Probabilistic Quorum Systems

Dahlia Malkhi[*] Michael K. Reiter[†] Avishai Wool[‡] Rebecca N. Wright[§]

Abstract

We initiate the study of *probabilistic quorum systems*, a technique for providing consistency of replicated data with high levels of assurance despite the failure of data servers. We show that this technique offers effective load reduction on servers and high availability. We explore probabilistic quorum systems both for services tolerant of benign server failures and for services tolerant of arbitrary (Byzantine) ones. We also prove bounds on the server load that can be achieved with these techniques.

1 Introduction

Quorums are tools for increasing the availability and efficiency of replicated data. A *quorum system* is a set of subsets of servers, every two of which intersect. Intuitively, the intersection property guarantees that if a "write" operation is performed at one quorum, and later a "read" operation is performed at another quorum, then there is some server that observes both operations and therefore is able to provide the up-to-date value to the reader. Thus, system-wide consistency can be maintained while allowing any quorum to act on behalf of the entire system. Compared with performing every operation at every server, using quorums reduces the load on servers and increases service availability despite server crashes.

Quorum systems are traditionally assessed by three measures of quality: load [NW98], fault tolerance [BG86], and failure probability (see [BG87, PW95]). The *load* of a quorum system is a measure of its efficiency: it is the rate at which the busiest server is accessed. The *fault tolerance* of a system is the maximum number of server failures for which there is still guaranteed to be a quorum containing no faulty servers. The *failure probability* is the probability that every quorum contains a faulty server, assuming that servers fail independently with a fixed probability. (Load, fault tolerance, and failure probability are defined precisely in Section 2.) The fault tolerance of any quorum system is bounded by half of the number of servers. Moreover, the failure probability typically grows to 1 when the individual failure probability of servers exceeds 1/2. Also, there is a tradeoff between low load and good fault tolerance, and in fact it is impossible to achieve optimality in both of them simultaneously.

To circumvent these limitations, we relax the intersection property of a quorum system to allow

[*]School of Computer Science and Engineering, the Hebrew University of Jerusalem, Israel; dalia@cs.huji.ac.il. Work done while at AT&T Labs–Research.

[†]Bell Labs, Lucent Technologies, Murray Hill, New Jersey, USA; reiter@research.bell-labs.com. Work done while at AT&T Labs—Research.

[‡]Bell Labs, Lucent Technologies, Murray Hill, New Jersey, USA; yash@acm.org

[§]AT&T Labs—Research, Florham Park, New Jersey, USA; rwright@research.att.com.

Appears in *Information and Computation*, Vol. 170, No. 2, 2001.

"quorums" chosen according to a specific *access strategy* to fail to intersect with some small probability ε. Accordingly, we call these *ε-intersecting quorum systems* and, though an abuse of terminology, continue to refer to the chosen sets of servers as "quorums." We henceforth refer to systems that satisfy the original definition of quorums as *strict* quorum systems. We then extend the definition of the three quality measures—load, fault tolerance and failure probability—to address the probabilistic nature of our set systems. By these measures, probabilistic quorum systems show dramatic improvement over strict quorum systems: Allowing even a small probability ε of non-intersection yields a clear advantage in the fault tolerance and failure probability of the system, while the load remains unchanged or improves.

We study probabilistic quorum systems further, in a model where servers may exhibit arbitrary (Byzantine) failures. Malkhi and Reiter [MR98a] adapted strict quorum systems to the task of masking Byzantine failures to improve the efficiency of Byzantine fault-tolerant data replication. They introduced *b-dissemination* quorum systems, in which any two quorums intersect in at least $b+1$ servers, to mask the arbitrary failure of b servers for self-verifying data, and *b-masking* quorum systems, in which two quorums intersect in at least $2b+1$ servers, to mask b arbitrary server failures for arbitrary data. In this paper, we also explore relaxing the intersection properties of these quorums, to achieve (b, ε)-dissemination and (b, ε)-masking systems. Again, we show that these systems offer substantial improvements over their strict counterparts in the measures described above.

1.1 Applications

Due to their relaxed intersection properties, our probabilistic quorums are most suitable for use when the consistency of replicated data may be relaxed to achieve greater availability of that data. Below we describe several examples of applications where this tradeoff is justified.

