Wireless Networks With Retransmission Diversity Access Mechanisms: Stable Throughput and Delay Properties

Goran Dimić, *Student Member, IEEE*, Nicholas D. Sidiropoulos, *Senior Member, IEEE*, and Leandros Tassiulas

***Abstract*—Building on the concept of retransmission diversity, a class of collision resolution protocols [(B)NDMA] has been introduced recently for wireless packet multiple access. These protocols provide the means for improved performance compared with random access and splitting-based collision resolution protocols at a moderate receiver complexity cost. However, stability of these protocols has not been established, and the available steady-state analysis is restricted to symmetric (common-rate) systems. In this paper, the stability region of (B)NDMA is formally analyzed. The tools used in the analysis range from a preliminary *dominant system* approach to the Foster–Lyapunov recurrence criterion and the (σ, ρ) deterministic fluid arrivals approach. It is rigorously established that maximum stable throughput is close to 1. This is followed by a simpler and more general steady-state analysis, bypassing the earlier generating function approach, using instead only balance equations. This approach allows dealing with asymmetry (multirate systems), yielding expressions for throughput and delay per queue. Finally, we generalize BNDMA and the associated analysis to multicode systems.**

***Index Terms*—Random access, signal processing aspects of network protocols: stability.**

I. Introduction

INCREASED interest in wireless data and multimedia services motivates research in improved random access protocols. These protocols are suitable for multiplexing bursty sources encountered in data transfer [5], [10]. At light traffic conditions, they provide average delay that is significantly smaller than that of fixed allocation schemes like time, frequency, or code-division multiple access (CDMA). However, they have relatively low maximum throughput and suffer from excessive delay under even moderate traffic.

The throughput/delay penalty of random access protocols is due to collisions of data packets. When a collision occurs, the received packets are discarded without recovering any data. New transmissions of the same packets must follow, possibly inducing secondary collisions, so that more slots are used than transmitted packets. Therefore, throughput decreases and delay can become excessive.

In wireline networks, it is possible to overcome this shortcoming of random access techniques through the use of carrier sensing and collision detection, as in wireline Ethernet local area networks. In carrier sense multiple access with collision detection (CSMA/CD), terminals sense the common bus for a carrier before transmitting and then listen for collisions during transmission. If a collision is detected, the transmission is immediately aborted. If propagation delay is small relative to packet duration, CSMA/CD alleviates the impact of collisions. This differentiates CSMA/CD from ALOHA-type access, which assumes that feedback is made available after packet transmission is complete. An alternative way of achieving higher throughput is by means of *collision multiplicity feedback*, wherein the number of collided packets is made available to the transmitting terminals at the end of the packet transmission. This can be exploited to optimize the retransmission probability, but delay performance remains poor at higher loads because collisions are still wasteful.

In wireless networks, it is possible to improve performance in two ways. One is to employ a certain fixed amount of spreading, which enables multipacket reception, but this comes at the price of bandwidth expansion. Another stems from the fact that collided packets are often received with disparate powers; if one of them has much higher power than the rest, then it can be correctly decoded. This is the so-called *capture* effect. Note, however, that one may not rely on capture alone because it is a random event.

A novel approach to the collision resolution (CR) problem has been proposed recently in [15]. The idea behind it is to generate diversity via immediate simultaneous retransmissions of *all* collided packets induced by the medium access control (MAC) layer protocol. The key fact is that if *the same* K packets collide again for a total of K times, then one collects K linear mixtures of the original packets. If these mixtures are linearly independent, then it is possible to recover the original packets by solving the associated linear system. This requires that packets contain certain known prefixes that enable detection and estimation of the mixing matrix.

The protocol in [15] was dubbed *network-assisted diversity multiple access* (NDMA). NDMA can achieve maximum throughput close to 1, while exhibiting low delay over a wide

Manuscript received July 7, 2002; revised April 8, 2003. G. Dimić and N. D. Sidiropoulos were supported by the ARL Communications and Networks CTA and NSF/Wireless IT and Networks CCR-0096164. Preliminary versions of parts of this paper appear in the *Proceedings of the ICASSP 2002* and the *Proceedings of the ISIT 2002*. The associate editor coordinating the review of this paper and approving it for publication was Dr. Athina Petropulu.

G. Dimić is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: goran@ece.umn.edu).

N. D. Sidiropoulos is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA. He is also with the Department of Electronic and Computer Engineering, Technical University of Crete, Crete, Greece (e-mail: nikos@telecom.tuc.gr).

L. Tassiulas is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: leandros@isr.umd.edu).

Digital Object Identifier 10.1109/TSP.2003.814471

range of loads. Maximum throughput is not 1, due to the use of orthogonal terminal ID sequences embedded in the packet header. The details will be clarified shortly, but we remark that it is also possible to solve the set of linear equations *blindly*, provided one more retransmission is requested, and a certain type of packet phase modulation is employed at the transmitters [17]. The protocol in [17] was dubbed blind NDMA (BNDMA). BNDMA retains throughput and delay characteristics similar to NDMA.

In the context of the random access schemes reviewed above, (B)NDMA can be viewed as an alternative means of boosting throughput close to one: Instead of requiring fast or elaborate feedback, bandwidth overexpansion or relying on the random capture effect, it works by exploiting the receiver complexity dimension. It improves throughput by paying the price of a moderately complex receiver [roughly, $O(K^3)$ for a K-fold collision]. Additional benefits include a low delay characteristic and applicability in a wireless environment, wherein CSMA/CD is not an option due to shadowing.

A. NDMA and BNDMA Protocols

Consider a discrete-time, slotted system with one base station (BS) and J users, synchronized to slot timing. Each user stores incoming packets in an infinite-capacity buffer. The average rate of the jth buffer arrival process is λ_j, and arrivals are independent across users. At the beginning of each slot, a user transmits one packet, provided that it is allowed to transmit and its buffer is nonempty.

Listen-while-you-talk is not feasible in the wireless environment. Therefore, the BS detects collisions and provides feedback to the users. It is important to spell out feedback assumptions. As is customary in slotted random access (e.g., slotted ALOHA or tree splitting), we assume 0/1/e feedback that is made available to the users at the beginning of each slot. The timing constraint can be met by time-division duplex (TDD) or time-division multiplexing two protocols. In the noiseless case, e feedback is in fact not necessary for (B)NDMA; we assume a noiseless system[1] to focus on network effects. In our context, 0 clears all terminals for transmission, whereas 1 enables those that transmitted in the previous slot and disables all others. Note that each terminal knows whether it has transmitted or not in the previous slot.

In NDMA, transmission of a packet by the jth user is detected at the BS by using a filter matched to the user's orthogonal ID, which is embedded in the packet header. Therefore, collision multiplicity is estimated as the total number of detected users [15]. In BNDMA, collision multiplicity is estimated at the BS using rank detection [17].

Once the BS detects a collision, it sets feedback to 1. All users who transmitted in the previous slot will retransmit the same packet, whereas all others will wait. Based on the collision multiplicity estimation, the BS decides how many retransmissions of the collided packets are necessary for CR. The slots used for the first transmission and subsequent retransmissions comprise a CR *epoch*.

[1]In practice, this can be approximated using suitable forward error control coding.

After T transmissions (initial collision and $T-1$ retransmissions), the discrete-time baseband-equivalent data model is

$$\boldsymbol{X}_{T\times N} = \boldsymbol{A}_{T\times K}\boldsymbol{S}_{K\times N} + \boldsymbol{W}_{T\times N}. \tag{1}$$

N denotes packet length, K is the number of collided packets, $\boldsymbol{X}$ is the received data matrix, $\boldsymbol{A}$ is the mixing matrix, $\boldsymbol{S}$ is the signal matrix whose rows are collided packets, and $\boldsymbol{W}$ is the white Gaussian noise matrix. For NDMA, it is assumed that the channel between every user and the BS is frequency-flat and block-fading: constant over each slot but different from slot to slot [15]. For BNDMA, the channel is assumed constant over each CR epoch. Frequency selectivity can be easily accommodated in both protocols with the inclusion of some slot guard time; no other modifications are needed.

In NDMA, the mixing matrix $\boldsymbol{A}$ is estimated using the known user IDs, and then, $\boldsymbol{S}$ is recovered. In BNDMA, the mixing matrix has Vandermonde structure. This is obtained by the following retransmission scheme.

- Before the first retransmission, each user randomly draws a digital carrier for the packet ω_i (for the ith user).
- In the rth retransmission, the ith user's carrier is multiplied by r, and the whole packet is multiplied by $e^{jr\omega_i}$.

Random selection of digital carriers ensures that the Vandermonde mixing matrix has full rank with probability 1. This allows use of an ESPRIT-like method for blind packet recovery [17].

For a K-fold collision, NDMA requires $K-1$ retransmissions [15], whereas BNDMA requires K retransmissions [17]. Note that NDMA and BNDMA *deterministically* achieve lower CR delay than any tree-splitting/first-come first-serve (FCFS) protocol for slotted ALOHA [5], including the dynamic tree algorithm [6], which requires online rate estimation and adaptation.

