# An Optimal Q-Algorithm for the ISO 18000-6C RFID Protocol

Yael Maguire and Ravikanth Pappu

***Abstract*—Passive radio-frequency identification (RFID) systems based on the ISO/IEC 18000-6C (aka EPC Gen2) protocol have typical read rates of up to 1200 unique 96-bit tags per second. This performance is achieved in part through the use of a medium access control algorithm, christened the Q-algorithm, that is a variant of the Slotted Aloha multiuser channel access algorithm. We analyze the medium access control algorithm employed by the ISO/IEC 18000-6C RFID air interface protocol and provide a procedure to achieve optimal read rates. We also show that theoretical performance can be exceeded in many practical use cases and provide a model to incorporate real-world data in read-rate estimation.**

***Note to Practitioners*—Estimating read-rates in RFID has always been something of a black art. At one end of the spectrum, in the pure-theory approach, rates are estimated by taking the duration per bit and calculating the total number of bits that can be decoded per second. This approach does not take any of the protocol overheads or real-world conditions into account. In the pure-experimental approach, a standard test case is used to relatively compare read-rates as several factors—tags, readers, firmware, protocols, etc., are varied. Neither of these approaches really provides any insight into the problem of estimating read rates for the general case.**

**In this paper, we take on this problem by developing a first-principles model of collision probability in the Gen2 medium access control layer. Collisions of tag responses are a dominant factor in determining read rates in Gen2 systems. Using this model, we show that the worst case efficiency of the protocol can be no less than 36.8%, i.e., it should be possible to see more than 36.8% of a given population of tags per unit time. We them develop a dynamic Q-algorithm that performs much better than the worst case, and show its performance relative to a static Q-algorithm.**

**We then relax the assumptions underlying the above algorithm so as to be able to incorporate real-world situations and provide a framework wherein practitioners can make some measurements of a particular situation and use our model to estimate expected read rates. Three important factors that need to be considered are: (i) the different decoding times for different types of slot-occupancy; (ii) the capture effect, wherein a two-occupancy slot is decoded as a valid tag because the backscatter powers are sufficiently different; and (iii) the distribution of backscatter powers. We develop a model to account for these three factors.**

**Although our models make several assumptions, we have designed and deployed readers that justify almost all of them. We are currently working on developing a deeper characterization of the backscatter power distribution of a population of tags. This will allow us to use the signal processing capability of our readers to disambiguate two-occupancy slots and boost read rates well-above those predicted by our model. This is the focus of our current research.**

***Index Terms*—Multiple-access, anti-collision, Gen2, ISO 18000-6C, analysis, optimal, radio-frequency identification (RFID).**

## I. Introduction

THE RATIO OF terrestrial radio and cellular telephone systems to the number of humans on earth is approaching unity and, in the past decade, a completely different kind of radio device has emerged and is poised to eclipse this ratio by three orders of magnitude. Rapid advances in CMOS technology have enabled the production of low-cost *tags* that are capable of reporting identity over a wireless link. These low-cost tags—usually costing on the order of tens of cents—are typically composed of a few thousand gates of silicon, and have little, if any, general purpose computing power available to them beyond responding to commands from an interrogator or *reader*. This asymmetry between the interrogators and the tags is further amplified by the fact that, in many applications, tags are passive, i.e., they do not have an on-board source of power; rather, they obtain it by harvesting power from the electric, magnetic or electromagnetic field generated by the interrogators.

There are a large variety of medium access control (MAC) methods employed in multistation communication systems today. Each of these methods makes certain assumptions—computing power, frequency selectivity, ability to sense a carrier or collision in the channel, clock synchronization, etc., about the capabilities of the participating stations. Viewed from this perspective, a typical passive Radio-frequency identification (RFID) scenario consisting of several tens or hundreds of tagged cases arrayed on a pallet represents tens or hundreds of stations within a few feet of each other with no spare computing resources, frequency selectivity, or the ability to synchronize with other stations. Another challenge arises from the fact that the population of stations is moving in space and can fade in and out of visibility depending on whether or not the station is able to receive power from the interrogator.

This paper analyzes the efficiency of the MAC layer ISO/IEC 18000-6C RFID air interface protocol, hereafter referred to as the Gen2 protocol, or simply Gen2 [1]. It is organized as follows. In Section II, we provide an overview of the application context where our efficiency analysis is most relevant. In Section III, we review the ALOHA MAC protocol, on which the Gen2 MAC layer is based. Then, we review the assumptions and the operation of the Gen2 MAC protocol and introduce the terminology and notation used in the remainder of this paper.

Manuscript received July 29, 2007. First published April 04, 2008; current version published December 30, 2008. This paper was recommended for publication by Associate Editor M. Marlin and Editor N. Viswanadham upon evaluation of the reviewers' comments.

The authors are with ThingMagic, Inc., Cambridge, MA 02142 USA (e-mail: yael@thingmagic.com; ravi@thingmagic.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASE.2008.2007266

TABLE I
A SAMPLING OF PASSIVE RFID USE CASES

| No. of tags | Duration in field | Speed | Use case |
|---|---|---|---|
| unknown | unknown | fast | Reading pallet tags |
| unknown | unknown | slow | Supermarket checkout |
| known | known | fast | Airport baggage conveyor |
| known | known | slow | Library book checkout |

In Sections V–VII, we formulate the efficiency mathematically, derive the optimal Q-algorithm, and present results.

## II. Application Contexts

Recent developments in RFID technology and corresponding international standards [1] have spurred the deployment of passive systems in applications ranging from inventory management of consumer packaged goods to tracking medical equipment in hospitals to counting poker chips on gaming tables. There are several different application contexts that are addressed by passive RFID systems. Each of these contexts might place a distinct performance requirement on the RFID system. As one way of categorizing these contexts, consider the following matrix in Table I.