The first application arose in the context of an electronic voting system designed by AT&T Labs for the country of Costa Rica. In this system, each voter is given a unique voter identifier when she registers to vote. On election day, she presents this voter ID to any one of over 1000 voting stations spread across the country in order to cast her vote. To prevent a voter ID from being used multiple times in one election, it is necessary to "lock" each voter ID country-wide when it is presented at any voting station. In order to preserve the integrity of the election, it is only necessary to prevent large-scale repeat voting. Therefore, it suffices for each repeated use of the same voter ID to be detected with high probability, so numerous repeat attempts will be detected with virtual certainty. We thus adopted a protocol among voting stations that uses probabilistic quorums for this purpose. Moreover, by using our dissemination or masking quorum constructions in the locking protocol, repeat usage of a voter ID can be prevented even if some number of voting stations do not follow the locking protocol (e.g., some stations have been altered by bribed election officials). At the same time, the use of our probabilistic quorum constructions ensures that the election progresses even in the presence of benign failures of significant numbers of voting stations. A prototype implementation of the Costa Rica electronic voting system was built over the Phalanx system [MR98b]. This implementation made use of various probabilistic quorum systems, including masking systems, for locking voter IDs.

The second application is maintaining the location of a mobile device, such as a cellular telephone. The location of a mobile device can be recorded in a variable that is replicated at several "location stores". This variable is updated (e.g., by the device itself) using a quorum-based protocol among the location stores when the device moves from cell to cell (cf. [HL99]). The ability of callers to access this information, even at the risk of it being stale, is the primary requirement for this application. A caller that receives stale location information can be forwarded by the stale cell to a more recent cell for that device, but the caller can make no progress if it receives no information about the device's current or recent whereabouts due to location store failures.

Finally, we note that a system built with probabilistic quorum systems can be strengthened by a properly designed *diffusion* mechanism, which propagates updates to replicated data "lazily", i.e., outside the critical path of client operations. Diffusion methods (also known as "epidemic" or "gossip" protocols) have been studied for both benign failure environments [DGH+87, AES97] and Byzantine environments [MMR99]. Coupled with a diffusion mechanism, the probability of inconsistency using probabilistic quorum constructions can be driven further toward zero when updates are sufficiently dispersed in time, making probabilistic quorum constructions useful in a wider variety of settings.

1.2 Related work

Strict quorum systems have been extensively studied and measured (cf., [Gif79, Tho79, Mae85, GB85, Her86, BG87, ET89, CAA90, AE91, NW98, PW97, PW95]). Byzantine quorum systems were introduced in [MR98a], and further studied in [MRW00, Baz97].

Because of the possibility of inconsistency admitted by probabilistic quorum systems, they are most attractive for systems in which some level of inconsistency can be tolerated, and in particular, where the efficiency and availability gained outweigh the cost of handling such inconsistencies. Other approaches to relaxing consistency in such application domains have been proposed. For example, Hayden and Birman implemented *pbcast* [HB95], a probabilistic broadcast with relaxed reliability guarantees for building fault-tolerant distributed applications. Malkhi et al. [MMR97] proposed a probabilistic secure broadcast with a relaxed consistency property for securely replicating services in a very large network.

In the domain of replicated database systems, several approaches that relax the strict *serializability* guarantee have been proposed. The goal of these efforts has been to decrease the contention between user transactions, and hence to increase the concurrency and decrease the abort rate. Krishnakumar and Bernstein [KB94] suggest N*-bounded-ignorance*, a relaxed consistency condition that permits $N+1$ conflicting transactions to be performed concurrently. Pu and Leff [PL91] propose *epsilon-serializability*, another relaxed consistency condition, that allows query-transactions (containing read operations only) to overlap update-transactions arbitrarily, and further allow content-dependent concurrency in update-transactions, based on the semantics of their operations. Wong and Agrawal [WA92] also make use of the semantics of the data items, and furthermore take into account the state of each transaction while it is executing.