We are interested in establishing stability of NDMA and BNDMA for a finite user population and buffered packets. Stability analysis is complicated because the queues are coupled, yielding a nonseparable multidimensional Markov chain. This difficulty also arises in the stability analysis of buffered slotted ALOHA [2], [12]–[14], [16], wherein a single necessary and sufficient stability condition is missing for $J > 3$. However, initial progress can be made by employing the *dominant system* approach, which was originally developed for slotted ALOHA [12]–[14], [16]. This is pursued in Section II. Then, in Sections III and IV, tight sufficient conditions for stability of NDMA and BNDMA, respectively, are established.

After establishing stability, in Section V we turn to steady-state analysis. This lays the foundation for estimating average delay, on a per-queue basis, in terms of average arrival rates of all users.

With the insight gained from stability and steady-state analyzes of BNDMA, a generalized BNDMA scheme with improved performance is proposed and analyzed in Section VI. Section VII presents a unified delay analysis of the xNDMA protocols. All analytic results are compared with simulations, which are described in Section VIII.

II. Preliminary Stability Analysis via Dominant System Approach

Let $\mathbf{s}(t) := [s_1(t), \ldots, s_J(t)]^T$ be the vector of queue lengths. Assuming Poisson arrivals, and given that the CR epoch is deterministically bounded (by J, $J+1$ for NDMA and BNDMA, respectively), it suffices to study the embedded Markov chain with transition times set at the beginnings of epochs. Let $\mathbf{s}(k) := [s_1(k), \ldots, s_J(k)]^T$ denote the vector of queue lengths at the beginning of the kth epoch. To see that this is indeed a Markov chain, note that [15], [17]

$$s_j(k+1) = \begin{cases} s_j(k) - 1 + n_j(k), & s_j(k) > 0 \\ n_j(k), & s_j(k) = 0 \end{cases} \tag{2}$$

for $j = 1, 2, \ldots, J$, where $n_j(k)$ is the number of new arrivals into queue j during the kth epoch. $n_j(k)$ is a random variable, with mean

NDMA:	$\lambda_j \left[\sum_{i=1}^{J} 1\{s_i(k) > 0\} + \delta \left(\sum_{i=1}^{J} s_i(k) \right) \right]$
BNDMA:	$\lambda_j \left[\sum_{i=1}^{J} 1\{s_i(k) > 0\} + 1 \right]$

Here, $\delta(x)$ is the Kronecker delta function, and $1\{\}$ is the indicator function. We adopt the following definition of stability:

Definition 1 (e.g., [12]): Queue j of the system is *stable* if

$$\lim_{k\to\infty} \Pr\{s_j(k) < x\} = F(x) \quad \text{and} \quad \lim_{x\to\infty} F(x) = 1 \tag{3}$$

$$\text{If } \lim_{x\to\infty} \liminf_{k\to\infty} \Pr\{s_j(k) < x\} = 1 \tag{4}$$

the queue is *substable*. A stable queue is also substable. If a queue is not substable, it is unstable. The system is stable if all the queues are stable. If at least one queue is unstable, the system is unstable.

In (3), $\inf$ stands for the greatest lower bound.

Proposition 1: The vector process $\mathbf{s}(k)$ is a homogeneous, irreducible, and aperiodic Markov chain with countable number of states.

The proof is straightforward (see [9]).

The definition of stability in (3) is equivalent to positive recurrence of the associated embedded Markov chain. In other words, the system is stable if and only if there is a positive probability mass function of $\mathbf{s}(k)$ when k tends to infinity. Substability as defined in (4) is equivalent to positive recurrence of the embedded Markov chain at the boundary of stability—when the average arrival rate is equal to the average service rate. Then, different initial conditions $\mathbf{s}(0)$ may yield different positive probability mass functions of $\mathbf{s}(k)$ when k tends to infinity. Hence, it is the worst case that decides whether the queue is substable or unstable. One can find more detailed explanations in [11].

A. Preliminary Conditions for Stability

Consider a dominant system in which every one of the J queues always transmits one packet at the beginning of a CR epoch, even if it has none in its queue, in which case, it transmits a dummy packet. Since this action increases the CR (and hence service) time for all queues without affecting arrivals, a queue in the dominant system always has at least as many buffered packets as it would have in the original system, *on a realization-by-realization basis*, provided both begin from the same initial state. It is said that the queues in the dominant system dominate the queues in the original system. Similar to the slotted ALOHA case [12], by virtue of Proposition 1, the original system is stable if and only if $\lim_{k\to\infty} \Pr\{s_j(k) = 0\} > 0, \forall j$. Note that this is equivalent to existence of a positive probability mass function, and therefore, it is also equivalent to definition of stability (3). Let the superscript O (D) denote the original (dominant) system. If the dominant system is stable, then $\lim_{k\to\infty} \Pr\{s_j^D(k) = 0\} > 0, \forall j$. Since

$$\lim_{k\to\infty} \Pr\{s_j^O(k) = 0\} > \lim_{k\to\infty} \Pr\{s_j^D(k) = 0\}$$

it follows that the original system is also stable.

Assume that the arrival process is Poisson, and consider any particular queue in the dominant system. This queue is equivalent to a slotted M/D/1 queue with service time J slots, for NDMA, or $J+1$ slots, for BNDMA. Note that the queues in the dominant system are decoupled. Loynes theorem states that if the arrival process and service process of a queue are stationary, and the average arrival rate is less than the average service rate, then the queue is stable; if the average arrival rate is greater than the average service rate, the queue is unstable; if they are equal, the queue can be either stable or substable [11]. Stationarity of arrivals is given, whereas service is deterministic in the dominant system, hence trivially stationary (note that this is not obvious in the original system). Therefore, a sufficient condition for stability is

$$\begin{aligned} &\text{NDMA: } \lambda_j < \frac{1}{J} \\ &\qquad\qquad\qquad \text{for } j = 1, 2, \ldots, J. \\ &\text{BNDMA: } \lambda_j < \frac{1}{J+1}. \end{aligned} \tag{5}$$

III. Stability of NDMA via Foster–Lyapunov Approach

A relaxed condition on the arrival rates that guarantees stability of an NDMA system can be obtained by using the Foster-Lyapunov approach [3].

Note that for the transmission of K collided packets, K slots are needed. After these packets are transmitted, another contention resolution period can start. Therefore, the system should be stable if on average less than one new packet arrives in the system during one slot. This intuition is confirmed in the following Theorem.

Theorem 1: The NDMA system with Poisson arrivals is stable if[2]

$$\sum_{j=1}^{J} \lambda_j < 1. \tag{6}$$

Proof: It has been shown that $\mathbf{s}(k)$ is an irreducible Markov chain. We will show that $\mathbf{s}(k)$ is ergodic under the above condition, by using Foster's criterion for ergodicity of a Markov chain (e.g., [3], reproduced here for convenience):

[2]If $\sum_{j=1}^{J} \lambda_j > 1$, then the system is clearly unstable.

Suppose that the chain is irreducible, and let $\boldsymbol{S}_0$ be a finite subset of the state space $\boldsymbol{S}$. Then, the chain is positive recurrent if for some V: $\boldsymbol{S} \to \mathbb{R}$ and some $\epsilon > 0$, we have $\inf_{\boldsymbol{q}} V(\boldsymbol{q}) > -\infty$ and

$$\sum_{\boldsymbol{w}\in\boldsymbol{S}} p_{\boldsymbol{qw}} V(\boldsymbol{w}) < \infty, \qquad \forall\, \boldsymbol{q} \in \boldsymbol{S}_0$$
$$\sum_{\boldsymbol{w}\in\boldsymbol{S}} p_{\boldsymbol{qw}} V(\boldsymbol{w}) \le V(\boldsymbol{q}) - \epsilon, \qquad \forall\, \boldsymbol{q} \notin \boldsymbol{S}_0$$

where $p_{\boldsymbol{qw}}$ is the probability of transition from state $\boldsymbol{q}$ to state $\boldsymbol{w}$.

If a chain is irreducible and aperiodic positive recurrent, then it is ergodic (e.g., [3]), which implies (3).

Note that $\sum_{\boldsymbol{w}\in\boldsymbol{S}} p_{\boldsymbol{qw}} V(\boldsymbol{w}) = E[V(\boldsymbol{s}(k+1)) \mid \boldsymbol{s}(k) = \boldsymbol{q}]$. In addition, $\sum_{\boldsymbol{w}\in\boldsymbol{S}} p_{\boldsymbol{qw}} V(\boldsymbol{w}) - V(\boldsymbol{q}) = E[V(\boldsymbol{s}(k+1)) - V(\boldsymbol{s}(k)) \mid \boldsymbol{s}(k) = \boldsymbol{q}]$. The right-hand side of the last equation resembles the notion of *drift* of a one-dimensional discrete Markov chain. Then, roughly speaking, one may think of Foster's criterion as a generalization of drift analysis: If for all large enough states (states out of a finite subset $\boldsymbol{S}_0$) drift is negative (c.f. the last condition), then the size of queues decrease so that the chain is stable.

