are the roots of the equation

$$\sum_{j=1}^{n-1} \frac{\alpha}{\lambda}\left(\frac{1+\alpha}{1+\alpha/\lambda}\right)^{j} = 1, \qquad (10)$$

which can be rewritten in the form

$$\frac{(\lambda+\alpha)^n - (1+\alpha)^n \lambda^{n-1}}{\lambda - 1} = 0. \qquad (11)$$

The left-hand side of (11) is a polynomial in λ of degree $(n-1)$, where the coefficient of λ^{n-1} is 1: it is the characteristic polynomial of $\bar{\bar{C}}$. By construction, it also follows that $\bar{1}^t \bar{v} = 1$. A symmetric reasoning is used to derive coordinates for the left eigenvector associated with λ, $\bar{w}$. If $\lambda = 1$ is an eigenvalue, then $w_i = i/[n(n-1)], i = 1, \ldots, n-1$, and that can only occur for $\alpha(n-1) = 1$. ■

Proposition 5 The following properties hold true:

1. *The matrix $\bar{\bar{C}}$ has a unique positive eigenvalue, denoted by λ_1.*
2. *If $\alpha(n-1) < 1$, then $\lambda_1 < \alpha(n-1)$. Else, if $\alpha(n-1) = 1$, then $\lambda_1 = 1$. Else, if $\alpha(n-1) > 1$, then $\lambda_1 > \alpha(n-1)$.*
3. *If λ is an eigenvalue of $\bar{\bar{C}}$ distinct from λ_1, then $|\lambda| < \min(1, \lambda_1)$.*
4. *All eigenvalues of $\bar{\bar{C}}$ are simple roots of the characteristic polynomial.*

Proof: 1. We have that $p(0) = -\alpha^n$ and $\lim_{x \to +\infty} p(x) = +\infty$. Therefore, there is at least one positive eigenvalue. Let λ_1 denote the smallest positive eigenvalue of $\bar{\bar{C}}$. The characteristic polynomial of $\bar{\bar{C}}$ can be written directly in terms of powers of x as

$$p(x) = x^{n-1} - \sum_{j=0}^{n-2} x^{j} \left[\sum_{l=0}^{j} \binom{n}{l} \alpha^{n-l} \right]. \qquad (12)$$