The first row of Table I describes a scenario where there is an unknown (but bounded) number of tags in the interrogation field of a reader for an *a priori* unknown duration. These tags are moving through the read field rapidly. Such a situation might occur in a retail distribution center where a pallet of tagged cases is being transported on a forklift through a portal where readers are located. *In this scenario, the reader is required to read as many tags as possible in the window when tags are visible to the reader*. This duration is a function of both the beamwidth of the reader antenna as well as the velocity of the forklift that is moving the pallet through the field. The duration is usually unknown because the forklift typically moves through the read field at an unknown velocity between one and six miles per hour. In the second row, we have a scenario that might occur at the checkout counter of a supermarket. The number of tags is still unknown, whereas the (human) shopper is moving through the read field much more slowly. In this case, the number of tags is typically lower than in the previous case, and the available duration is higher. The use case described in the third row is fundamentally different from the both the first and second cases. Here, the number of tags is known *a priori*, as is the duration that is available to the reader to read the tag. However, because the tags are moving at high velocities, the duration is extremely short. As an example, with a conveyor that is moving single packages at 3 m per second (approximately 600 feet per min), a reader with a read field width of 1 m has about 330 ms to successfully read a tag. Here, the goal of the reader is to guarantee a single read in a short duration. This goal requires a different reading strategy than that for the first case.

Our analysis in this paper will focus on the first case presented in Table I—*the reader is required to read an unknown number of tags as fast as is allowed by the protocol.*

## III. The ALOHA MAC Protocol

The ALOHA protocol was the first system to employ broadcast radio communications to allow simultaneous computer and user communications [2]. In this protocol, the simplest version has a number of users with access to a shared broadband channel. Any user may communicate at any time. If a user does not receive an acknowledgement within a reasonable amount of time, it will consider the receipt of that packet unsuccessful and retransmit the packet. Making the assumption that user transmission follows a Poisson traffic model, the expected rate of traffic of data per interval $\tau$, $R$, with collisions that occur between any two users over $2\tau$ intervals is

$$C = R \exp(-2R). \tag{1}$$

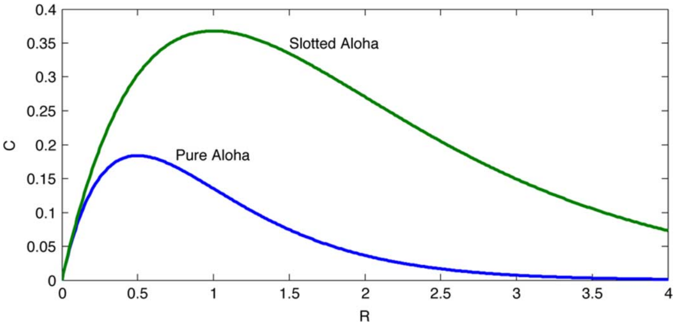

Fig. 1. Capacity for the two different types of multiuser wireless communications protocols using ALOHA. Without carrier sense and collision detection, one cannot achieve the Shannon bound of 1.

The capacity of this pure Aloha model is maximized when $R = 1/2$ and the maximum capacity is $1/2e$. For the case where specific time intervals for communication are delineated or slots are created, collisions that occur between any two users is restricted to a single time interval $\tau$. Therefore, the capacity is

$$C = R \exp(-R). \tag{2}$$

The capacity is maximized when $R = 1$ and the maximum capacity is $1/e$, which is double the pure ALOHA case. A plot of these is shown in Fig. 1.

## IV. Factors Affecting Gen2 Performance

### A. A Brief Review of the MAC Subsystem of the Gen2 Protocol

The Gen2 protocol is fully specified in [1]. In a typical scenario, one interrogator remotely powers up a population of $T$ tags and engages in an anti-collision protocol to disambiguate them and read their unique identities and any data payload stored in the tags. An interrogator manages a population of tags using three basic operations—select, inventory, and access. Selection is the process by which an interrogator divides all available tags into different subpopulations. The inventory process, which is the focus of our interest in this paper, enables the interrogator to identify all tags within a given subpopulation. Finally, access commands allow an interrogator to complete specific transactions like reading and writing identity and data with a particular tag.

We will use Fig. 2 to introduce several important terms and concepts in the Gen2 protocol. The protocol begins with a *Select* command that restricts the interrogator's attention to a given population of tags. After time $T_4$ of selection, the interrogator is required to issue the all-important *Query* command. Embedded in this command is a parameter $Q$ which can take on values from

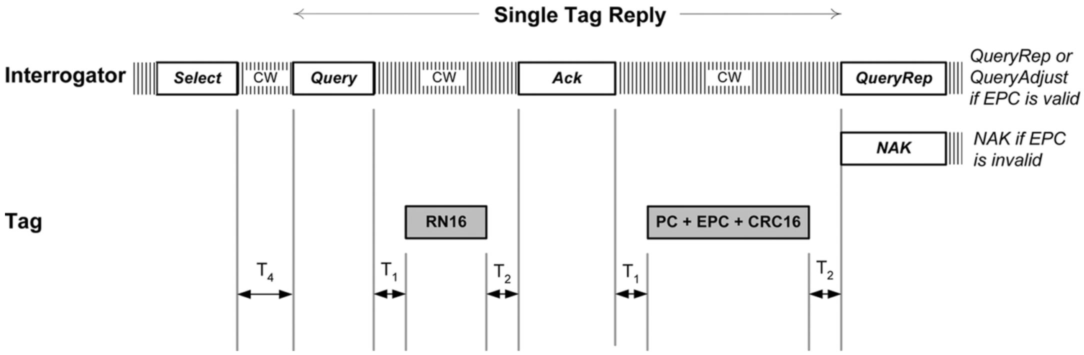

Fig. 2. Timeline of operations involved in identifying a single Gen2 tag. This figure is reproduced from [1].

0 through 15. The role of $Q$ in determining the efficiency of the protocol is more fully described below. If $Q = 0$ and there is only one tag in the field of the interrogator, the tag responds to the query by backscattering an RN16, a uniformly distributed 16-bit random number, within time $T_1$. The interrogator broadcasts the same RN16 within time $T_2$. Once the tag which sent the RN16 receives this acknowledgment, or ACK, it backscatters its ID, labeled $\mathrm{PC} + \mathrm{EPC} + \mathrm{CRC}$ in Fig. 2, which is then decoded by the reader. If, instead of $Q$ being 0, it had a finite value $Q = n$, then the single tag in the field responds to a *Query* in one of $2^n$ *slots* with uniform probability of selecting a slot. Assuming our tag selected a slot counter value other than 0 (this is a uniformly distributed random number, see V below), it needs to receive a *QueryRep* command from the interrogator to cause it to decrement its slot counter by one. This process continues until the slot counter is 0, after which the procedure described in the preceding paragraph is followed. Since passive tags are very simple devices without the ability to detect a collision in the channel and backoff (i.e., CSMA/CD [3]), *a priori* random slot selection is a substantially lower cost and power alternative to implementing a multiple-access mechanism.