Consider the function $V(\boldsymbol{s}) = \sum_{j=1}^{J} s_j$, which is defined on the state space $\boldsymbol{S} = \mathbb{Z}_+^J$ of the Markov chain, where $\mathbb{Z}_+$ denotes the non-negative integers. For all $\mathbf{s} \in \mathbf{S}$, we have

$$V(\mathbf{s}(k)) \ge 0 \tag{7}$$

and

$$\begin{aligned} &E[V(\mathbf{s}(k+1)) \mid \mathbf{s}(k)] \\ &\quad = E\left[\sum_{j=1}^{J} s_j(k+1) \,\middle|\, \mathbf{s}(k)\right] \\ &\quad = E\left[\sum_{j=1}^{J} (s_j(k) - 1\{s_j(k) > 0\} + n_j(k)) \,\middle|\, \mathbf{s}(k)\right] \\ &\quad = V(\mathbf{s}(k)) - E\left[\sum_{j=1}^{J} 1\{s_j(k) > 0\} \,\middle|\, \mathbf{s}(k)\right] \\ &\qquad + E\left[\sum_{j=1}^{J} n_j(k) \,\middle|\, \mathbf{s}(k)\right] < \infty \end{aligned} \tag{8}$$

because the third term is always finite due to Poisson arrivals $n_j(k)$ and bounded epoch length. Here, $1\{.\}$ is the indicator function.

Consider the last condition. We have

$$\begin{aligned} &E[V(\mathbf{s}(k+1)) - V(\mathbf{s}(k)) \mid \mathbf{s}(k)] \\ &\quad = E\left[\sum_{j=1}^{J} (s_j(k+1) - s_j(k)) \,\middle|\, \mathbf{s}(k)\right] \\ &\quad = \sum_{j=1}^{J} E[n_j(k) - 1\{s_j(k) > 0\} \mid \mathbf{s}(k)] \\ &\quad = \sum_{j=1}^{J} \lambda_j l(k) - \sum_{j=1}^{J} (1\{s_j(k) > 0\} \mid \mathbf{s}(k)) \end{aligned} \tag{9}$$

where $l(k)$ is the kth epoch length in slots, and it is equal to the number of transmitted packets $\sum_{j=1}^{J} 1\{s_j(k)\}$. Hence

$$E[V(\mathbf{s}(k+1)) - V(\mathbf{s}(k)) \mid \mathbf{s}(k)] = l(k)\left(\sum_{j=1}^{J} \lambda_j - 1\right).$$

Let $\boldsymbol{S}_0 = \{\mathbf{0}\}$. Then, $\forall\, \mathbf{s}(k) \notin \boldsymbol{S}_0$, we have that $l(k) \ge 1$. Therefore, $\forall\, \mathbf{s}(k) \notin \boldsymbol{S}_0$, if (6) holds, then there exists ϵ, where $0 < \epsilon < 1 - \sum_{j=1}^{J} \lambda_j$, such that

$$E[V(\mathbf{s}(k+1)) - V(\mathbf{s}(k)) \mid \mathbf{s}(k)] < -\epsilon. \tag{10}$$

It follows that all conditions of the Foster's criterion (7), (8), and (10) are satisfied. Therefore, we conclude that $\boldsymbol{s}(k)$ is ergodic. Hence, (6) is sufficient for stability.[3] ■

IV. Stability of BNDMA via (σ, ρ) Deterministic Fluid Approach

In order to strengthen the BNDMA stability result obtained via the dominant system approach, we first attempted to show that BNDMA satisfies the conditions of the monotone separable framework of Baccelli and Foss [4] (see also [1]). In [4], the authors provide a rigorous framework for application of the "saturation rule." While we were successful in demonstrating that BNDMA conforms to this framework, calculating a key constant $(\gamma(0))$ in [4, Th. 1] turned out to be a formidable task, due to the dependence of the queues. For this reason, we adopt an alternate route, first suggested by Cruz [7], in which packet arrivals are assumed to satisfy certain deterministic constraints along each sample path. This approach conforms to a leaky bucket rate control mechanism, and the ensuing analysis captures the essence of the protocol without distractions due to intricate asymptotic probabilistic behavior.

Specifically, at this point, we depart from the Poisson arrivals assumption and revert to the following alternative:[4] The number of packet arrivals to the jth queue over the time interval $[s, t)$, denoted by $n_j(s, t)$, satisfies

$$n_j(s, t) \le \lambda_j(t - s) + \sigma_j. \tag{11}$$

Note that it is customary to use ρ_j instead of λ_j for the slope in this context [7]; however, since this slope serves as an upper bound on the long-term average rate, we prefer to use λ_j for simplicity and uniformity with the rest of the paper. $0 \le \sigma_j < \infty$ is a measure of burstiness [7]. Since we are dealing with a slotted system, time is measured in slots, and the variables s, t are integer. The initial state of the jth queue, which is denoted by $s_j(0)$, is assumed to be a finite non-negative integer but is otherwise arbitrary.

In our present deterministic fluid context, stability means that every queue in the system remains bounded. Thus, we aim to show that under a suitable condition on the λ_js, the state

[3]It can be shown that Theorem 1 holds under more general conditions, i.e., stationary ergodic arrivals. The idea is that the sum of the queue lengths in the NDMA system can be shown to be bounded above by J plus the length of a single server queue with sum input. From Loynes' Theorem (e.g., [11]), the single server queue is stable under (6). This argument is more general, but the Foster–Lyapunov approach sheds more light into the system dynamics because it is tied to the familiar concept of drift.

[4]This is done only for the purposes of a tractable stability analysis; throughout the rest of the paper, the usual Poisson arrivals assumption is in effect.

(backlog) of every queue in the system will remain bounded for all time, irrespective of initial conditions. As it turns out, it is easier to prove that every queue in the system will empty out in finite time infinitely often, irrespective of initial conditions and the state of other queues in the system. This, in turn, implies that every queue remains bounded.

We have the following result.

Theorem 2: The B-NDMA system is stable if

$$\sum_{j=1}^{J} \lambda_j + \max_{j=1}^{J}(\lambda_j) < 1. \tag{12}$$

Proof: The proof is by induction, and the proof of each step is by contradiction. Without loss of generality, we assume that

$$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_J. \tag{13}$$

In addition, if $s_i(0)+\sigma_i > s_j(0)+\sigma_j$ for $i<j$, then we may increase the burstiness allowance σ_j to σ_j' such that $s_i(0)+\sigma_i \le s_j(0)+\sigma_j'$. This assures that

$$s_j(0)+\lambda_j t+\sigma_j \le s_{j+1}(0)+\lambda_{j+1}t+\sigma_{j+1} \tag{14}$$

for all j and $t \ge 0$.

• `Show that queue 1 remains bounded for all time.`

Consider the first queue, and assume that it *never empties*. Then, it remains continuously backlogged, and since the BNDMA "server" is work-conserving, the first queue transmits at all times. Over the time interval $[0, T)$, queue 1 accumulates at most $q_1^{\max} := s_1(0) + \lfloor \lambda_1 T \rfloor + \sigma_1$ packets, where $\lfloor x \rfloor$ denotes the largest integer not greater than x. If it transmits continuously, then it can be active over at most[5] $(J+1)q_1^{\max} = (J+1)\lfloor \lambda_1 T \rfloor + (J+1)[s_1(0)+\sigma_1]$ contiguous slots, where $(J+1)$ is the longest possible CR epoch length. From (12) and (13), it follows that $\lambda_1 < 1/(J+1)$. Noting that $\lfloor \lambda_1 T \rfloor \le \lambda_1 T$, we write $\lfloor \lambda_1 T \rfloor = (1/J+1)T - \delta_1 T$, $\delta_1 > 0$. This yields that queue 1 can be active in at most $t_{1,a} = T + (J+1)[s_1(0)+\sigma_1 - \delta_1 T]$ slots. Noting that the transmission time of packets is quantized in epochs of length $J+1$ slots, if $T \ge (J+1)\lceil (s_1(0)+\sigma_1)/\delta_1(J+1) \rceil$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x, then $t_{1,a} < T$. Since $t_{1,a} < T$ for finite T, it contradicts the assumption that queue 1 never empties.

It follows that queue 1 will empty in finite time, irrespective of initial conditions and the state of other queues in the system. We may now repeat this exact argument to claim that queue 1 will empty out infinitely often as time tends to infinity, irrespective of initial conditions and the state of other queues in the system. Thus, queue 1 remains bounded for all time.

• `Induction hypothesis:` Queues 1 to $j-1$ remain bounded for all time.

• `Induction step:` Show that, under the induction hypothesis, queue j also remains bounded for all time.

Let us again begin by showing that queue j will empty out once in finite time. Let $q_i^{\max} = s_i(0) + \lfloor \lambda_i T \rfloor + \sigma_i$ denote the maximum number of packets transmitted by the ith queue over time T, where $i = 1, \ldots, j$. Assume that the jth queue never empties. It therefore transmits continuously in every slot. Under this scenario and using (14), the longest possible activity burst of the jth queue is obtained for the following distribution of epochs.

- $v_1 = q_1^{\max}$ epochs of length $J+1$ when all queues transmit.
- $v_2 = q_2^{\max} - q_1^{\max}$ epochs of length J when all queues except the first queue transmit (the first queue is idle because it is empty).
- $v_3 = q_3^{\max} - q_2^{\max}$ epochs of length $J-1$ when all queues except the first and second queue transmit, and so on, until only the jth and higher ordered queues remain transmitting.
- $v_j = q_j^{\max} - q_{j-1}^{\max}$ epochs of length $J+2-j$.