Another setting in which the use of replicated variables to give probably correct results has proved to be useful is the efficient simulation of a PRAM on an asynchronous system [KPRR92, AR92]. Specifically, Kedem et al. [KPRR92] use replicated variables in a way that a correct copy can be reliably identified and probably exists. They then use these variables to create a global counter that processors use to determine whether they are roughly synchronized with other processors, and behave appropriately if they are not. Aumann and Rabin [AR92] exhibit a clock construction in an asynchronous system with multiple processors that use a shared memory to create an object that correctly behaves as a clock with high probability. They use the clock to ensure that processors stay synchronized throughout the computation. In both cases, the protocols to read and write the replicated variables are somewhat complex due to the need to detect or mask incorrect copies.

Unlike these previous works, which are tailored to specific application requirements, in our work we strive for a general technique for replicating data with a high degree of simplicity, efficiency and fault tolerance. Consequently, our techniques are very different from those used in these previous works.

1.3 Our results

We begin by defining and exploring the limits of ε-intersecting quorum systems. In particular, we show a lower bound on the load of ε-intersecting quorum systems that is within a small constant factor of the bound for strict quorum systems. Thus, ε-intersecting quorum systems cannot yield substantial improvements on the load in general. In contrast, we show that ε-intersecting quorum systems *can* yield substantial improvements on the load when high fault tolerance is also needed. For any ε, we provide a simple construction of an ε-intersecting quorum system that demonstrates optimal load and fault tolerance—$O(1/\sqrt{n})$ and $\Omega(n)$, respectively, for a system of n servers—thereby circumventing a tradeoff between optimal load and fault tolerance inherent in strict quorum systems. In addition, our construction has failure probability better than that of any strict quorum system.

For an environment in which servers may experience Byzantine failures but servers store only *self-verifying* data, i.e., data that servers can suppress but not un-detectably alter (such as digitally signed data), we investigate (b, ε)-*dissemination quorum systems*. We demonstrate a dramatic improvement in both the load and the fault tolerance in this setting: strict b-dissemination quorums systems can be constructed for $b \leq \lfloor \frac{n-1}{3} \rfloor$ arbitrarily faulty servers, and the load of a dissemination quorum system with such resilience is at least $\frac{2}{3}$. Using essentially the same construction as we use to demonstrate ε-intersecting systems, we demonstrate a (b, ε)-dissemination quorum system resilient to the arbitrary failure of any constant fraction of the servers and with outstanding failure probability, whose load is $O(1/\sqrt{n})$. For large n, this construction provides a considerable advantage over strict dissemination quorum system constructions.

Finally we define and explore (b, ε)-*masking quorum systems*, which can mask b arbitrary server failures for arbitrary forms of data. Using techniques that diverge from the previous, we derive and prove correct a general (b, ε)-masking quorum system construction that can mask any $b < \frac{n}{2}$ Byzantine failures with an arbitrarily small ε. Our construction beats the $\Omega(\sqrt{b/n})$ lower bound on the load of any strict masking system [MRW00]. For instance, we demonstrate a system that can mask up to $b = \sqrt{n}$ Byzantine failures with a load of only $O(n^{-0.3})$. Moreover, we show lower bounds on the load of any (b, ε)-masking quorum system that demonstrate that our construction is asymptotically load-optimal when $b = \omega(\sqrt{n})$.[1] Note that this also demonstrates that our lower bounds are tight in this case. We also show that our construction offers excellent failure probability.

The contributions of this paper can be summarized as follows:

- We initiate the study of probabilistic quorum systems and provide formal definitions for them. We extend the traditional definitions of three measures of quality—load, fault tolerance and failure probability—to address probabilistic constructions. We define three types of probabilistic quorum systems: ε-intersecting quorum systems that tolerate benign failures only; (b, ε)-dissemination quorum systems tolerant of b arbitrary server failures with self-verifying data; and

[1] ω is the little-oh analog of Ω, namely $f(n) = \omega(g(n))$ if $f(n)/g(n) \to \infty$ as $n \to \infty$.

(b, ε)-masking quorum systems that tolerate b arbitrary server failures with arbitrary data. We present protocols for using any quorum system meeting our definitions to implement a replicated variable whose semantics approximate that of a *multi-reader, single-writer safe variable* [Lam86].