The situation is much the same with multiple tags in the field, with one important difference. Since tags have no way of coordinating their responses with each other, there is now a finite probability of tags selecting the same slot, i.e., their responses collide in the channel. It is up to the interrogator to detect the collision and issue a new *Query* or *QueryAdjust* command to resolve this collision. The latter command simply increments or decrements the value of $Q$, thereby doubling or halving the number of slots available to the tags.

The correct choice of $Q$ in any round of the protocol is critical to achieving optimal read rates. Intuitively, if the number of slots relative to the number of selected tags is very high, then the probability that there is collision in any given slot is low, and the number of correctly decoded tags is close to the number of selected tags. The drawback here is that there are many empty slots, which leads to wasted time. On the other hand, if the number of slots relative to the number of selected tags is low, then lots of collisions can be expected, which leads to more rounds being required to complete the inventory process. The challenge, then, is to develop an algorithm to control $Q$ that (a) minimizes the total time required to inventory a population of tags by minimizing the number of empty slots and collided slots and (b) gracefully handles a dynamic population of tags. We tackle this problem in the next section.

### B. Other Important Factors Affecting Gen2 Read Rates

Although the correct choice of $Q$ is a crucial factor in achieving high read rates with unknown populations of tags, it is by no means the only factor. We digress here briefly to enumerate other factors that play important roles in determining overall system performance. We look at these factors using the following three scenarios.

1) *One reader, one tag*: Here, we assume that only one reader and one tag are present. The tag might be affixed to some other object of unknown electromagnetic characteristics and is assumed to have a 0 dBiL isotropic antenna. We assume that there is no interference from other readers or other devices in the reader's frequency band. In such a scenario, read performance is dominated by two fundamental limits. The *forward link limit* is the minimum power required to keep a passive tag powered up and communicate with it. Today's Gen2 tags require between 25 and 100 $\mu$W to reliably power up. Experimental passive tags with power-up thresholds as low as 12.6 $\mu$W have been reported recently in the microwave band [4]. Noting that power transmitted from the reader drops off at $1/r^2$, where $r$ is the read range, the forward-link limit determines the maximum range at which a passive RFID tag can be reliably read.

   The other important limit in the one-reader, one-tag scenario is the *return link limit*. Put simply, this limit is determined by the ability of the reader to reliably distinguish the tag's backscatter signal from its own noise floor. Since a reader's noise floor is implementation dependent, we will instead provide a lower fundamental limit, which is determined by the thermal noise power received at the receiver front end. The thermal noise power at the input to the receiver is given by

$$P_{\mathrm{in}}\,(\mathrm{dBm}) = 10\log(k_B T B/1\ \mathrm{mW}) \qquad (3)$$
$$= -173.8\ \mathrm{dBm} + 10\log(B) \qquad (4)$$

   where $k = 1.381 \times 10^{-23}$W/Hz/K is Boltzmann's constant, $T = 300°$K is the room temperature, and $B = 128$ kHz is the double-sided bandwidth of a tag response for a Miller backscatter with $M = 4$. The means that the noise floor is approximately $-123$ dBm. The signal-to-noise ratio (SNR) required for successful decoding of 128 bits is about 11.7 dB above this for orthogonal, ASK signals. A practical receiver does not have an effective noise temperature of 300 K, but a higher value given by the analog noise figure and the implementation margin relative to a perfect demodulator. Noise figures are usually 15–30 dB above this fundamental limit, meaning that successful tag reading must produce backscatter at the receiver above about $-96$ dBm. If the tag backscatter power received is lower than this limit, the tag will not be successfully decoded. As of this writing, well-designed passive RFID systems are forward link limited.

2) *One reader, many tags*: Here we assume that several tags are added to the one-reader, one-tag system. This is the primary case that we will be dealing with in this paper. We will assume that the resulting system is neither forward link limited, i.e., all tags in the field are powered up reliably, nor reverse link limited. We also assume that the signal processing system in the reader can decode any tag that is powered up. We have discussed this case in detail in Subsection IV-A.

3) *Many readers, many tags*: Finally, we describe the most general case that occurs in practice—many simultaneously operating readers, and many groups of tags passing through readers' fields of view at arbitrary velocities and with arbitrary durations. Such a scenario might be observed at the distribution center of many major retailers, where readers at inbound dock doors are continuously scanning for tags. Read performance in this scenario is affected by several factors. First, the objects which are tagged might be so-called *RF unfriendly* objects, i.e., they absorb or reflect electromagnetic radiation, leaving the tags unable to be powered up. A second important factor is reader–reader interference [5]. This occurs when several readers are transmitting simultaneously and are visible to other readers in the vicinity. This causes the effective noise power at the input to the reader to be increased substantially, thereby causing tag backscatter signals to be drowned out. This effect can be mitigated by designing reader front ends with sufficient filtering capability to reject co-channel and adjacent channel jammers. Finally, there is the phenomenon of reader-tag interference. Existing Gen2 tags get confused when two or more readers address them simultaneously at carrier frequencies that are less than approximately 1 MHz apart [6]. The primary reason for this effect is that the tags does not possess any degree of frequency selectivity. While this is not a reader issue *per se*, it is certainly a factor in lowering read rates in the many reader, many tag scenario.

## V. Assumptions and Notation

We make the following assumptions in our analysis.

- Tags select their slots with a uniform probability in the range $[0, 2^Q - 1]$. This assumption is reasonable in practice because tags are required to select $Q$-bit subsets from the RN16s, which are required to be almost uniform with the probability of any particular RN16 being generated is between $0.8/2^{16}$ and $1.2/2^{16}$. As an example, consider Fig. 3 which clearly shows that commercially available tags meet this requirement.