This yields the following upper bound on the length of time over which queue j can remain continuously active

$$t_{j,a} \le \sum_{l=1}^{j} (J+2-l)v_l.$$

By substituting the v_ls, we obtain

$$t_{j,a} \le \sum_{l=1}^{j-1}[s_l(0)+\sigma_l] + (J+2-j)[s_j(0)+\sigma_j] + \sum_{l=1}^{j-1}\lfloor \lambda_l T \rfloor + (J+2-j)\lfloor \lambda_j T \rfloor. \tag{15}$$

From (12) and (13), it follows that

$$\lambda_j < \frac{1}{J+2-j}\left[1 - \sum_{l=1}^{j-1}\lambda_l\right]$$

which yields

$$\lfloor \lambda_j T \rfloor = \frac{1}{J+2-j}\left[1 - \sum_{l=1}^{j-1}\lambda_l\right] T - \delta_j T, \qquad \delta_j > 0.$$

Therefore, (15) becomes

$$\begin{aligned} t_{j,a} &\le \sum_{l=1}^{j-1}[s_l(0)+\sigma_l] + (J+2-j)[s_j(0)+\sigma_j] \\ &\quad + [1-(J+2-j)\delta_j]T \\ &= T + (J+2-j)\delta_j \\ &\quad \cdot \left\{ \frac{\sum_{l=1}^{j-1}[s_l(0)+\sigma_l] + (J+2-j)[s_j(0)+\sigma_j]}{(J+2-j)\delta_j} - T \right\}. \end{aligned}$$

It follows that if

$$T \ge \left\lceil \frac{\sum_{l=1}^{j-1}[s_l(0)+\sigma_l] + (J+2-j)[s_j(0)+\sigma_j]}{(J+2-j)\delta_j} \right\rceil$$

where T is chosen to consist of an integer number of epochs of lengths $J+2-j, \ldots, J+2$, according to v_ls for $l = 1, \ldots, j$, then $t_{j,a} < T$, which contradicts the assumption that queue j never empties. Hence, queue j will empty once in finite time.

[5]Recall that no packet is lost in transmission.

So far, we have not used the induction hypothesis; it comes into play at this point. In order to repeat the above argument to show that queue j will empty in infinite time *infinitely often*, we need to have that the backlog of all lower ordered queues is bounded at the beginning of subsequent queue-j evacuation intervals (c.f. the last inequality). This is assured by the induction hypothesis, and thus, the proof is complete. ■

V. Steady-State Analysis

Assuming stability, let $P_{e,j} := \lim_{k\to\infty} \Pr\{s_j(k) = 0\}$, $(j = 1, 2, \ldots, J)$ be the probability that queue j is empty at the beginning of an epoch in the steady state. It is shown in [15] and [17] that by knowing P_e (in a symmetric system, $P_{e,j} = P_e, \forall j$), one can find an approximate distribution of CR epoch lengths. Then, by finding the first and second moments of CR epoch lengths, one can approximate delay. Therefore, the steady-state analysis provided relations between P_e and the average arrival rate λ so that delay could be expressed in terms of λ.

Analogously, our goal is to find a relation between $P_{e,j}$ and the vector of average arrival rates $(\lambda_1, \ldots, \lambda_J)$. This will serve us in finding delay for each user, as elaborated upon in Section VII.

We now revert to Poisson arrivals in order to focus on steady-state behavior. Note that in the steady-state, a queue must have the same average number of incoming and outgoing packets during the average epoch length, which is denoted $E[l]$.[6] Note that $(1 - P_{e,j})$ is the average number of transmitted packets by queue j during $E[l]$. Therefore, the balance equations are

$$\lambda_j E[l] = 1 - P_{e,j}, \qquad \text{for } j = 1, 2, \ldots, J \tag{16}$$

and, therefore, also

$$\sum_{j=1}^{J} \lambda_j E[l] = \sum_{j=1}^{J} (1 - P_{e,j}). \tag{17}$$

Let $\rho := \sum_{j=1}^{J}(1 - P_{e,j})$, which is the average total number of transmitted packets during $E[l]$. Since each active user transmits exactly one packet, ρ is also the average number of active users during $E[l]$. From the protocols [15] and [17], it follows that in the kth epoch

$$\text{NDMA:} \quad l(k) = \begin{cases} 1, & \text{if 0 users transmit} \\ \#\ active\ users(k), & \text{otherwise} \end{cases}$$
$$\text{BNDMA:}\ l(k) = \#\ active\ users(k) + 1. \tag{18}$$

Note that for BNDMA:

$$l = \sum_{j=1}^{J} 1\{s_j(k) > 0\} + 1$$

which implies by linearity of expectation

$$E[l] = \sum_{j=1}^{J} E1\{s_j(k) > 0\} + 1$$
$$= \sum_{j=1}^{J} (1 - P_{e,j}) + 1 = \rho + 1$$

[6]This is a slight abuse of notation, where $E[l]$ stands for $E[l(k)]$. This convention is used throughout the rest of the paper.

whereas for NDMA

$$l = 1 \times 1\{\mathbf{s}(k) = \mathbf{0}\} + \sum_{j=1}^{J} 1\{s_j(k) > 0\}$$

which again implies by linearity of expectation

$$E[l] = \Pr\{\mathbf{s}(k) = \mathbf{0}\} + \sum_{j=1}^{J}(1 - P_{e,j}) = \Pr\{\mathbf{s}(k) = \mathbf{0}\} + \rho.$$

We have already noted that the queues are coupled so that Markov chain is nonseparable multidimensional. Therefore, it is difficult to find the exact expression for $\Pr\{\boldsymbol{s}(k) = \mathbf{0}\}$. To make the analysis tractable, we will assume that the queues are independent, so that $\Pr\{\boldsymbol{s}(k) = \mathbf{0}\} = \prod_{j=1}^{J} P_{e,j}$. This is justifiable for low traffic loads when there is small probability of contention among users and epochs are short. However, at high loads, this assumption can lead to inaccurate estimates of $P_{e,j}$. Hence, we will use simulations to verify our analysis.

By using the independence assumption, we have that $P_{e,j}$ for $j = 1, 2, \ldots, J$ satisfy

NDMA:	$\lambda_j \left(\prod_{i=1}^{J} P_{e,i}\right) + \left(1 - \sum_{i=1}^{J} \lambda_i\right) P_{e,j} - \left(1 - \sum_{i=1}^{J} \lambda_i\right) = 0$
BNDMA:	$P_{e,j} = 1 - \dfrac{\lambda_j}{1 - \sum_{i=1}^{J} \lambda_i}$

. (19)

Remark 1: Note that the independence assumption is only invoked to derive (19) [NDMA]; for BNDMA, the independence assumption is not necessary, and hence, (19) [BNDMA] is *exact*.

The uniqueness of solution of the above systems of equations is established in the following proposition.

Proposition 2: $P_{e,j}$ has a unique solution in (0, 1), $\forall j$, if

NDMA:	$\sum_{i=1}^{J} \lambda_i < 1$
BNDMA:	$\sum_{i=1}^{J} \lambda_i + \max_{i=1}^{J}(\lambda_i) < 1$

.

The proof is given in the Appendix, which also shows that (19) [NDMA] boils down to polynomial rooting in (0, 1).

This generalizes the steady-state analysis in [15] and [17] to the asymmetric (multirate) case in a much simpler way. The formulas in (19) allow direct calculation of ρ and $E[l]$, from which throughput can be calculated. In addition, given the above formula for $P_{e,j}$, delay can be accurately approximated (see Section VII).

Note that the conditions of Proposition 2 are *necessary* for stability in the usual Markovian sense. For NDMA, it is obvious that the system is unstable if $\sum_{i=1}^{J} \lambda_i > 1$. For BNDMA, if the condition of Proposition 2 is not satisfied, then, at least for the queue with the highest arrival rate (19), [BNDMA] does not

yield a positive solution for $P_{e,j}$. This contradicts stability in the Markovian sense and shows that, at least for Poisson arrivals, the total load that is less than $J/(J+1)$ packets per slot may not be sufficient for stability, even though a K-fold collision is resolved in $K+1$ slots in BNDMA.

VI. Generalized BNDMA

At a given total offered traffic load $\sum_{i=1}^{J}\lambda_i$, in a BNDMA system, a queue with higher arrival rate would have smaller stability margin (lower $P_{e,j}$) than a queue with lower arrival rate (c.f. (19) [BNDMA]). Such a high-rate queue will have more backlogged packets, higher queuing delay, and therefore higher total average delay, than a low-rate queue (analysis follows in Section VII). To improve the stability margin and decrease delay, we have proposed that the jth queue in BNDMA be allowed to transmit up to $m_j \geq 1$ packets simultaneously at the beginning of a CR epoch, provided it is nonempty [8]. Note that this allows *limited* contention between packets from the same queue.

To preserve the CR method used in BNDMA (in particular, to ensure that the mixing matrix $\boldsymbol{A}_{T\times K}$ is Vandermonde and has full rank w.p. 1), independent phase modulation is employed for different packets of the same queue that are transmitted during the same epoch (in addition to independent phase modulation across users).

The embedded Markov chain state-transition equation [with the same notation as in (2)] becomes

$$s_j(k+1)=\begin{cases} s_j(k)-m_j+n_j(k), & s_j(k)\geq m_j \\ n_j(k), & 0\leq s_j(k)<m_j \end{cases} \quad \text{for } j=1,2,\ldots,J. \tag{20}$$

Note that given a state, the number of packets transmitted during the following epoch, and, consequently, the epoch length, are deterministic. Therefore, $\mathbf{s}(k)$, which is given in (20), is a homogeneous, irreducible, aperiodic Markov chain with a countable state-space (analogously to Proposition 1).