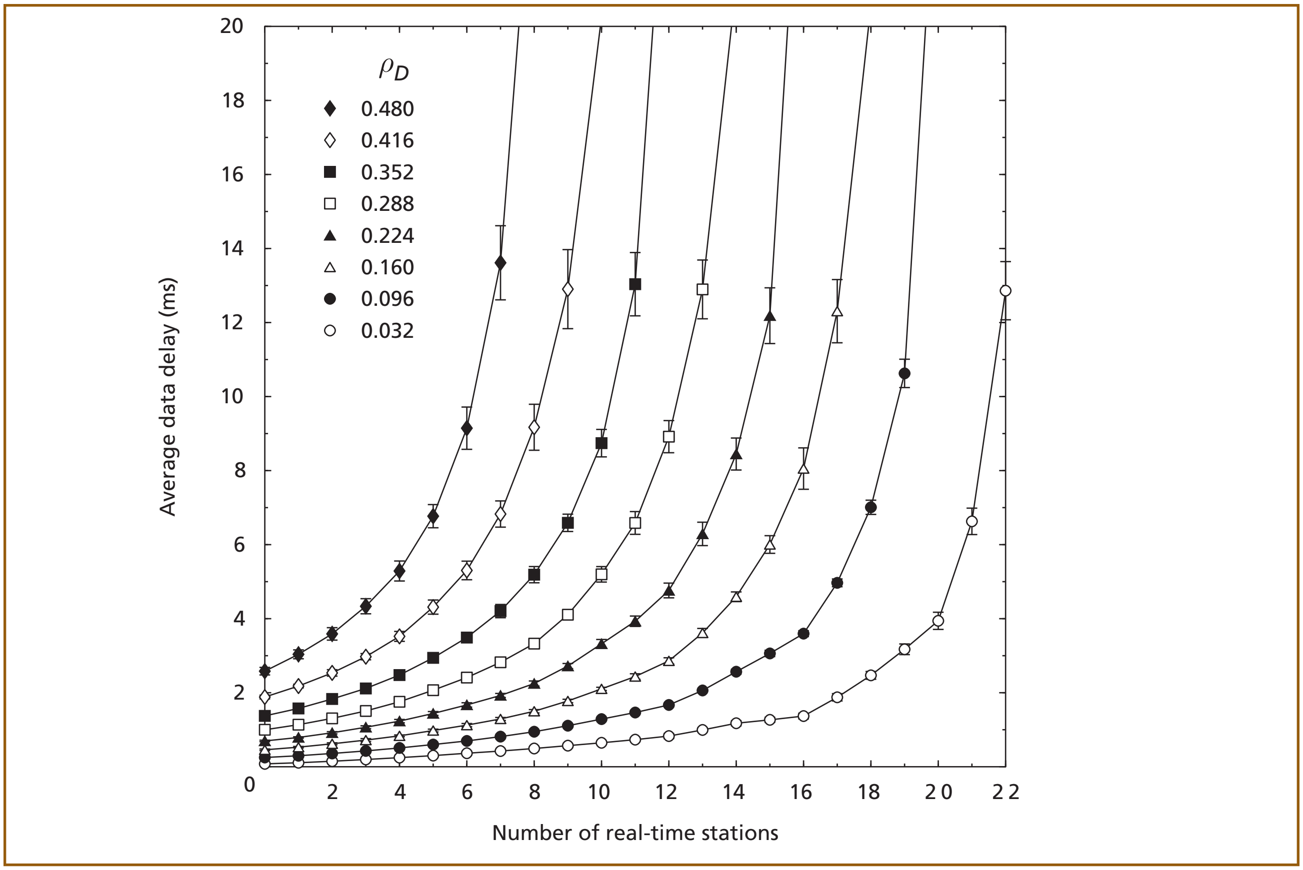

Figure 5.
Average data packet delay as a function of the number of real-time stations for fixed data loads.

Take x to be positive and $x > \lambda_1$:

$$p(x) = x^{n-1}\left\{1 - \sum_{j=0}^{n-2} \frac{1}{x^{n-j-1}}\left[\sum_{l=0}^{j} \binom{n}{l}\alpha^{n-l}\right]\right\}$$

$$> x^{n-1}\left\{1 - \sum_{j=0}^{n-2} \frac{1}{\lambda_1^{n-j-1}}\left[\sum_{l=0}^{j} \binom{n}{l}\alpha^{n-l}\right]\right\} \quad (13)$$

$$= \left(\frac{x}{\lambda_1}\right)^{n-1} p(\lambda_1) = 0 .$$

Consequently, $p(x) > 0$ for $x > \lambda_1$, which implies that λ_1 is the only positive eigenvalue of $\bar{\bar{C}}$.

2. Using l'Hôpital's rule, we have that $p(1) = (1+\alpha)^{n-1}\left[1 - \alpha(n-1)\right]$. Therefore, if $\alpha(n–1) = 1$, then $\lambda_1 = 1$. Now, define

$$q(\alpha) \triangleq \frac{p(\alpha(n-1);\alpha)\left[\alpha(n-1)-1\right]}{\alpha^{n-1}}$$

$$= \alpha n^n - (1+\alpha)^n (n-1)^{n-1}, \quad (14)$$

where $p(x;\alpha)$ makes explicit the dependence of the characteristic polynomial on the parameter α. The function $q(\alpha)$ assumes its maximum value of 0 for $\alpha = 1/(n–1)$: $q(\alpha) < 0$ for $\alpha > 0$ and $\alpha \neq 1/(n–1)$. If $\alpha(n–1) < 1$, then $q(\alpha) < 0$ implies $p(\alpha(n–1);\alpha) > 0$, and hence $\lambda_1 < \alpha(n–1)$. If $\alpha(n–1) > 1$, then $p(\alpha(n–1);\alpha) < 0$, and hence $\lambda_1 > \alpha(n–1)$.