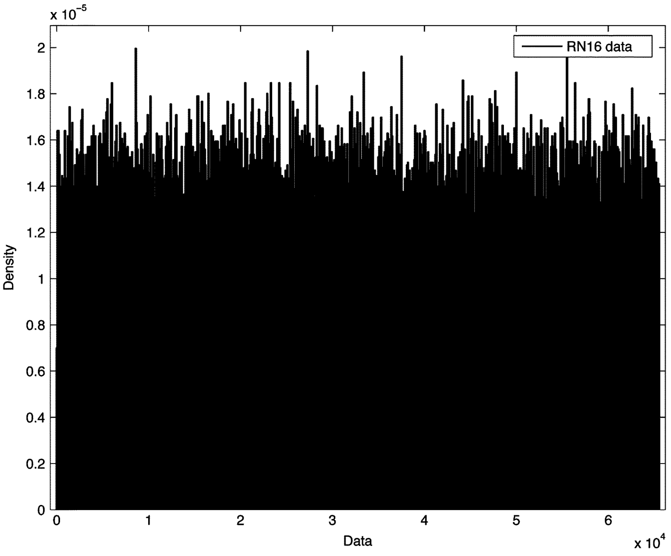

Fig. 3. Histogram of $2^{16}$ RN16 values obtained from a commercially available Gen2 tag showing that the random number generator in this tag is producing uniformly distributed random numbers.

- The bit error rate in decoding backscattered signals is zero. This assumption implies that any tag that can be powered up is decoded with probability unity. This assumption, although counterintuitive, is also reasonable in practice, because typical RFID deployments tend to be forward-link limited. The primary challenge is to deliver enough power to the tags to keep them alive and participating in the protocol. When this challenge is overcome, decoding the response is usually accomplished with unity probability.
- All types of slots—empty, singly occupied, and multiply occupied—require the same time to decode. *This assumption allows us to deal with time in discrete units of slots*. This assumption is usually violated in practice because interrogators might choose to deal with each of these types of slots differently. One possible approach is to determine slot occupancy by measuring return signal power in each slot and making a decision to process or abandon the return signal in the slot based on empirical measurements. Given this, our analysis will yield a lower bound on the efficiency.
- The interrogator will not end a round prematurely. The Gen2 protocol allows the interrogator to abandon a round at any given time, which can be used in practice to improve read rates when the number of tags is not known *a priori*. This will not be used in this analysis.

We use the following notation in this paper.

- A slot with $k$ responses is referred to as a $k$-occupied slot.
- The expected number of $k$-occupied slots when there are $T$ tags present and the number of slots is $s = 2^Q$ is denoted by $E_k(T, s)$. Where there is no chance of confusion, we will abbreviate $E_k(T, s)$ to simply $E_k$.
- The time required to decode or decide not to decode a $k$-occupied slot is denoted by $t_k$.

## VI. Our Analysis

The metric we are trying to optimize is to maximize the number of reads per second i.e., the number of 1-occupancy slots per unit time. A general expression for the expected number of $k$-occupancy slots is derived in Appendix I and reproduced below for convenience.

$$E_k(T,s) = \binom{T}{k}(1/s)^{(k-1)}(1-1/s)^{(T-k)}. \tag{5}$$

For $k = 1$, (5) reduces to

$$E_1(T,s) = T(1-1/s)^{(T-1)}. \tag{6}$$

Let us denote the number of of 1-occupancy slots per unit time by $\mathcal{L}(T,s)$. This is given by

$$\mathcal{L}(T,s) = \frac{E_1(T,s)}{t_0 s} = \frac{E_1}{t_0 s} \tag{7}$$

where $t_0$ is the time per slot. Our goal is to find an $s = \hat{s}$, and thereby a $Q = \log_2(\hat{s})$, such that $\mathcal{L}(T,\hat{s})$ is maximized. Thus, our optimization problem is succinctly stated as

$$\hat{s} = \arg\max_s \mathcal{L}(T,s) \tag{8}$$

To solve this, we recognize that the logarithm of a function is monotonic with the function as long as the function is always positive. This is true in our case, so we maximize the logarithm of (8)

$$\begin{aligned}\log\mathcal{L}(T,s) &= \log Tt_0 - \log s + (T-1)\log(1-1/s)\\ \frac{\partial \log\mathcal{L}(T,s)}{\partial s} &= -1/s + (T-1)1/(1-1/s)(1/s^2) = 0\\ &\implies \hat{s} = T. \end{aligned} \tag{9}$$

Note that $\hat{s}$ is independent of the time required to decode a slot. Substituting this solution into (7), the efficiency $\mathcal{L}$ becomes

$$\mathcal{L}(T,\hat{s}) = \frac{(1-1/T)^{(T-1)}}{t_0}. \tag{10}$$

As $T \rightarrow \infty$, $\mathcal{L}(T,\hat{s}) \rightarrow 1/t_0 e \approx 0.368 t_0^{-1}$. Without loss of generality, we can assume $t_0 = 1$ (i.e., treat time in units of slots). From (10), it is clear that 36.8% of all slots yield decoded tags. This result is identical to that achievable with the slotted ALOHA protocol [2], where the number of slots is infinite. Finally, as shown in Fig. 4, (10) converges rapidly to $1/e$. This implies that for any number of tags $T > 20$, a well-designed anti-collision algorithm should read all tags in no more than 2.72 $T$ slots on average.

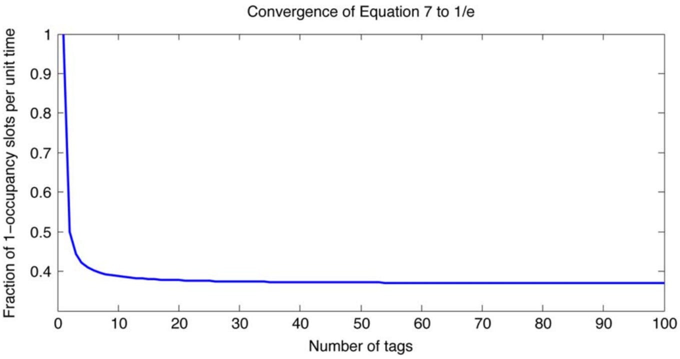

Fig. 4. This plot shows the rapid convergence of (10) to $1/e$. This implies that for any number of tags $T > 20$, a well-designed anti-collision algorithm should read all tags in no more than 2.72 $T$ slots on average.