- We provide practical constructions for each class of quorum system. Our constructions have outstanding behavior in all measures: they have higher fault tolerance than strict ones; they achieve better failure probability, and in particular, can achieve vanishingly small failure probability even when the individual component failure probability is more than $1/2$, thus beating the failure probability of any strict quorum system; and they maintain these properties simultaneously with optimal load. For Byzantine environments, our probabilistic dissemination and masking constructions can also beat the general lower bound on the load of their strict counterparts.
- We show lower bounds on the load of each type of probabilistic quorum system that demonstrate the load-optimality of our constructions.

The rest of this paper is structured as follows. In Section 2, we present the definitions of strict quorum systems and of the various traditional quality measures. Section 3 defines ε-intersecting quorum systems and extends the traditional quality measures to the probabilistic setting, proves a lower bound on the load of any ε-intersecting quorum system, and presents a construction exhibiting very good load, fault tolerance and failure probability. Section 4 introduces (b, ε)-dissemination quorum systems and provides a construction tolerant of the Byzantine failure of any constant fraction of the servers. In Section 5, we define (b, ε)-masking quorum systems, present a lower bound on the load of any such system, and present a probabilistic masking quorum system construction. We demonstrate the advantages of our techniques for particular system sizes in Section 6 and conclude in Section 7.

2 Strict quorum systems

In this section, we give a brief review of strict quorum systems. We assume a *universe* U of servers, $|U| = n$, and a distinct set of clients. Servers that obey their specifications are *correct*. A (Byzantine) *faulty* server, however, may deviate from its specification arbitrarily. When working with Byzantine failures, we assume that up to b servers may exhibit Byzantine failures. At times we restrict our attention to *crash* failures only, where a server fails by simply halting; we will be explicit when we do so. Throughout this paper we assume that clients behave according to their specifications.

Definition 2.1 *A* set system $\mathcal{Q}$ *over a universe* U *is a set of subsets of* U*.*

Definition 2.2 *A* (strict) quorum system $\mathcal{Q}$ *over a universe* U *is a set system over* U *such that for every* $Q, Q' \in \mathcal{Q}$, $Q \cap Q' \neq \emptyset$*. Each* $Q \in \mathcal{Q}$ *is called a* quorum*.*

Intuitively, because every two quorums intersect, when a client reads the replicated data, it is sure to receive the last written value from some server, namely the one that is both in its read quorum and in the last write quorum. In a typical access protocol in which values are written with timestamps, the reader can detect the most up-to-date value as the one with the highest associated timestamp.

Remark: In the domain of replicated database systems, it is common to differentiate between the collection of read-quorums $\mathcal{R}$ and the collection of write-quorums $\mathcal{W}$ [BHG87]. The intersection requirement is then $R \cap W \neq \varnothing$ and $W \cap W' \neq \varnothing$ for all $R \in \mathcal{R}$ and $W, W' \in \mathcal{W}$. For simplicity we shall not make the distinction between the two types of quorums.

Traditionally, three measures were defined to assess the quality of quorum systems: the load, the fault tolerance, and the failure probability of the system.

2.1 Load

Clients pick quorums to access in accordance with some *access strategy*, which defines the likelihood that a quorum is chosen for any given access.

Definition 2.3 *An* access strategy *(or simply a* strategy*) w for a set system $\mathcal{Q}$ specifies a probability distribution on the elements of $\mathcal{Q}$. That is, $w : \mathcal{Q} \to [0,1]$ satisfies $\sum_{Q \in \mathcal{Q}} w(Q) = 1$.*

The load of a quorum system, defined by Naor and Wool [NW98], captures the probability of accessing the busiest server. Load is a measure of efficiency. All other things being equal, systems with lower load can process more requests than those with higher load.