3. If λ, $\lambda \neq \lambda_1$, is an eigenvalue of $\bar{\bar{C}}$, then the Perron-Frobenius theorem[12] asserts that $|\lambda| < \lambda_1$; the same result could be proven, in our specific case, pursuing the reasoning that led to *1* for x complex, $x \neq \lambda_1$ and $|x| \geq \lambda_1$. We would arrive at the conclusion that $|p(x)| > 0$, implying that no eigenvalue of $\bar{\bar{C}}$ can have modulus greater than or equal to λ_1. Now, let λ, $\lambda \neq \lambda_1$, be an eigenvalue and assume that $|\lambda| \geq 1$. We have

$$(|\lambda| + \alpha)^n > |\lambda + \alpha|^n = (1+\alpha)^n |\lambda|^{n-1}, \quad (15)$$

where the strict inequality follows from the fact that λ cannot be positive. With $|\lambda| = 1$, we get

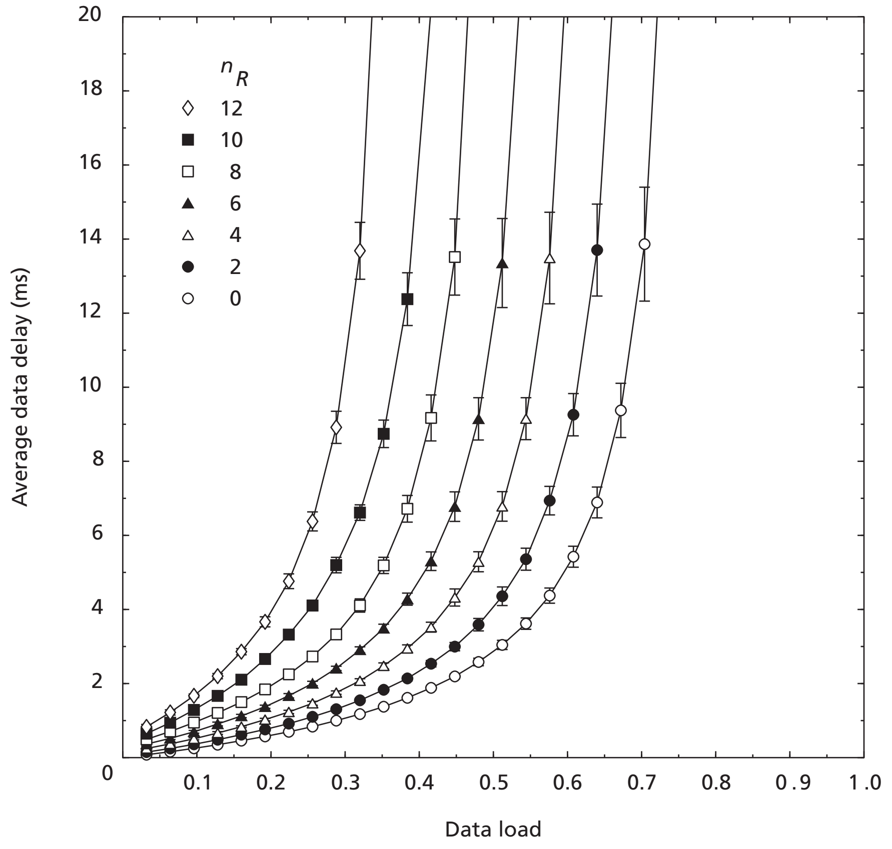

Figure 6.
Average data packet delay as a function of the data load for fixed numbers of real-time stations.

$(1+\alpha)^n > (1+\alpha)^n$, which is false. Therefore, we must have $|\lambda| > 1$. From (15), we can write

$$\frac{\left(|\lambda|+\alpha\right)^n-(1+\alpha)^n|\lambda|^{n-1}}{|\lambda|-1}=p\left(|\lambda|\right)>0. \qquad (16)$$

This implies $|\lambda| > \lambda_1$, which is in contradiction with the Perron-Frobenius theorem. Therefore, we must have $|\lambda| < 1$, thus concluding the proof.

4. The result is proven by showing that no root of $p(x)$ is a root of $p'(x)$. ■

Proposition 4 allows us to obtain the full spectral decomposition of matrix $\bar{\bar{C}}$ in closed form. For our purposes here, it suffices to decompose matrix $\bar{\bar{C}}$ around the principal idempotent matrix corresponding to eigenvalue λ_1, $\bar{\bar{E}}_1$:

$$\bar{\bar{C}}=\lambda_1\bar{\bar{E}}_1+\bar{\bar{C}}_1, \qquad (17)$$

with

$$\bar{\bar{E}}_1=a_1\bar{v}_1\bar{w}_1^t,$$

and $a_1=\left(\bar{v}_1^t\bar{w}_1\right)^{-1}=(\lambda_1/\alpha)\times(\lambda_1-1)/\left[\lambda_1-\alpha(n-1)\right]$ for $\alpha(n-1)\neq 1$ and $a_1=\left(\bar{v}_1^t\bar{w}_1\right)^{-1}=2\,(n-1)$ for $\alpha(n-1)=1$. The eigenvalues of the matrix $\bar{\bar{C}}_1$ are those of $\bar{\bar{C}}$, with the exception of λ_1 and the addition of the eigenvalue 0. In particular, those eigenvalues have all modulus less than one. Matrices $\bar{\bar{E}}_1$ and $\bar{\bar{C}}_1$ are orthogonal, and thus

$$\bar{\bar{C}}^k=\lambda_1^k\bar{\bar{E}}_1+\bar{\bar{C}}_1^k,\quad k=1,2,\ldots. \qquad (18)$$

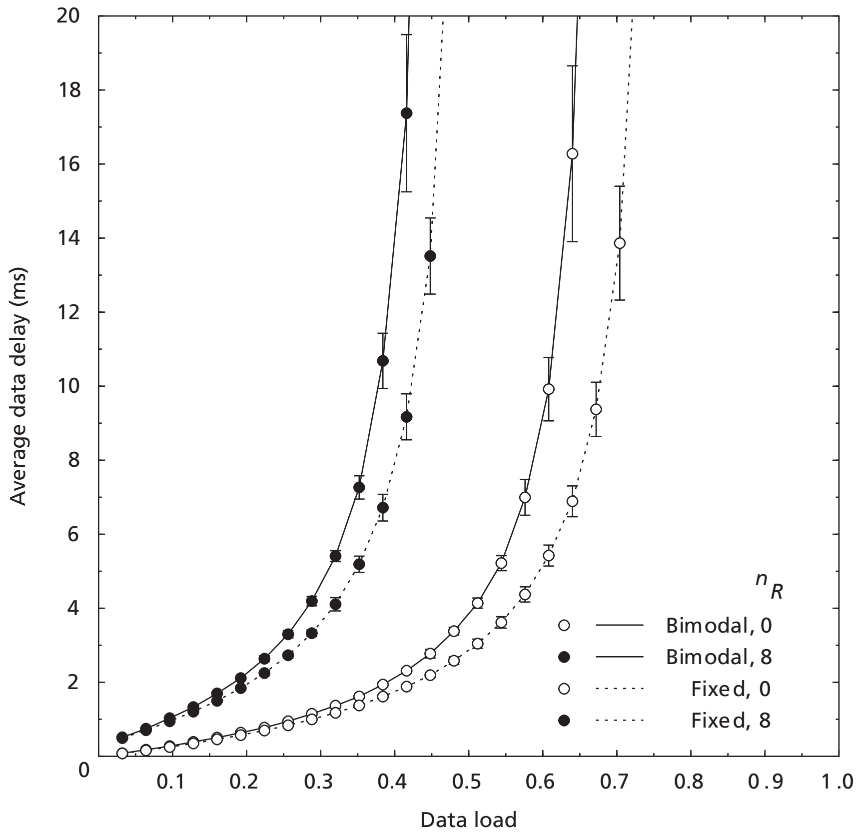

Figure 7.
Comparison of 8,000-bit fixed-length data packets and data packets obtained from a bimodal distribution, one half containing 960 bits and the other half containing 9,600 bits.

Proposition 6 *The vector of access delays in round k, $\bar{d}_k$, can be written as:*

$$\bar{d}_k = \left[a_1 \lambda_1^k \left(\bar{w}_1^t \bar{d}_0 - \frac{\varepsilon}{\lambda_1 - 1} \right) \bar{v}_1 + \bar{\bar{C}}_1^k \bar{d}_0 + \frac{\varepsilon}{\alpha(n-1)-1} \left(\bar{\bar{I}} - \bar{\bar{C}}_1^k \right) \bar{1} \right]_+ ,$$

for $\alpha(n-1) \neq 1$, and

$$\bar{d}_k = \left[2\left(\bar{w}_1^t \bar{d}_0 - \varepsilon(k-1) \right) \bar{1} + \bar{\bar{C}}_1^k \bar{d}_0 - 2\varepsilon \left(\bar{\bar{I}} - \bar{\bar{C}}_1^k \right) \left((n-1)\, \bar{w}_1 + \frac{n+1}{3n} \bar{1} \right) \right]_+$$

for $\alpha(n-1) = 1$.