## VII. Optimal Q-Algorithm

### A. Theory

We will now use the result from the previous section to calculate the best $Q$ value for a round and describe a recursive algorithm to read all available tags in the shortest amount of time. Given a total of $T$ tags in a population or an expected number of tags $\langle t\rangle = T$, since an interrogator knows how many tags are read per round, we can describe an optimal recursive algorithm to determine the value of $Q$ for each round.

1) Count the number of tags read in the current round, $r_n$ and set this value to $Tr_n$. The first round is $r_1$.
2) Set $T_{n+1} = T_n - Tr_n$. The seed value for this difference equation is the known number of tags to be read or the number of tags expected to be seen $T_1 = T$.
3) For round $r_n$, set $Q_n = \lfloor \log_2 T_n \rceil$ and have the interrogator run the round with this $Q_n$ value.
4) Repeat steps (1) though (3) until $Tr_n = 0$.

Given the results from the previous section, the value of $Q_1$ will be chosen as the optimal value for round $r_1$ given that the number of tags, $T$ is known. If the number of tags which singly occupy a slot are read with $\mathrm{BER} \rightarrow 0$, then at the end of the round, the number of total tags in the field $(T - Tr_1)$ is still known and, therefore, the subsequent round, $r_2$ will also have the optimal value of $Q_2$. By recursion, at the end of any round $n$, the total number of tags is known and hence the choice of $Q_{n+1}$ (from (9)) will always be the optimal choice.

### B. Results

In this section, we provide results of applying the algorithm in Section VII-A to a population of tags on a pallet of consumer packaged goods such as cooking oil, peanut butter, soap, and other marvels of the consumer age. The total number of tags in the pallet is unknown, and the reader is set up to read for a fixed duration of 390 ms. This time is chosen so as to be below the 400 ms threshold for frequency-hopping as specified by the U.S. Federal Communications Commission. Setting the time to 390 ms removes frequency as a variable in the read-rate performance of the reader. Additionally, the query command is set up to cause the tags to respond exactly once within 390 ms, in order to ensure that multiple reads of the same tag do not artificially inflate the read-rate. Finally, we expect to see no more than 97 tags in 390 ms because each tag requires approximately 4 ms to be successfully decoded.

Several experiments were carried out. The first set of experiments set the value of $Q$ to be fixed at one of 2, 3, 4, 5, 6, 7. For each $Q$, a single read of duration 390 ms was performed and the slot occupancy of each slot (i.e., zero, single, or multiple) was

TABLE II
STATIC Q

| Q | Zero | Single | Multiple | Slots | Efficiency |
|---|---|---|---|---|---|
| 7 | 84 | 79 | 2 | 165 | 0.479 |
| 6 | 23 | 86 | 11 | 120 | 0.717 |
| 5 | 7 | 78 | 32 | 117 | 0.667 |
| 4 | 2 | 46 | 90 | 138 | 0.333 |
| 3 | 0 | 19 | 136 | 155 | 0.123 |
| 2 | 2 | 5 | 155 | 162 | 0.031 |

TABLE III
VARIABLE Q WITH $Q_{\min} = 2$ AND $Q_{\max} = 7$

| $Q_{init}$ | Zero | Single | Multiple | Slots | Efficiency |
|---|---|---|---|---|---|
| 6 | 59 | 82 | 6 | 147 | 0.558 |
| 5 | 25 | 86 | 9 | 120 | 0.717 |
| 4 | 7 | 72 | 41 | 120 | 0.600 |
| 3 | 6 | 76 | 35 | 117 | 0.650 |
| 2 | 1 | 71 | 45 | 117 | 0.601 |

recorded. The second set of experiments allowed $Q$ to vary between $Q_{\min} = 2$ and $Q_{\max} = 7$. The initial value of $Q = Q_{\text{init}}$ was set up to be one of 2, 3, 4, 5, 6. The results are shown in Tables II and III

In both tables, we compute the efficiency as the fraction of singly occupied slots. Some observations from the above results follow.

- For the static Q case, maximum performance is attained at $Q = 6$, as opposed to the intuitively appealing value of $Q = 7$. $Q = 7$ allows 128 slots, which would seem to be more than sufficient to accommodate 97 tags. However, as is clear from the table, the number of 0-occupancy slots is about 50% of the total number of slots—lots of wasted time. Better performance is obtained at $Q = 6$, while $Q = 5$ yields similar results as $Q = 7$, for a different reason. In the $Q = 5$ case, the number of multiply occupied slots is higher than in the $Q = 7$ case, while the number of 0-occupancy slots is substantially lower.
- The efficiency $Q = 2, 3, 4$ is fairly low resulting in a lower percentage of reads. This implies that the number of tags read is much lower than expected. While this could be due to several reasons, for the *one-reader, many-tag* case that we are considering in this paper, it is due to a very large fraction of collisions as is seen in the multiple-occupancy column.
- For the variable $Q$ case, we note that the performance if the optimal algorithm is clearly better than that of the static $Q$ case. This is illustrated in Fig. 5.

### C. Extending the Optimal Algorithm

In practice, the algorithm in Section VII-A is not always optimal because of the following relaxation of assumptions.

- In some RFID applications, the total number of tags is not known in advance.
- In passive RFID, not all tags in the field will be powered at a given time. Due to the location of the tags on materials, the pathloss from reader antenna to the tag may exceed the power-up threshold of the tag. Furthermore, this threshold may be frequency dependent due to multipath fading. These issues usually contribute to lack of knowledge of the total number of tags in the field.