Definition 2.4 *Let w be a strategy for a set system $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ over a universe U. For a server $u \in U$, the* load *induced by w on u is $l_w(u) = \sum_{Q_i \ni u} w(Q_i)$. The* load *induced by a strategy w on $\mathcal{Q}$ is $L_w(\mathcal{Q}) = \max_{u \in U}\{l_w(u)\}$. The load of $\mathcal{Q}$ is $L(\mathcal{Q}) = \min_w\{L_w(\mathcal{Q})\}$, where the minimum is taken over all strategies.*

It is known that for any quorum system $\mathcal{Q}$ over n servers, $L(\mathcal{Q}) \geq \max\{\frac{1}{c(\mathcal{Q})}, \frac{c(\mathcal{Q})}{n}\}$ where $c(\mathcal{Q})$ is the size of the smallest quorum in $\mathcal{Q}$ [NW98]. In particular, this implies that for any quorum system $\mathcal{Q}$, $L(\mathcal{Q}) \geq 1/\sqrt{n}$.

We note that load is a best case definition. The load of the quorum system will be achieved only if an optimal access strategy is used, and only in the case that no failures occur. A strength of this definition is that load is a property of a quorum system, and not of the protocol using it. A comparison of the definition of load to other seemingly plausible definitions is given in [NW98].

2.2 Fault Tolerance

Fault tolerance and failure probability capture the resilience of the system to crash failures. The fault tolerance of a quorum system $\mathcal{Q}$ is the size of the smallest set of servers that intersects all quorums in $\mathcal{Q}$.

Definition 2.5 *For a set system $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$, define $\mathcal{S} = \{S : S \cap Q_i \neq \varnothing$ for all $1 \leq i \leq m\}$. Then the* fault tolerance *of $\mathcal{Q}$ is $A(\mathcal{Q}) = \min_{S \in \mathcal{S}} |S|$.*

Thus, a quorum system $\mathcal{Q}$ is resilient to the failure of any set of $A(\mathcal{Q}) - 1$ or fewer servers. In particular, the failure of at least $A(\mathcal{Q})$ servers is necessary to disable every quorum in the system, and some particular set of $A(\mathcal{Q})$ failures can in fact disable them all.

Moreover, the intersection property implies that the failure of any full quorum in $\mathcal{Q}$ will disable all quorums (i.e., $A(\mathcal{Q}) \leq c(Q)$), and so by the aforementioned lower bound on load, $A(\mathcal{Q}) \leq nL(\mathcal{Q})$. Therefore, there is a tradeoff between load and fault tolerance in strict quorum systems, and in particular, it follows that any strict quorum system with optimal load of $\Theta(1/\sqrt{n})$ has fault tolerance of (only) $O(\sqrt{n})$.

2.3 Failure Probability

The failure probability of a quorum system is the probability that the system is disabled when individual servers crash independently with a fixed probability.

Definition 2.6 *The* failure probability $F_p(\mathcal{Q})$ *of* $\mathcal{Q}$ *is the probability that every* $Q \in \mathcal{Q}$ *contains at least one crashed server, under the assumption that each server in* U *crashes independently with probability* p.

A strict quorum system $\mathcal{Q}$ has a "good" failure probability if $\lim_{n\to\infty} F_p(\mathcal{Q}) = 0$ when $p < \frac{1}{2}$ [PW95]. When $p \geq \frac{1}{2}$ then $F_p(\mathcal{Q}) \geq p \geq \frac{1}{2}$ for strict quorum systems, and typically $F_p(\mathcal{Q}) \to 1$ when $p > \frac{1}{2}$.

2.4 Finding a Live Quorum

In a quorum system that has a good failure probability, it is very likely that some live quorum (i.e., a quorum containing no crashed servers) exists. However, if the identities of the crashed servers are unknown to the client, the client needs to *find* such a live quorum. Designing techniques for efficiently finding a live quorum in case of crash failures is an active research topic. Clearly it is always possible for a client to find a live quorum if one exists by querying all the servers, but this has a high cost to the servers, and effectively increases the load to 1 on all the servers.

Several metrics for quantifying the difficulty of finding a live quorum can be found in the literature. In [PW96], Peleg and Wool define the *probe complexity* as the worst-case cost of adaptively discovering a live quorum. In [Baz96], Bazzi defines the *cost of failure*, as the number of extra server queries per crashed server the client discovers. And most recently, [Baz99] suggests accessing a superset of many quorums in the hope of catching at least one live quorum and controlling the load simultaneously.