Proof: Proposition 3 is our starting point. If $\alpha(n-1) \neq 1$, then we have

$$\sum_{l=0}^{k-1} \bar{\bar{C}}^l = (\bar{\bar{I}} - \bar{\bar{C}}^k)\, (\bar{\bar{I}} - \bar{\bar{C}})^{-1}. \qquad (19)$$

Using the equation for matrix $\bar{\bar{C}}$ given in Proposition 2, we write

$$\bar{\bar{C}}\,\bar{1} = \alpha(n-1)\bar{\xi} - \alpha\bar{\bar{B}}\,\bar{1} = \left[\alpha(n-1) - 1\right]\bar{\xi} + \bar{1}, \qquad (20)$$

implying that

$$\left(\bar{\bar{I}} - \bar{\bar{C}} \right)^{-1} \bar{\xi} = \frac{1}{1-\alpha(n-1)} \bar{1}. \qquad (21)$$

Also,

$$\bar{w}_1^t \left(\bar{\bar{I}} - \bar{\bar{C}} \right)^{-1} \bar{\xi} = \frac{1}{1-\alpha(n-1)} \bar{w}_1^t \bar{1} = \frac{1}{1-\lambda_1}, \qquad (22)$$

and the result of the proposition for the case $\alpha(n-1) \neq 1$ follows from the decomposition of $\bar{\bar{C}}$, (17), and straightforward manipulation of the results above. The case $\alpha(n-1) = 1$ has a similar treatment, but the details are omitted. ■

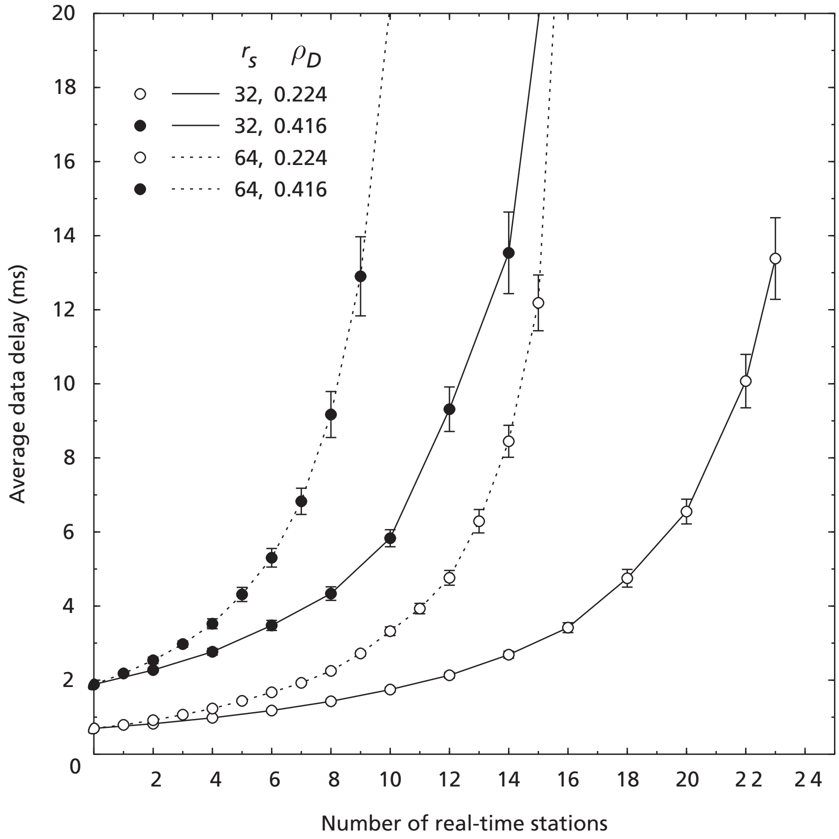

Figure 8.
Comparison of 64- and 32-kb/s source rates.

We can now easily derive stability conditions for the operation of the system.