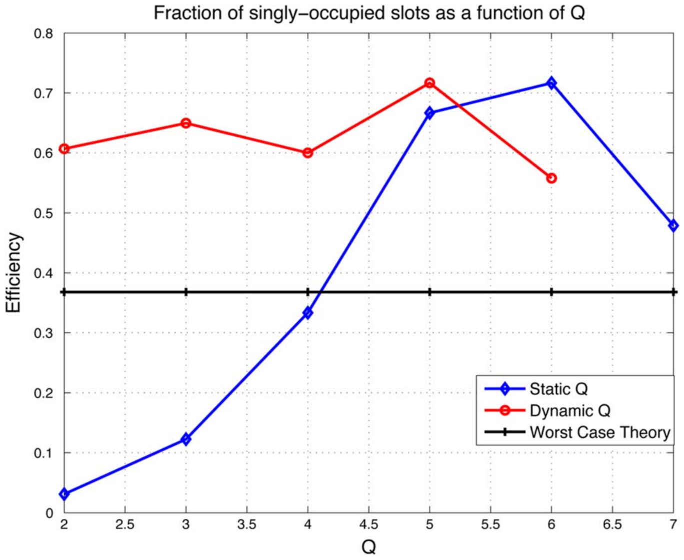

Fig. 5. This plot compares the efficiencies of the static and dynamic Q algorithms. The horizontal line at 0.368 is the worst-case theory value derived in (9).

- The lengths in time of the different types of slots are not equal. An efficient RFID interrogator can usually estimate whether a slot is empty, singly occupied or multiply occupied. If a slot is empty or collided, an interrogator only needs to wait a time $T_1$ (shown in Fig. 2) before commanding the population of tags to decrease their slot counter.
- With a wireless protocol, multiple occupancy slots can still result in the tag being read. Tag backscatter signals are usually received with differing powers at the interrogator's receive antenna depending on their pathloss relative to the antenna. An interrogator can determine the RN16 of the strongest tag and request that it backscatter the $\mathrm{PC} + \mathrm{EPC} + \mathrm{CRC}$. This phenomenon is known as the *capture effect* and is discussed in [7].

Given the last two conditions, a new efficiency metric to include these effects must be defined. Let us denote by $a_k$ the probability that a $k$-occupancy slot is successfully decoded as a single tag by an interrogator. Note that it is not important which of the $k$ responding tags was decoded, just that one of them was successfully decoded. For example, if $a_2 = 0.1$, it means that 10% of all two-occupancy slots are successfully decoded by the interrogator. Also, from our assumptions above, $a_1 = 1$.

We can then define an extended efficiency $\mathcal{L}_{\text{ext}}$ as

$$\begin{aligned}\mathcal{L}_{\text{ext}}(T, s)\\ &= \frac{a_1 E_1 + a_2 E_2 + \dots}{t_0 E_0 + ((1 - a_1) t_2 + a_1 t_1) E_1 + \dots}\\ &= \frac{\sum_{k=1}^{T} a_k E_k}{t_0 E_0 + t_1 \sum_{k=1}^{T} a_k E_k + t_2 \sum_{k=1}^{T} (1 - a_k) E_k}. \end{aligned} \tag{11}$$

The times $t_k$ are the times required to decode $k$-occupied slots. We assume that $t_m = t_n$ for $m, n \geq 2$. This is a reasonable assumption because it does not make sense for the interrogator to expend cycles distinguishing various types of collided slots. Referring to Fig. 2, we note that an optimal interrogator

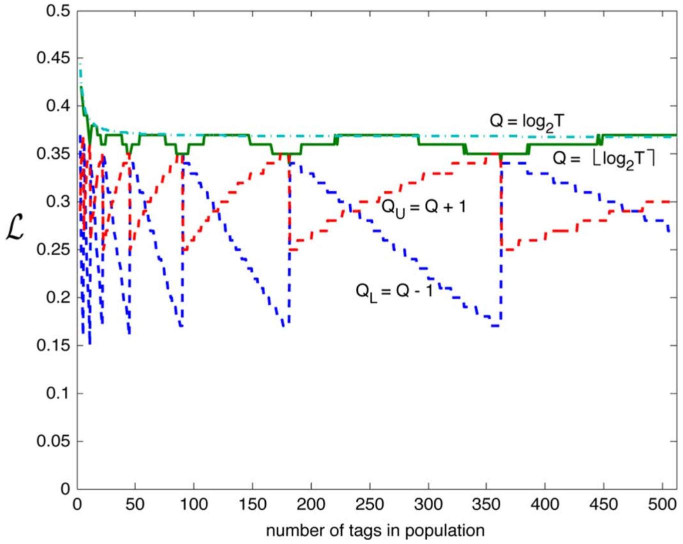

Fig. 6. Plot of the efficiency versus number of tags using the formula in (9), a rounded version of this (required for the Gen2 protocol) and the integers that bound it. This figure shows that the choice of $Q$ calculated is optimal when all types of slots (empty, single occupancy, multiple occupancy) are equal in time. Note that there is some loss of efficiency due to the Gen2 protocol requirement that a round must have a number of slots equal to a power of 2.

will have $t_0 = t_2 = \min T_1$ and $t_1 = 2\min T_1 + 2\min T_2+$ time to backscatter the $\mathrm{RN16} + \mathrm{PC} + \mathrm{EPC} + \mathrm{CRC}$. The actual value of $\min T$ depends on the particular mode of operation of Gen2 and is fully specified in [1]. Further, note that (7) is a special case of (11), where $a_k = 0$ for $k \geq 2$ and $t_0 = t_1 = t_2 = 1$.

The optimization problem is the same as before [(8)]

$$\hat{s} = \arg\max_{s} \mathcal{L}_{\text{ext}}(T, s). \tag{12}$$

Since $\mathcal{L}_{\text{ext}}$ is a fractional sum of exponential terms, (11) does not admit a simple closed form analytical solution. However, (11) provides the ability to incorporate real-world measurements of tag-backscatter power into a read-rate calculation and thereby provides an empirical predictor of expected read rates in various use cases. The next section uses numerical analysis and real-world data to look at this (11) in more detail.