A. Stability of Generalized BNDMA

Note that a generalized BNDMA (G-BNDMA) system as described above can be viewed as splitting the jth user's queue in m_j subqueues, where metering is used for the assignment of incoming packets to each subqueue. Moreover, each subqueue transmits exactly one packet when it is nonempty and allowed to transmit. Under the same fluid traffic model as in the proof of the stability Theorem 2, metering induces "decimated" constraints on the subqueues, i.e., subqueue i of queue j receives at most $(\lambda_j/m_j)T+constant$ packets over a time interval of length T. Then, Theorem 2 immediately yields the following stability result for G-BNDMA:

Corollary 1: The G-BNDMA system is stable if

$$\sum_{j=1}^{J}\lambda_j+\max_{j=1}^{J}\left\{\frac{\lambda_j}{m_j}\right\}<1. \tag{21}$$

B. Steady-State Analysis of G-BNDMA

We again revert to Poisson arrivals to the queues in order to discuss the steady-state behavior of G-BNDMA. Due to metering, arrivals to *subqueues* are not independent Poisson, which significantly complicates the analysis. Therefore, we approximate metering with random packet assignment, which preserves Poisson distribution and independence of arrivals to subqueues. Since each subqueue can transmit no more than one packet during an epoch, a G-BNDMA system with J queues with arrival rates λ_j ($j=1,2,\ldots,J$) is approximated by a BNDMA system with $\sum_{i=1}^{J}m_i$ terminals with arrival rates λ_j/m_j for all m_j subqueues of the jth queue of the original G-BNDMA system.

Let $P_{e,j}$ denote the probability that a subqueue of the jth queue of the system with random packet assignment is empty in the steady state. The balance equation (16) becomes

$$\frac{\lambda_j}{m_j}E[l]=1-P_{e,j},\qquad \text{for } j=1,2,\ldots,J \tag{22}$$

so that $\sum_{j=1}^{J}m_j(\lambda_j/m_j)E[l]=\sum_{j=1}^{J}m_j(1-P_{e,j})$. By following the same steps as in Section V, we obtain

$$P_{e,j}=1-\frac{\lambda_j/m_j}{1-\sum_{i=1}^{J}\lambda_i},\qquad j=1,2,\ldots,J. \tag{23}$$

Clearly, by increasing m_j, the stability margin is improved.

VII. Delay

Delay in a symmetric ($\lambda_1=\cdots=\lambda_J=\lambda$) NDMA or BNDMA system can be closely approximated [15], [17] by modeling each queue as an M/G/1 queue with server vacation. This M/G/1 queue's service time is equal to the average length of an epoch in which a particular queue transmits a packet—relevant epoch—$E[l_r]$, and vacation time is equal to the average length of an epoch in which a particular queue is idle—irrelevant epoch—$E[l_i]$. We reproduce the delay formula for convenience

$$D=E[l_r]+\frac{\lambda E[l_r^2]}{2(1-\lambda E[l_r])}+\frac{E[l_i^2]}{2E[l_i]}.$$

The first and second moments of relevant and irrelevant epoch length are functions of P_e [15], [17] obtained from the probability mass functions of l_r and l_i

NDMA:	$P(l_r=b)=\binom{J-1}{b-1}(1-P_e)^{b-1}P_e^{J-b}$, $1\leq b\leq J$ $P(l_i=b)=\binom{J-1}{b}(1-P_e)^{b}P_e^{J-b-1}+P_e^{J}\delta(b-1)$, $1\leq b\leq J-1$
BNDMA:	$P(l_r=b)=\binom{J-1}{b-2}(1-P_e)^{b-2}P_e^{J-b+1}$, $2\leq b\leq J+1$ $P(l_i=b)=\binom{J-1}{b-1}(1-P_e)^{b-1}P_e^{J-b}$, $1\leq b\leq J$

Note that the above distributions assume independence of queues. This has already been assumed and discussed in Section V.

From the point of view of a particular user, every CR epoch in an asymmetric NDMA or BNDMA system is either relevant or irrelevant, just like a CR epoch in a symmetric system. Therefore, delay in an asymmetric NDMA or BNDMA system is also approximated by modeling each queue as an M/G/1 queue with server vacations, where service time is equal to the average relevant epoch length, and vacation time is equal to the average irrelevant epoch length.

To find approximate delay expressions for an asymmetric G-BNDMA system, we use the random packet assignment approximation as in the steady-state analysis (see Section VI-B). Hence, we analyze delay of a BNDMA system with $\sum_{j=1}^{J} m_j$ terminals with rates λ_j/m_j $(j = 1, 2, \ldots, J)$ for all m_j subqueues of the jth queue of the original system. Each subqueue can transmit at most one packet during an epoch. Note that by using this approximation, the average delay is the same for all subqueues of the same queue because they have the same $P_{e,j}$ (see Section VI-B). Therefore, we need only find approximate delay expressions for each queue. Hence, we develop a unified approach to approximating delay of any asymmetric NDMA or BNDMA or G-BNDMA system, where different queues may have different $P_{e,j}$s and, hence, different relevant and irrelevant epoch length distributions and delay.

Let $\boldsymbol{s}^a(k) = [s_{(1,1)}(k), \ldots, s_{(1,m_1)}(k), s_{(2,1)}(k), \ldots, s_{(2,m_2)}(k), \ldots, s_{(J,m_J)}(k)]^T$ be the state of the approximate G-BNDMA model, where $s_{(j,z)}(k)$ denotes the state of the zth subqueue of the jth queue at the beginning of the kth epoch. Let $l_{r,j}$ and $l_{i,j}$ denote the relevant and irrelevant epoch length of any subqueue of the jth queue. Note that by setting $m_j = 1$ for all $j = 1, \ldots, J$, $\boldsymbol{s}^a(k)$ accommodates NDMA and BNDMA as well.

Thus, delay is estimated by using the following approximation:

$$D_j = E[l_{r,j}] + \frac{\frac{\lambda_j}{m_j} E[l_{r,j}^2]}{2\left(1 - \frac{\lambda_j}{m_j} E[l_{r,j}]\right)} + \frac{E[l_{i,j}^2]}{2E[l_{i,j}]} \qquad j = 1, 2, \ldots, J. \tag{24}$$

The first and second moments of $l_{r,j}$ and $l_{i,j}$ depend on the steady-state behavior of all the queues and, hence, are functions of $(P_{e,1}, P_{e,2}, \ldots, P_{e,J})$. The probability mass functions $P(l_{r,j} = b)$ and $P(l_{i,j} = b)$ sum up the probabilities of all the realizations in which all the subqueues, except for the subqueue of interest transmit t packets, given the epoch length b. Based on NDMA and BNDMA protocols, t and b are related as follows:[7]

NDMA:	relevant epoch:	$t = b - 1$
	irrelevant epoch:	if $b = 1$, then $t = 0$ and $t = 1$ else $t = b$
BNDMA:	relevant epoch:	$t = b - 2$
	irrelevant epoch:	$t = b - 1$

(25)

[7]In NDMA, if epoch is irrelevant and $b = 1$, the only realization in which $t = 0$ packets are transmitted (all-zero state) is counted together with realizations in which $t = 1$ packet is transmitted.

Again, assume independence of queues. Let $I_{(j,z)}(k) := 1\{s_{(j,z)}(k) > 0\}$, and define

$$I(k) = \left[I_{(1,1)}(k), \ldots, I_{(1,m_1)}(k), I_{(2,1)}(k) \ldots, I_{(J,m_J)}(k)\right]^T$$

$$S_{r,t}(j^*, z^*) := \left\{ I \;\middle|\; t = \sum_{j=1}^{J} \sum_{\substack{z=1 \\ (j,z)\neq(j^*,z^*)}}^{m_j} I_{(j,z)}, \text{ and } I_{(j^*,z^*)} = 1 \right\}$$

$$S_{i,t}(j^*, z^*) := \left\{ I \;\middle|\; t = \sum_{j=1}^{J} \sum_{\substack{z=1 \\ (j,z)\neq(j^*,z^*)}}^{m_j} I_{(j,z)}, \text{ and } I_{(j^*,z^*)} = 0 \right\} \tag{26}$$

where (j^*, z^*) denotes the index of the subqueue of interest. Hence, $S_{r,t}$ and $S_{i,t}$ are the sets of all permutations of t active subqueues during a relevant or an irrelevant epoch, respectively. This and (25) gives the following distribution of relevant and irrelevant epoch length[8]

$$P(l_{r,j} = b) = \sum_{I \in S_{r,t}(j^*,z^*)} \prod_{j=1}^{J} \prod_{\substack{z=1 \\ (j,z)\neq(j^*,z^*)}}^{m_j} (1 - P_{e,j})^{I_{(j,z)}} P_{e,j}^{1 - I_{(j,z)}}$$

$$P(l_{i,j} = b) = \sum_{I \in S_{i,t}(j^*,z^*)} \prod_{j=1}^{J} \prod_{\substack{z=1 \\ (j,z)\neq(j^*,z^*)}}^{m_j} (1 - P_{e,j})^{I_{(j,z)}} P_{e,j}^{1 - I_{(j,z)}}. \tag{27}$$

Note that providing a BNDMA system with the simultaneous multiple packet transmission from each queue not only improves the stability margin, but it can also decrease delay. This is corroborated by simulations. In particular, in a symmetric system, the following holds.