Proposition 7 *The system is unconditionally stable if and only if*

$$\alpha \le \frac{1}{n-1}.$$

In addition, if $\alpha > 1/(n-1)$*, the system is stable if and only if*

$$T < \frac{\varepsilon}{\lambda_1 - 1}.$$

Proof: The proof follows directly from the expressions in Proposition 6, together with the fact that $\lim_{k\to+\infty} \bar{\bar{C}}_1^k = \bar{\bar{0}}$, where $\bar{\bar{0}}$ is the null matrix. In the case that $\alpha(n-1) > 1$, it follows from Proposition 6 that the vector of access delays will converge to zero if and only if

$$\bar{w}_1^t \bar{d}_0 - \frac{\varepsilon}{\lambda_1 - 1} < 0. \qquad (23)$$

From Proposition 1 we can conclude that, whatever the initial vector $\bar{\varepsilon}$, we always have $\bar{d}_0 \le T\bar{\xi}$, and thus $\bar{w}_1^t \bar{d}_0 \le T$; the equality is satisfied for $\bar{\varepsilon} = \bar{0}$. Therefore, if $T < \varepsilon/(\lambda_1-1)$, the system is stable. On the other hand, if $T \ge \varepsilon/(\lambda_1-1)$, the initial conditions $\bar{\varepsilon} = \bar{0}$ prevent $\bar{d}_0$ from satisfying (23), and the system is unstable. ■

Let n_{fit} and n_{stab} denote the maximum number of real-time stations that satisfy (1) and the criterion of unconditional stability derived in Proposition 7, respectively. The maximum number of real-time stations that can be accommodated in an unconditionally stable system, n_{max}, is thus

$$n_{max} = \min(n_{fit}, n_{stab}). \qquad (24)$$

Results and Discussion

The previous section showed that once the maximum length of the data packets is known, we can determine stability conditions for the operation of the

Table I. Nominal values for the parameters of the system.

Parameter	Symbol	Nominal value
Channel bit rate	r_c	2,000 kb/s
Long spacing	t_{long}	40 μs
Medium spacing	t_{med}	20 μs
Observation interval	t_{obs}	16 μs
Duration of white slots	t_{wslot}	20 μs
Duration of black slots	t_{bslot}	20 μs
Overhead bits	h	200 bits
Interaccess interval	t_{acc}	21 ms
Source rate	r_s	64 kb/s
Maximum delay	w_{max}	25 ms
Number of data stations	n_D	10
Initial backoff window	f_{data} (0)	16
Bits per data packet	b_{data}	8,000 bits

Table II. Maximum number of real-time stations in an unconditionally stable system.

	r_s = 64 kb/s			r_s = 32 kb/s		
w_{max} (ms)	n_{fit}	n_{stab}	n_{max}	n_{fit}	n_{stab}	n_{max}
15	**17**	31	17	29	**19**	19
25	**22**	47	22	39	**27**	27
35	**24**	63	24	44	**35**	35

Table III. Stability conditions for α > 1/(n–1), with r_s = 32 kb/s and w_{max} = 25 ms.

Number of real-time stations	Stable if T is less than
28	394.0 ms
30	28.4 ms
32	11.1 ms
34	5.2 ms
36	**2.3 ms**

real-time access procedures. If those procedures are stable, then real-time stations have bounded access delays. This is a consequence of the collision-free character of their contention mechanism. The next section of this paper presents stability results for the operation of the real-time access procedures.

Another important aspect of network performance is the impact of real-time access procedures on the average delay of data packets. The section "Average Delay of Data Packets" assesses the impact through simulation.

The channel bit rate is denoted by r_c , and each packet transmission from either a real-time or a data station has an overhead of h bits. **Table I** gives nominal values for the system parameters.

Stability Results for Real-Time Traffic

We assume that real-time stations generate bits at the constant rate of r_s. Further, a real-time station discards information bits that have been delayed for transmission by more than w_{max} seconds, with $w_{max} > t_{acc}$. The real-time packets are of fixed length, and they convey a maximum of $r_s w_{max}$ source bits. Therefore,

$$t_{pkt} = \frac{h + r_s w_{max}}{r_c}, \tag{25}$$

and we always take $t_{unit} = t_{inter}$.

In **Table II**, we present values of n_{fit}, n_{stab}, and n_{max} for $(w_{max} - t_{acc}) = 4$ ms. We observe that for r_s = 64 kb/s, the maximum number of real-time stations is constrained by n_{fit}, whereas for r_s = 32 kb/s, the maximum number of real-time stations is constrained by n_{stab}. Indeed, for a given value of w_{max}, the length of the real-time packets, t_{pkt}, decreases as the source rate, r_s, decreases. This implies that t_{inter} also decreases and that α increases proportionally.