### *D. Simulation Results of the Extended Algorithm*

Fig. 6 shows a plot of a numerical simulation of $\mathcal{L}(T, s)$ for various values of $T$ and an $s$ value calculated from (9). Simulated tags were randomly associated with a slot and the probability of being able to read tags was 1.0 for 1-occupied slots and 0.0 for $k$-occupied slots for $k \geq 2$. Ten thousand runs for each value of $T$ were executed to keep the variance on the estimate of the efficiency low. Two other curves were calculated for $Q$ values one higher (i.e., $Q_U = Q_{\max} + 1$) and one lower (i.e., $Q_L = Q_{\max} - 1$) than (9). It is readily apparent that using a rounded approximation of (9) is a good approximation to the optimal solution up to the quantity of tags shown here, 512. The limit on the number of tags were chosen to represent the maximum number of tags which would conceivably be in the read field of an interrogator in a given protocol round. Any loss of efficiency is due to the coarseness of trying to make the number of slots in a round a power of 2.

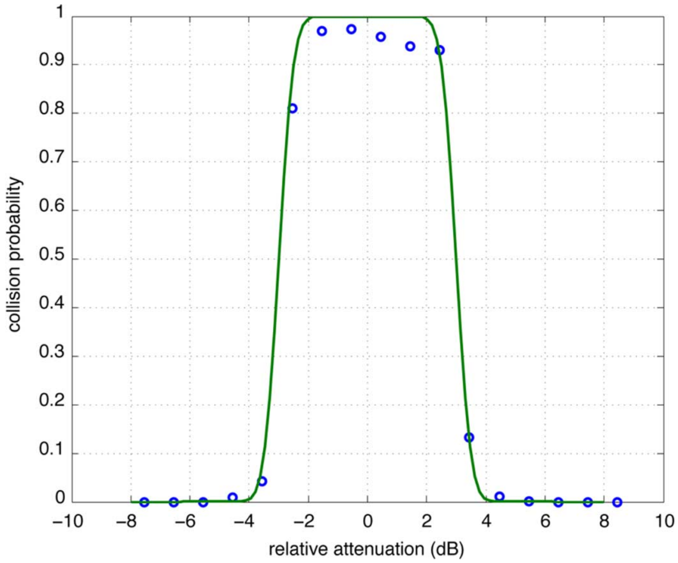

Fig. 7. Demonstrating the capture effect in passive RFID. Plot of the measured collision probability, $p_c$ versus relative attenuation between two tags in a box communicating with a UHF RFID reader. The circles are actual data points and the solid line is a curve fit.

Investigating the case when multiply occupied slots can be read with nonzero probability, consider (11) with $t_0 = t_1 = t_2$. It can be simplified to yield

$$\mathcal{L}_{\text{ext}}(T, s) = \frac{\sum_{k=0}^{T} a_k E_k}{t_0 s} = \frac{E_1}{t_0 s} + \frac{\sum_{k=1}^{T} a_k E_k}{t_0 s} \tag{13}$$

because $\sum_{k=0}^{T} E_k = s$. This is definitely greater than $(E_1)/(t_0 s)$ from (7). Therefore, the efficiency of reading can be improved over $\approx 1/e$ if tags can be read in collided slots (i.e., captured). Intuitively, this should be possible if the margin in backscatter power of a single tag over the rest of the population exceeds an implementation-dependent threshold.

*Measuring $a_2$:* An experiment was conducted to get a feel for the magnitude of the signal difference between two tags to observe the capture effect. We measured the receive signal strength of an orthogonal ASK Miller decoder to be able to resolve a single tag response from two that are simultaneously responding (i.e., to look at $a_2$). Two tags were individually placed in isolated metal boxes with near-field antennas and a variable attenuator was included in the path to one tag box to be able to systematically vary the relative power between the tags from the perspective of the reader receiver. The data from this experiment is shown in Fig. 7. For a relative attenuation of about 3–4 dB, the receiver is able to successfully distinguish one tag's RN16 from the other. This experiment gives us a feel for the capture effect.

One more piece of experimental information is required to be able to calculate $a_k$, that being the distribution of tag power in a specific application. This distribution will have moments that depend on the material tags are adhered to, the environment and the pathloss of the reader to the tag population. If we assume each tag has a identical and identically distributed (i.i.d.) probability distribution function $p_{p,2}(\alpha)$, centered at a mean power, $\mu$, with respect to received signal power $\alpha$

$$a_2 = \int_A (1 - p_c(\alpha)) p_{p,2}(\alpha - \mu)\, d\alpha \tag{14}$$

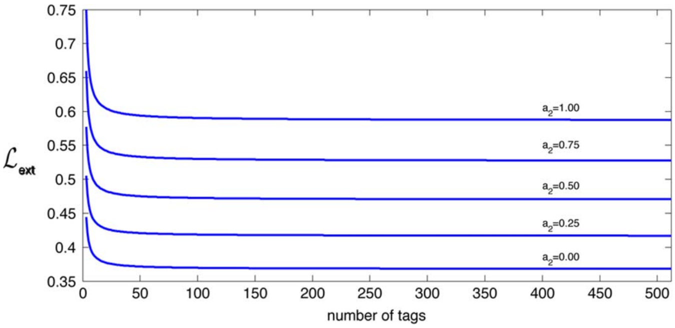

Fig. 8. Efficiency curves $\mathcal{L}_{\text{ext}}$ from (15) showing that the ability to resolve tags during double occupancy slots can improve efficiency from less than 40% to close to 60%. The lowest curve is identical to the curve from Fig. 4.

for coefficients $a_k, k > 2$, these coefficients will be determined by higher order integrals. For the purposes of illustration, let us consider the case where double occupancy has a nonzero $a_2$, while all higher order coefficients are equal to 0. That is

$$\mathcal{L}_{\text{ext}} = \frac{E_1 + a_2 E_2}{t_0 s}. \tag{15}$$

A plot of this efficiency for various values of $a_k$ is shown in Fig. 8. As can be seen, a significant improvement in efficiency can be achieved by simply being able to decode two-occupancy slots. Therefore, from an interrogator-design perspective, designing in sufficient signal-processing capability to be able to successfully decode at least a two-occupancy slot offers improvements in efficiency of up to 50%.