Proposition 3: Consider the approximate delay expression for the G-BNDMA system in (24), assuming independence of queues, random packet assignment, and M/G/1 queue with server vacation approximation. In the noiseless case, if $\lambda_j = \lambda$, $\forall\, j$ and $m_j = m$, $\forall\, j$, then the average delay D decreases as m increases, for all values of offered traffic load λJ.

The proof is given in the Appendix. Note, however, that there are practical limitations on m.

VIII. Simulation Results

We performed Monte Carlo (MC) simulations of the NDMA, BNDMA and G-BNDMA systems. The values of probabilities that the jth queue is empty $P_{e,j}$ and the delay of the jth queue packets D_j obtained by simulation (in the noiseless case) are compared with the analytic results. The delay of each system is compared with the delay of slotted ALOHA with first-come-first-serve (FCFS) splitting protocol for collision

[8]Recall that in the G-BNDMA case, $P_{e,j}$ denotes the probability that a subqueue of the jth queue is empty.

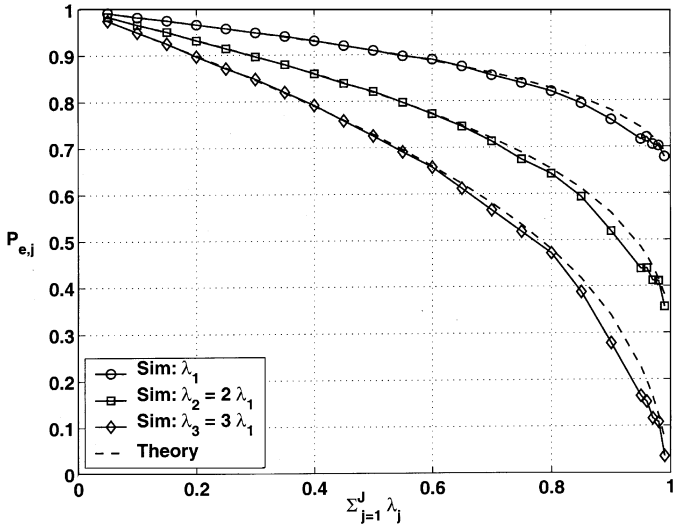

Fig. 1. NDMA: $P_{e,\,j}$ versus total load.

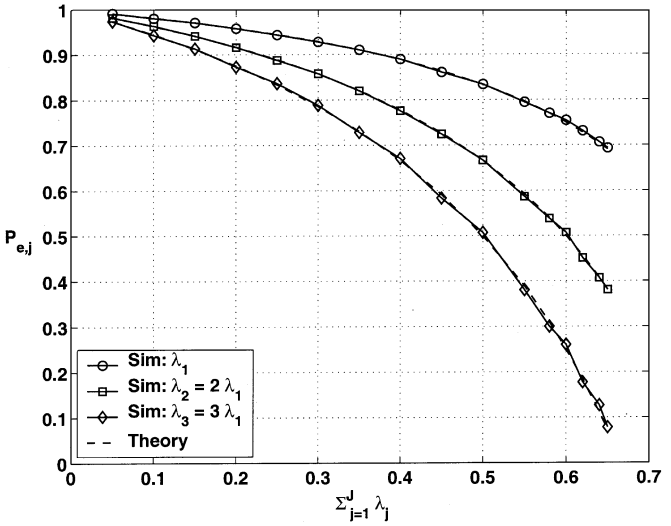

Fig. 3. BNDMA: $P_{e,\,j}$ versus total load.

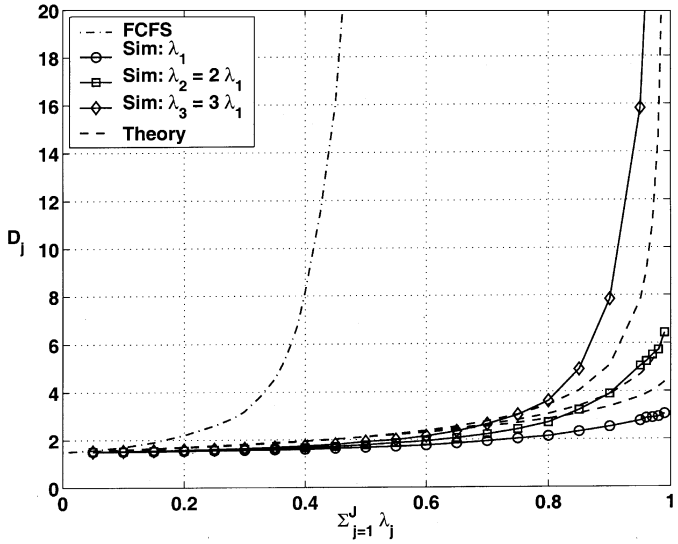

Fig. 2. NDMA: Delay versus total load.

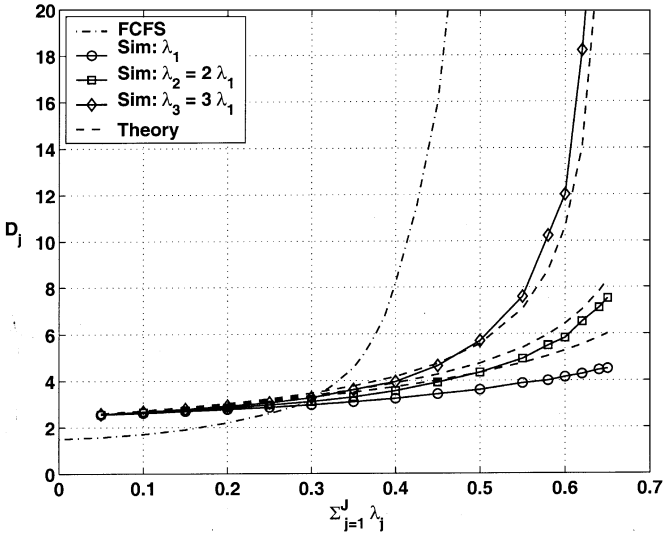

Fig. 4. BNDMA: Delay versus total load.

resolution, which was obtained by simulation. The simulated (G/B)NDMA systems have $J = 3$ queues with arrival rates $(\lambda_1, \lambda_2, \lambda_3) = (\lambda_1, 2\lambda_1, 3\lambda_1)$. $P_{e,\,j}$ and D_j are plotted against total offered traffic load $\sum_{j=1}^{J} \lambda_j$. Note that in the noiseless case, throughput is equal to offered load for all four protocols considered (NDMA, BNDMA, G-BNDMA, FCFS splitting), provided a system is stable.

Figs. 1 and 2 depict $P_{e,\,j}$ versus $\sum_{j=1}^{J} \lambda_j$ and D_j versus $\sum_{j=1}^{J} \lambda_j$, respectively, for NDMA. Fig. 2 also shows a comparison of NDMA versus FCFS (dash–dotted line). The full lines denote MC simulation results, and the dashed lines depict analytic results. We see that at low to medium traffic loads, $P_{e,\,j}$s are accurately estimated, which corroborates our assumption that queues are practically independent. However, at medium to high traffic loads, queues are coupled so that the actual $P_{e,\,j}$s are lower than estimated. At low to medium traffic loads, simulation results for delay are close to analytic results and even slightly better. Note that analytic results for NDMA delay are based on the M/G/1 queue with server vacation approximation, which is valid if service time (relevant epoch length, $l_{r,\,j}$) and vacation time (irrelevant epoch length, $l_{i,\,j}$) are independent. However, both depend on offered load and must be dependent [15]. Correct estimation of $P_{e,\,j}$s at low to medium traffic loads and dependency of epoch lengths yield actual delay that is lower than estimated. At high traffic loads, low-rate users' delay remains smaller than estimated due to dependency of $l_{r,\,j}$ and $l_{i,\,j}$ in the M/G/1 model. However, high-rate users' delay is significantly higher than estimated. To explain this, note that actual values of $P_{e,\,j}$s are smaller than estimated so that actual values of $E[l_{r,\,j}]$ and $E[l_{i,\,j}]$ are larger than estimated. In addition, due to dependency of queues, epoch lengths do not have binomial distribution. Simulations show that actual distribution yields higher values of $E[l_{r,\,j}^2]$ and $E[l_{i,\,j}^2]$, given $P_{e,\,j}$. Thus, for high-rate users, high actual values of the second moments significantly increase total delay and make it higher than estimated. Interestingly, for medium-rate users, overestimation of delay due to M/G/1 approximation and underestimation of delay due to underestimation of the first and second moments of epoch lengths cancel each other, yielding accurate delay estimation.