As noted, we have that $n_{stab} < n_{fit}$, for r_s = 32 kb/s. Consequently, for any number of real-time stations, n, satisfying $n_{fit} \geq n > n_{stab}$, the stability properties of the system depend on the magnitude of the initial perturbation. **Table III** presents the values of T below which the system remains stable. Those values should be contrasted with the transmission time of a data packet—4.1 ms for a data packet of 8,000 bits, with 200 bits of overhead.

Average Delay of Data Packets

We keep the assumptions of the previous section, but we denote the number of real-time stations in the system by n_R and allow real-time packets to be of variable length. No padding bits are used; just the information bits that have been generated between the last access instant and the current one are conveyed, up to a maximum of $(h + r_s w_{max})$ bits. In addition, there are

n_D data stations in the wireless LAN, and data packets arrive at each data station according to a Poisson process with parameter λ_D / n_D packets per second. Initially, the number of data bits per data packet is a constant, denoted by b_{data}. The data load, ρ_D, and the real-time load, ρ_R, are defined as $\rho_D \triangleq \lambda_D b_{\text{data}} / r_c$ and $\rho_R \triangleq n_R r_s / r_c$, respectively, and their sum is the total load, ρ. The performance results of this section were obtained by simulation. Each simulation run corresponds to 4 to 7 minutes of real-time operation of the protocol. Each point was obtained by running 10 independent replicas of the system and averaging the results. The error bars in the plots mark the 95% confidence intervals.

Figure 4 shows the average data packet delay as a function of the fraction of real-time load, $\rho_R / (\rho_R + \rho_D)$, for various values of the total load, ρ. It is interesting to observe that the average data packet delay remains approximately constant, or even decreases slightly, as we trade data load for real-time load. One could expect the data delays to increase with the fraction of real-time load, because the latter traffic has priority over data. However, the data access procedures anticipate collisions and, as a consequence, use the channel less efficiently than the collision-free access procedures of real-time traffic. As we increase the real-time load and decrease the data load, a larger volume of traffic gets priority over data. That traffic, however, is efficiently served, causing essentially no alteration on the average data packet delay. This result reassures us that the priority given to real-time traffic does not significantly affect the average data packet delay, provided that the total load remains constant.

In **Figure 5**, we plot the average data packet delay for constant data load as a function of the number of real-time stations. We see, for example, that if the maximum tolerable delay for data packets is 10 ms, then we can support 10 real-time stations at a data load of 0.352 and 14 real-time stations at a data load of 0.224. **Figure 6** shows how the throughput-delay curves of data traffic are affected by different numbers of real-time stations. As an example, we have that if the maximum allowed data delay is again 10 ms, we can support data loads of 0.416 and 0.544, for a system with 8 and 4 real-time stations, respectively.

Figure 7 compares the data packet delays between the cases of fixed-length data packets and data packets obtained from a bimodal distribution. For the bimodal distribution case, we have assumed that half the data packets have 960 bits and the other half have 9,600 bits. Clearly, the delay performance is worse when data packets have a bimodal distribution. The reasons for this are twofold. First, the average number of bits per data packet is lower in the bimodal distribution case, while both the transmission and the multiple access overheads remain constant. Second, whenever a collision occurs, the time during which the channel is unusable depends on the maximum, rather than on the average, length of the packets involved in that collision. We also observed that, at a level of about 1%, we could have clipping of real-time traffic when the data packets follow a bimodal distribution, as described. This happens because the transmission time of a long data packet, 4.9 ms, exceeds the elasticity of real-time packets, given by $(w_{\max} - t_{\text{acc}}) = 4$ ms, and so some data packets may force the real-time stations to discard their oldest information bits.

Finally, **Figure 8** shows the increase in the number of real-time stations that can be supported by the system when the source rate is decreased from 64 kb/s to 32 kb/s. For instance, at the 10-ms level of average data packet delay, the number of real-time stations that can be accommodated in the system increases from 8 to 12 at a data load of 0.416, and from 14 to 21 at a data load of 0.224. The number of real-time stations is never doubled when the source rate is decreased by half, because the transmission overhead remains constant and t_{inter} gets smaller, implying longer black bursts for the same access delays of real-time stations.