## VIII. Conclusion

Our analysis of the efficiency of the MAC subsystem ISO/IEC 18000-6C RFID air interface showed that an algorithm exists to achieve optimal read rates from a population of RFID tags. Starting from a general slot occupancy probability distribution function, we defined the optimization problem for maximizing read rates and calculated the analytical solution. This solution converges to the Slotted ALOHA result when the number of slots approaches infinity. One source of practical loss of efficiency was the requirement of the Gen2 specification to specify slot lengths be equal to a power of two rather than any integer. In light of practical optimizations, we generalized the efficiency function to include unequal times for different slot occupancies afforded by the Gen2 specification. A further generalization was considered to include the fact that typical wireless systems exhibit the capture effect. From this observation and experimental support, the efficiency function was evaluated to show that at least a 50% improvement in efficiency can be achieved by reading tags in collided slots. Methods of disambiguating two-occupancy slots to successfully decode at least one tag merit further investigation, and are the subject of our future research efforts.

## Appendix I
## Derivation of $E_k(t, s)$

The general expression for $E_k(t, s)$ may be derived by considering a ball and urn model [8] wherein $t$ tags are randomly distributed into $s$ slots. We are asking for the probability that a slot is $k$-occupied. Start with a definition of $E_1(t, s)$

$$E_1(t, s) \equiv \langle \text{single occupancy}(t, s) \rangle \tag{16}$$

where the expected value is denoted by $\langle\rangle$. Use linearity of expectation [9] to get

$$E_1(t, s) = s \times p(\text{urn 1 has exactly one tag in it}). \tag{17}$$

This probability distribution $p$ is given by a binomial distribution [10]. There are $\binom{t}{1}$ ways of selecting a tag to occupy a slot. The probability that the slot will be selected by one tag and passed over by $(t-1)$ tags is $(1/s)(1-1/s)^{(t-1)}$. Combining these results, we obtain the overall probability

$$p = \binom{t}{1} 1/s(1-1/s)^{(t-1)} \tag{18}$$

therefore

$$E_1(t, s) = sp = t(1-1/s)^{(t-1)} \tag{19}$$

extending this to $k$-occupancy yields

$$E_k(t, s) = \binom{t}{k} (1/s)^{(k-1)} (1-1/s)^{(t-k)}. \tag{20}$$

## Acknowledgment

The authors would like to thank A. Harrow for insights into urn problems and deriving the compact general slot occupancy probability distribution function presented in this text. They also thank K. Yogeeswaran, H. Reddy, B. Fiegel, S. Shah, and K. Asher for providing tools and insights used to gather experimental data. Finally, they thank several anonymous reviewers for providing detailed reviews which tremendously improved this paper.

## References

[1] EPC Global, "EPC® radio-frequency protocols class-1 generation-2 UHF RFID protocol for communications at 860 MHz–960 MHz version 1.1.0," 2006.

[2] N. Abramson and F. Kuo, Eds., "The ALOHA System," in *Computer Networks*. Englewood Cliffs, NJ, Prentice-Hall, 1973.

[3] *Information Processing Systems: Local Area Networks—Part 3. Carrier Sense Multiple Access With Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications*, ANSI/IEEE Standard 802.3-1990 edition, 1992, ANSI, International Standard ISO/IEC 8802-3, IEEE product number: SH13482 ed..

[4] V. Pillai *et al.*, "An ultra-low-power long range battery/passive RFID tag for UHF and microwave bands with a current consumption of 700 na at 1.5 v," *IEEE Trans. Circuits Syst.*, vol. 54, no. 7, pp. 1500–1511, 2007.

[5] K. S. Leong, M. L. Ng, and P. H. Cole, "The reader collision problem in RFID systems," in *Proc. IEEE Int. Symp. Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications*, 2005, vol. 1, pp. 658–661.

[6] K. S. Leong, M. L. Ng, A. R. Grasso, and P. H. Cole, "Dense RFID reader deployment in Europe using synchronization," in *J. Commun.*, Nov./Dec. 2006, vol. 1, no. 7, pp. 9–15.

[7] C. T. Lau and C. Leung, "Capture models for mobile packet radio networks," *IEEE Trans. Commun.*, vol. COM-40, pp. 917–925, May 1992.

[8] K.-T. Fang, , S. Kotz, N. L. Johnson, and C. B. Read, Eds., "Occupancy problems," in *Encyclopedia of Statistical Sciences*. New York: Wiley, 1985, pp. 402–406.

[9] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, 1991.

[10] D. J. C. MacKay, *Information Theory, Inference and Algorithms.* Cambridge, U.K.: Cambridge Univ. Press, 2002.

**Yael Maguire** received the M.S. degree in media arts and sciences from the Massachusetts Institute of Technology (MIT), Cambridge, MA, for his work towards scaling NMR quantum computing to tabletop systems, the Undergraduate degree from Queen's University, Canada, in engineering physics, and the Ph.D. degree in the area of near-field electromagnetic sensing for biology, chemistry, and electronics from Media Laboratory, MIT, Cambridge, MA, where he invented a new type of sensor for molecular detection using NMR.

He is a founder and co-CTO at ThingMagic, Inc., where he co-developed the Mercury platform of software-defined radio RFID readers. He has won numerous scholarships and awards, most recently being recognized as one of Technology Review's "Top 35 innovators in the world under 35." His research has focused on the fundamental ties between information processing and physics, quantum computing, biophysics, and microelectromechanical systems (MEMS).

**Ravikanth Pappu** received the B.S. degree in electronics and communication engineering from Osmania University, India, the M.S. degree in electrical engineering from Villanova University, Villanova, PA, the M.S. degree in media arts and science and the Ph.D. degree in 2001 from the Massachusetts Institute of Technology (MIT), Cambridge, MA, for the invention of physical one-way functions.

He is a founder and Head of Advanced Development, ThingMagic, Inc., where he works on designing, implementing, and deploying cutting edge RFID systems in environments ranging from distribution centers to pickup trucks. Most recently, he led the design and implementation of the Tool Link asset tracking system in collaboration with Ford Motor Company and Dewalt. While at MIT, he co-created the first dynamic holographic video system with haptic interaction (aka the HoloDeck). In 2003, he was honored as one of Technology Review's top 100 innovators under the age of 35. He has published over 25 papers and is a named inventor on 12 U.S. and international patents. His research interests are in RFID and sensor networks, cryptography, and optical engineering.