Figs. 3 and 4 show results for BNDMA with the same notation as for NDMA. Note that for BNDMA, the steady-state analysis in Section V is exact, i.e., it takes queue dependence into

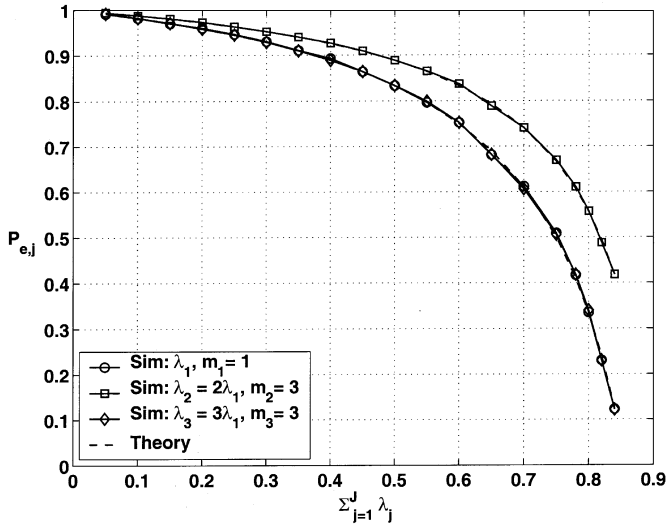

Fig. 5. Gen. BNDMA: $P_{e,j}$ versus total load.

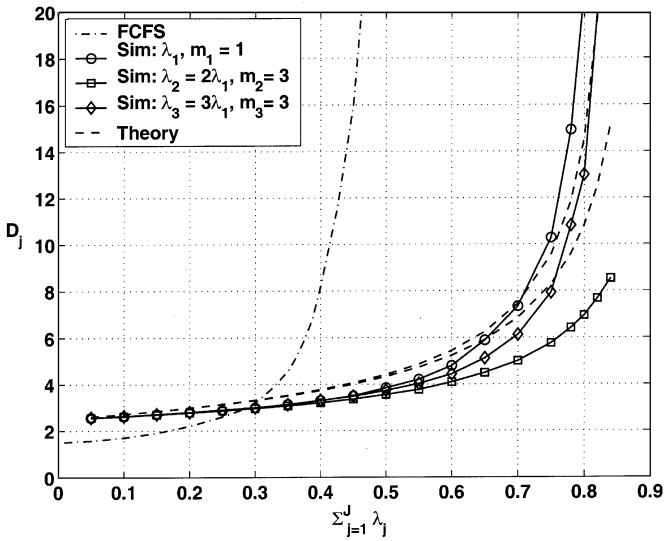

Fig. 6. Gen. BNDMA: Delay versus total load.

account. Hence, analytic $P_{e,j}$ results are accurate, even at high loads. Delay is lower than estimated, except for the highest rate user at high offered load. Lower actual delay is due to dependence of $l_{r,j}$ and $l_{i,j}$ in the M/G/1 approximation, which yields a pessimistic estimate of delay, given correct estimates of $P_{e,j}$s. In addition, note that at high traffic loads, epoch length distribution is not binomial. The actual distribution yields higher values of $E[l_{r,j}^2]$ and $E[l_{i,j}^2]$, given $P_{e,j}$. This increases actual delay and even compensates for delay overestimation due to M/G/1 approximation.

For generalized BNDMA, we used the following values for multipacket transmissions $\mathbf{m} = (m_1, m_2, m_3) = (1, 3, 3)$. Note that theoretic results for both $P_{e,j}$ and D_j are based on independent queues and random packet assignment (RPA) approximations, whereas delay analysis also includes the M/G/1 with server vacation approximation. The results are presented in Figs. 5 and 6. Clearly, multipacket transmissions improve performance compared to BNDMA. The discussion on B-NDMA applies here as well. Note that the estimated delay for users 1 and 3 is the same because they have the same $P_{e,j}$s. However, actual delay of user 3 is lower. This is due to RPA approximation, which yields pessimistic delay estimation for users with multipacket transmission. In addition, note that user 2 has the lowest λ_j/m_j ratio and, hence, highest $P_{e,j}$ and lowest delay.

IX. Conclusions

A unified stability and steady-state analysis of a class of collision resolution protocols with retransmission diversity has been provided. This bridges a gap in earlier analyses. For NDMA, a unique sufficient and necessary condition for stability is obtained, assuming Poisson arrivals.[9] It proves that NDMA has maximum throughput that approaches 1. For BNDMA, a sufficient stability condition is obtained for deterministic fluid arrivals, whereas the same condition is necessary when Poisson arrivals are assumed. Based on these results, a generalization of BNDMA, which allows multiple packet transmission from the same queue, is proposed. It is proven that the latter protocol has increased maximum stable throughput, which can be made close to 1. The tools used in the stability analysis range from a preliminary dominant system approach to the Foster–Lyapunov criterion and Cruz's (σ, ρ) deterministic fluid approach. The stability, steady-state, and delay analyses are extended to asymmetric (multirate) systems. Simulations show that the analysis provides good approximation of the delay performance for the whole class of protocols.

[9]The behavior of the system at the stability region boundary was not considered.

Appendix

Proof (Proposition 2): For NDMA, first note that we consider only the case when $\lambda_j > 0, \forall j$. Otherwise, (19) can be reduced to a system with $J - N_0$ equations in $J - N_0$ unknowns, where N_0 is the number of queues with $\lambda_j = 0$. Let $r = 1 - \sum_{i=1}^{J} \lambda_i$.

By multiplying the nth equation by λ_j/λ_n and subtracting it from the jth equation, $\forall j$, $j \neq n$, the following equivalent system is obtained:

$$\lambda_n \left(\prod_{i=1}^{J} P_{e,i} \right) + r P_{e,n} - r = 0$$

$$r \left[\left(P_{e,j} - \frac{\lambda_j}{\lambda_n} P_{e,n} \right) - \left(1 - \frac{\lambda_j}{\lambda_n} \right) \right] = 0$$

$$\text{for } j = 1, \ldots, J, \text{ and } j \neq n.$$

If $\sum_{i=1}^{J} \lambda_i < 1$, then $r > 0$, so that all equations except the nth can be divided by r. Without loss of generality, suppose that $n = 1$. Let $a_j = \lambda_j/\lambda_1$, for $j = 1, \ldots, J$. Further, assume that λ_1 is the highest arrival rate [$\lambda_1 = \max_j(\lambda_j)$, one can always enumerate queues in such a way]. Therefore

$$0 < a_j \leq 1, \qquad \forall j.$$

Substitute $P_{e,j} = f_j(P_{e,1}, a_j)$, for $j = 2, \ldots, J$ into $\prod_{i=1}^{J} P_{e,i}$:

$$\prod_{i=1}^{J} P_{e,i} = P_{e,1} \prod_{i=2}^{J} (a_i P_{e,1} + (1 - a_i)) = \sum_{i=1}^{J} b_i P_{e,1}^{J+1-i}$$

where

$$\begin{aligned}
b_1 &= \left(\prod_{i=2}^{J} a_i \right) \\
b_2 &= \sum_{i_1=2}^{J} (1 - a_{i_1}) \left(\prod_{\substack{i=2 \\ i \neq i_1}}^{J} a_i \right) \\
b_3 &= \sum_{i_1=2}^{J} \sum_{i_2 > i_1} (1 - a_{i_1})(1 - a_{i_2}) \left(\prod_{\substack{i=2 \\ i \neq i_1, i \neq i_2}}^{J} a_i \right) \\
&\vdots \\
b_k &= \sum_{i_1=2}^{J} \sum_{i_2 > i_1} \sum_{i_3 > i_2} \cdots \sum_{i_k > i_{k-1}} \left[\prod_{p=1}^{k} (1 - a_{i_p}) \right] \\
&\quad \cdot \left(\prod_{\substack{i=2 \\ i \neq i_1, \ldots, i \neq i_k}}^{J} a_i \right) \\
&\vdots \\
b_J &= \prod_{i=2}^{J} (1 - a_i).
\end{aligned}$$

Therefore, the equivalent system of equations becomes

$$\begin{aligned}
\sum_{i=0}^{J} c_i P_{e,1}^{J-i} &= 0, \\
P_{e,j} &= a_j P_{e,1} - (1 - a_j), \quad \text{for } j = 2, \ldots, J
\end{aligned}$$

where $c_i = \lambda_1 b_{i+1}$, for $i = 0, 1, \ldots, J-2$, $c_{J-1} = \lambda_1 b_J + r$, and $c_J = -r$. Hence, $c_0 > 0$, $c_i \geq 0$, for $i = 1, 2, \ldots, J-2$, $c_{J-1} > 0$, and $c_J < 0$. It follows that

$$\begin{aligned}
\lim_{P_{e,1} \to 0} \sum_{i=0}^{J} c_i P_{e,1}^{J-i} &= c_J < 0 \\
\lim_{P_{e,1} \to 1} \sum_{i=0}^{J} c_i P_{e,1}^{J-i} &= \sum_{i=0}^{J} c_i = \sum_{i=0}^{J-2} c_i + c_{J-1} + c_J \\
&= \sum_{i=0}^{J-2} c_i + b_J > 0.
\end{aligned}$$

For $P_{e,1} \in (0, 1)$, we have

$$\begin{aligned}
\frac{d}{dP_{e,1}} \left(\sum_{i=0}^{J} c_i P_{e,1}^{J-i} \right) &= \frac{d}{dP_{e,1}} \left(\sum_{i=0}^{J} c_{J-i} P_{e,1}^{i} \right) \\
&= \sum_{i=1}^{J} c_{J-i} P_{e,1}^{i-1} > 0.
\end{aligned}$$

Therefore, $P_{e,1}$ has a unique solution in $(0, 1)$. Since $0 < a_i \leq 1$, $\forall\, i$, $P_{e,i}$ also has a unique solution in (0, 1). The complete solution boils down to polynomial rooting of $\sum_{i=0}^{J} c_i z^{J-i}$ in (0, 1).