Conclusions

We have described distributed multiple access procedures that give real-time stations priority access to CSMA/CA-based wireless LANs. In the proposed procedures, real-time stations contend for access to the channel with black bursts and transmit packets with variable numbers of source bits. The proposed scheme has several important characteristics. It can be overlaid on an IEEE 802.11 standard implementation, without changing the access procedures of data stations. Real-time stations access the channel in a round robin order. Contention among real-time stations is minimized, because each black burst contention period produces a unique winner. Real-time stations have nonpreemptive priority over data stations, and the proposed scheme can support real-time stations with distinct bandwidth requirements. This scheme can also operate with negative acknowledgments, thus reducing overhead and providing robustness against hidden stations.

We have determined conditions under which the proposed access procedures for real-time stations are stable, thus implying bounded access delays. To assess the impact of real-time traffic on the average delays of data packets, we have also undertaken a simulation study. This study shows that a wireless LAN operating at 2 Mb/s can accommodate about 14 stations at a real-time source rate of 64 kb/s, for a data utilization of 0.224 with a maximum accepted data delay of 10 ms. The number of real-time stations is increased to 21 when the source rate is decreased to 32 kb/s.

References

1. P802.11, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, Draft Standard, IEEE 802.11, July 1995.
2. J. D. DeTreville, "A Simulation-Based Comparison of Voice Transmission on CSMA/CD Networks and on Token Bus," *AT&T Bell Labs Technical Journal*, Vol. 63, No. 1, January 1984, pp. 33-55.
3. E. Friedman and C. Ziegler, "Packet Voice Communications over PC-Based Local Area Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 7, No. 2, February 1989, pp. 211-218.
4. T. A. Gonsalves and F. A. Tobagi, "Comparative Performance of Voice/Data Local Area Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 7, No. 5, June 1989, pp. 657-668.
5. N. F. Maxemchuk, "A Variation on CSMA/CD That Yields Movable TDM Slots in Integrated Voice/Data Local Networks," *Bell System Technical Journal*, Vol. 61, No. 7, January 1982, pp. 1527-1550.
6. G. J. Nutt and D. L. Bayer, "Performance of CSMA/CD Networks Under Combined Voice and Data Slots," *IEEE Transactions on Communications*, Vol. COM-30, No. 1, January 1982, pp. 6-11.
7. F. Tobagi, "Carrier Sense Multiple Access with Message Based Priority Functions," *IEEE Transactions on Communications*, Vol. COM-30, No. 1, January 1982, pp. 185-200.
8. M. A. Visser and M. El Zarki, "Voice and Data Transmission over an 802.11 Wireless Network," *Proceedings of PIMRC'95*, Toronto, Canada, September 1995, pp. 648-652.
9. N. F. Maxemchuk and A. Netravali, "Voice and Data on a CATV Network," *IEEE Journal on Selected Areas in Communications*, Vol. JSAC-3, No. 2, March 1985, pp. 300-311.
10. F. Tobagi and L. Kleinrock, "Packet Switching in Radio Channels: Part II—The Hidden Terminal Problem in Carrier Sense Multiple Access and the Busy-Tone Solution," *IEEE Transactions on Communications*, Vol. COM-23, No. 12, December 1975, pp. 1417-1433.
11. J. Gilbert and L. Gilbert, *Linear Algebra and Matrix Theory*, Academic Press, San Diego, 1995.
12. H. Minc, *Nonnegative Matrices*, John Wiley & Sons, New York, 1988.

(Manuscript approved November 1996)

JOÃO L. SOBRINHO, who recently left the company, was

a member of technical staff in the Wireless Networking and Mobile Computing Research Department of Bell Labs in Nieuwegein, The Netherlands, when this work was performed. He conducted research on wireless LANs, specifically on integrating real-time and data traffic, and on efficient multicellular operation. Mr. Sobrinho received the Licenciatura and Ph.D. degrees in electrical and computer engineering, both from the Instituto Superior Técnico, Lisbon, Portugal.

A. S. KRISHNAKUMAR is head of the Wireless

Networking and Mobile Computing Research Department at Bell Labs in Nieuwegein, The Netherlands. The first Bell Labs research group located outside the United States, his department is responsible for high-speed wireless data networks and wireless multimedia integration. Mr. Krishnakumar received a B.Tech. degree in electronics from the Indian Institute of Technology, Madras, India, and an M.S. degree in electrical engineering from Northwestern University, Evanston, Illinois. He also earned an M.S. in statistics and a Ph.D. in electrical engineering from Stanford University, Palo Alto, California. ◆