2.5 Byzantine systems

As discussed in Section 1, quorum systems are generally insufficient to guarantee consistency in case of Byzantine server failures. Malkhi and Reiter extended quorum systems to handle Byzantine failures [MR98a]: a *b-dissemination* quorum system increases quorum overlap to $b+1$ servers, which suffices to mask faulty server behavior for self-verifying data; a *b-masking* quorum system further increases quorum overlap to $2b+1$ servers, masking faulty server behavior for any type of data[2].

Definition 2.7 *Let* $\mathcal{Q}$ *be a set system. Then*

- $\mathcal{Q}$ *is a* b-dissemination quorum system *if* $A(\mathcal{Q}) \geq b$ *and* $|Q \cap Q'| \geq b+1$ *for every* $Q, Q' \in \mathcal{Q}$.
- $\mathcal{Q}$ *is a* b-masking quorum system *if* $A(\mathcal{Q}) \geq b$ *and* $|Q \cap Q'| \geq 2b+1$ *for every* $Q, Q' \in \mathcal{Q}$.

For example, if a b-masking quorum system is used, then when a client performs a read operation at some quorum Q, the value written in the last preceding write operation, say to Q', is returned by at least $b+1$ correct servers, namely servers in the set $(Q \cap Q') \setminus B$ where B is the set of faulty servers. Any other returned value is either an old value, which can be detected by its earlier timestamp, or

[2]The original definition of [MR98a] allows more general failure configurations than we do here. The simplified definition presented here suffices for our purposes.

a "made-up" value returned only by servers in B. So, if the client discards any values that were returned by b or fewer servers, and then chooses from the remaining values the one with the most recent timestamp, then the client is guaranteed to obtain the correct value [MR98a].

There are known lower bounds on the load of strict, b-dissemination and b-masking quorum systems, and upper bounds on the attainable resilience for each system type [NW98, MR98a, MRW00]. In Table 1 we summarize these bounds concisely.

	strict	b-disseminating	b-masking
$L(\mathcal{Q}) \geq$	$\sqrt{\frac{1}{n}}$	$\sqrt{\frac{b+1}{n}}$	$\sqrt{\frac{2b+1}{n}}$
$b \leq$	N/A	$\lfloor \frac{n-1}{3} \rfloor$	$\lfloor \frac{n-1}{4} \rfloor$

Table 1: Bounds on the load and resilience of different quorum system types.

3 ε-intersecting quorum systems

In this section, we introduce probabilistic quorum systems and their properties. We first formally define ε-intersecting quorum systems, and show a protocol for using them. Next, we extend the definition of the three quality measures to account for the probabilistic nature of our systems. We then prove a lower bound on the load of ε-intersecting quorum systems, which shows that their relaxed consistency cannot yield substantial improvements on the load in general. Finally, we show that ε-intersecting quorums are not subject to the load/fault tolerance tradeoff, by demonstrating a construction over a universe of n servers that has a load of $\Theta(1/\sqrt{n})$ and fault tolerance of $\Theta(n)$, for which ε vanishes as n grows. We show that our construction has exceptionally good failure probability for essentially limitless component failure probabilities, for appropriate system sizes. The failure probability of our construction is provably better than that of any strict quorum system.

3.1 Definitions and usage of ε-intersecting quorum systems

We begin by defining ε-intersecting quorum systems. $\mathcal{Q}$ is an ε-intersecting quorum system if the total access probability of pairs of intersecting quorums is at least $1-\varepsilon$. Formally, we have the following.

Definition 3.1 *Let $\mathcal{Q}$ be a set system, let w be an access strategy for $\mathcal{Q}$, and let $0 < \varepsilon < 1$ be given. The tuple $\langle \mathcal{Q}, w \rangle$ is an* ε-intersecting quorum system *if $\mathbb{P}(Q \cap Q' \neq \varnothing) \geq 1 - \varepsilon$, where the probability is taken with respect to the strategy w.*

Abusing terminology slightly, we still call elements of $\mathcal{Q}$ quorums.

To demonstrate the utility of this definition, we now show a simple protocol which borrows from the protocols of Gifford [Gif79] and Thomas [Tho79] for accessing replicated data by a single writer and multiple readers, but with the distinction that it uses an ε-intersecting quorum system. Each server stores a copy of the replicated variable x and an associated timestamp value t that will be updated by clients. Write and read operations proceed as follows:

Write: For a client to write the value v to x, it

1. chooses a quorum Q according to the strategy w,
2. chooses a timestamp t greater than any timestamp it has chosen in the past, and
3. updates x and the associated timestamp at each server in Q to v and t, respectively.