For BNDMA, from (19) [BNDMA] and $P_{e,j} > 0$, $\forall\, j$ (stability is assumed), it follows that $\sum_{i=1}^{J} \lambda_i + \lambda_j < 1$, $\forall\, j$, which gives $\sum_{i=1}^{J} \lambda_i + \max_{i=1}^{J}(\lambda_i) < 1$. ■

Proof (Proposition 3): Under the conditions of Proposition 3, the average number of transmitted packets during an average length epoch is

$$\rho = mJ(1 - P_e) = \frac{\lambda J}{1 - \lambda J}. \tag{28}$$

Note that ρ is not a function of m. From the definition of relevant and irrelevant epoch lengths, we find

$$\begin{aligned}
E[l_r] &= 2 + \left(1 - \frac{1}{mJ}\right)\rho, \quad E[l_i] = 1 + \left(1 - \frac{1}{mJ}\right)\rho \\
E[l_r^2] &= 4 + 5\left(1 - \frac{1}{mJ}\right)\rho + \left(1 - \frac{1}{mJ}\right)\left(1 - \frac{2}{mJ}\right)\rho^2 \\
E[l_i^2] &= 1 + 3\left(1 - \frac{1}{mJ}\right)\rho + \left(1 - \frac{1}{mJ}\right)\left(1 - \frac{2}{mJ}\right)\rho^2.
\end{aligned} \tag{29}$$

Let $J \geq 2$ (multiuser system) and $m \geq 1$. We will show that $(\partial/\partial m) D_m < 0$, $\forall\, \lambda J > 0$.

It can be shown that

$$\begin{aligned}
\frac{\partial}{\partial m} D = -\Bigg[& \frac{\lambda^2}{2m^3} \frac{E[l_r^2]\left(E[l_r] - \frac{\rho}{mJ}\right)}{\left(1 - \frac{\lambda}{m} E[l_r]\right)^2} \\
& + \frac{\rho^2}{2m^4 J^3} \frac{1 - \frac{1}{mJ}}{\left(1 - \frac{\rho}{mJ}\right) E[l_i]^2} B(\rho) \Bigg]
\end{aligned} \tag{30}$$

where

$$\begin{aligned}
B(\rho) &= a(mJ)\rho^2 + b(mJ)\rho + c(mJ) \\
a(mJ) &= (mJ)^2 - 2mJ + 2 \\
b(mJ) &= 2mJ(mJ - 2), \quad c(mJ) = 2(mJ)^2.
\end{aligned} \tag{31}$$

From (28), it follows that $0 < \rho \leq mJ \Leftrightarrow 0 < \lambda J < (\lambda J)_{\max} = mJ/(mJ + 1)$. Since the aim is to prove that $(\partial/\partial m) D_m < 0$, $\forall\, \lambda J < (\lambda J)_{\max}$, it suffices to show that $E[l_r] - (\rho/mJ) > 0$, $(1 - (1/mJ)) > 0$ and $B(\rho) > 0$ for $0 < \rho \leq mJ$.

1) $mJ \geq 2 \Rightarrow (1 - (1/mJ)) > 0$.
2) $E[l_r] \geq 2$ and $\rho \leq mJ \Rightarrow E[l_r] - (\rho/mJ) > 0$.
3) $a(mJ) = (mJ - 1)^2 + 1 > 0\ \forall mJ \Rightarrow B(\rho)$ is always convex–U. The discriminant of $B(\rho)$ is $D_B = -4(mJ)^4$. Hence, $B(\rho)$ has no real roots, so that $B(\rho) > 0$, $\forall\, \rho$.

It follows that $(\partial/\partial m) D(\lambda J) < 0$, $\forall\, \lambda \in (0, m/(mJ + 1))$. ■

REFERENCES

[1] E. Altman, "The Baccelli–Foss saturation rule for stability for continuous time input processes," in *Proc. 32nd Allerton Conf. Commun., Contr., Comput,*. Urbana, IL: Allerton House, Univ. Illinois at Urbana-Champaign, Sept. 1994, pp. 396–403.

[2] V. Anantharam, "The stability region of the finite-user slotted ALOHA protocol," *IEEE Trans. Inform. Theory*, vol. 37, pp. 535–540, May 1991.

[3] S. Asmussen, *Applied Probability and Queues*. New York: Wiley, 1987.

[4] F. Baccelli and S. Foss. (1993) On the saturation rule for the stability of queues. INRIA Rep. 2015, Paris, France. [Online]. Available: www.nsu.ru/mmf/tvims/foss/saturat.pdf.

[5] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1992.
[6] J. I. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 505–515, Sept. 1979.
[7] R. Cruz, "A calculus for network delay, Part I," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114–131, Jan. 1991.
[8] G. Dimić and N. D. Sidiropoulos, "Multicode multicarrier random access," in *Proc. 34th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Oct. 29–Nov. 1 2000, pp. 1230–1234.
[9] ——, "Stability analysis of collision resolution protocols with retransmission diversity," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, May 13–17, 2002, pp. III-2133–2136.
[10] A. Ephremides and B. Hajek, "Information theory and communication networks: An unconsumated union," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2416–2434, Oct. 1998.
[11] R. Loynes, "The stability of a queue with nonindependent inter-arrival and service times," *Proc. Camb. Philos. Soc.*, vol. 58, pp. 497–520, 1962.
[12] W. Luo and A. Ephremides, "Stability of N interacting queues in random-access systems," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1579–1587, July 1999.
[13] R. Rao and A. Ephremides, "On the stability of interacting queues in a multi-access system," *IEEE Trans. Inform. Theory*, vol. 34, pp. 918–930, Sept. 1988.
[14] W. Szpankowski, "Stability conditions for some multiqueue distributed systems: Buffered random access systems," *Adv. Appl. Probab.*, vol. 26, pp. 498–515, 1994.
[15] M. Tsatsanis, R. Zhang, and S. Banerjee, "Network-assisted diversity for random access wireless networks," *IEEE Trans. Signal Processing*, vol. 48, pp. 702–711, Mar. 2000.
[16] B. Tsybakov and V. Mikhailov, "Ergodicity of slotted ALOHA system," *Probl. Pered. Inform.*, vol. 15, no. 4, pp. 73–78, 1979.
[17] R. Zhang, N. D. Sidiropoulos, and M. Tsatsanis, "Collision resolution in packet radio networks using rotational invariance techniques," *IEEE Trans. Commun.*, vol. 50, pp. 146–155, Jan. 2002.

Goran Dimić (S'98) received the Diploma in electrical engineering from the University of Belgrade, Belgrade, Serbia and Montenegro, in 1999 and M.S. degree from the University of Minnesota, Minneapolis, in 2001. He is currently working toward the Ph.D. degree at the University of Minnesota.

His research interests are in the area of signal processing for communications and networking.

Nicholas D. Sidiropoulos (SM'99) received the Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park (UMCP), in 1988, 1990 and 1992, respectively.

From 1988 to 1992, he was a Fulbright Fellow and a Research Assistant at the Institute for Systems Research (ISR), UMCP. From September 1992 to June 1994, he served his military service as a Lecturer in the Hellenic Air Force Academy, Athens, Greece. From October 1993 to June 1994, he also was a member of the technical staff, Systems Integration Division, G-Systems Ltd., Athens. He has held Postdoctoral (1994–1995) and Research Scientist (1996–1997) positions at ISR-UMCP, before joining the Department of Electrical Engineering, University of Virginia, Charlottesville, in July 1997 as an Assistant Professor. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, and Professor with the Department of Electronic and Computer Engineering, Technical University of Crete, Crete, Greece. His current research interests are primarily in multiway analysis and its applications in signal processing for communications and networking. He consults for industry in the areas of crosstalk cancellation and equalization for DSL modems, smart antennas, and frequency hopping communications.

Dr. Sidiropoulos received the NSF/CAREER award in 1998 and a best paper award from the IEEE Signal Processing (SP) Society in 2001. He is a member of the Signal Processing for Communications Technical Committee (SPCOM-TC) of the IEEE SP Society, has served as Associate Editor for the IEEE Signal Processing Letters from 2000 to 2002, and currently serves as Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Leandros Tassiulas was born in Katerini, Greece, in 1965. He received the B.E.E. degree from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1987 and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1989 and 1991, respectively.

From 1991 to 1995, he was an Assistant Professor with the Department of Electrical Engineering, Polytechnic University, Brooklyn, NY. In 1995, he joined the Department of Electrical Engineering, University of Maryland, where he is now an Associate Professor. He holds a joint appointment with the Institute for Systems Research and is a member of the Center for Satellite and Hybrid Communication Networks, established by NASA. His research interests are in computer and communication networks, with emphasis on wireless communications (terrestrial and satellite systems) and high-speed network architectures and management and control and optimization of stochastic systems in parallel and distributed processing.

Dr. Tassiulas coauthored a paper that received the INFOCOM 1994 Best Paper Award. He received a National Science Foundation (NSF) Research Initiation Award in 1992, the NSF Faculty Early Career Development Award in 1995, and the Office of Naval Research Young Investigator Award in 1997.