Read: For a client to read x, it

1. chooses a quorum Q according to the strategy w,
2. queries each server in Q to obtain a set of value/timestamp pairs $V = \{\langle v_u, t_u \rangle\}_{u \in Q}$,
3. chooses the pair $\langle v, t \rangle$ in V with the highest timestamp, and
4. chooses v as the result of the read operation.

We are aware of no standard definition of variable semantics that can be used to prove correctness of the above protocol, due to the possibility (albeit small) that a read quorum does not intersect the most recent write quorum. The following theorem nevertheless clarifies its utility, showing that the protocol approximates a *multi-reader, single-writer, safe variable* [Lam86]. Safe variables are but one example of useful shared data abstractions implemented using probabilistic quorums; other replicated data objects can be either constructed using probabilistic quorum systems directly (e.g., locks [MR98b]) or using variables for building blocks (e.g., atomic variables, borrowing from the techniques of [Lam86, IS92]).

Theorem 3.2 *Consider a multi-reader, single-writer variable replicated using the above access protocol with an ε-intersecting quorum system. If a read operation is not concurrent with any write operation and only crash failures occur, then with probability at least $1 - \varepsilon$ the read returns the value written by the last preceding write operation.*

Proof: Consider the last write operation prior to the read operation. Since there is only one writer, it follows by the specification of the write protocol that it has the highest timestamp of any write operation that precedes the read. Moreover, with probability at least $1 - \varepsilon$, the quorum Q' picked in this write operation and the quorum Q picked in the current read operation satisfy $Q \cap Q' \neq \emptyset$. So, with probability at least $1 - \varepsilon$, this value/timestamp pair appears in V and thus the correct value will be returned by the read. □

Remark: The definition of an ε-intersecting quorum system contains an access strategy, which is chosen to achieve the desired bound ε on non-intersection between two quorums chosen according to the strategy. Other access strategies on the same set system may fail to achieve the same intersection guarantee, as can be trivially demonstrated by a strategy that chooses each of two nonintersecting quorums with probability 1/2. Thus, for an ε-intersecting quorum system to obtain the advertised probability of intersection when used in a protocol, the specified access strategy must be enforced.

3.2 Measures of quality

We now turn to adapting the various measures of quorum systems defined in Section 2 to probabilistic quorum systems. The definition of load carries over immediately.

Definition 3.3 *Let $\langle \mathcal{Q}, w\rangle$ be an ε-intersecting quorum system. Then the* load *of $\langle \mathcal{Q}, w\rangle$ is $L(\langle \mathcal{Q}, w\rangle) = L_w(\mathcal{Q})$.*

However, the definitions of fault tolerance and failure probability, as formulated for strict quorum systems, are unsatisfactory in a probabilistic setting. To demonstrate this, we show how to convert any ε-intersecting quorum system $\langle \mathcal{Q}, w\rangle$ into a new system $\langle \mathcal{Q}', w'\rangle$ which has essentially the same consistency guarantee $1-\varepsilon$ but with an artificially inflated fault tolerance. The set system $\mathcal{Q}'$ is created by simply adding every possible singleton set as a quorum: $\mathcal{Q}' = \mathcal{Q} \cup \{\{u_1\}, \ldots, \{u_n\}\}$. For any $\gamma \ll \varepsilon$, the strategy w' is defined by $w'(Q) = (1-\gamma)w(Q)$ for all $Q \in \mathcal{Q}$, and $w'(\{u_i\}) = \gamma/n$ for all the singletons. Since the singleton quorums $\{u_i\}$ are used with such low probability, it is easy to see that $\langle \mathcal{Q}', w'\rangle$ is ε'-intersecting, with $\varepsilon' \approx \varepsilon$. However, the only way to disable all the quorums of $\mathcal{Q}'$ is to have all the servers crash, so $A(\mathcal{Q}') = n-1$. Likewise, the failure probability of $\mathcal{Q}'$ is unreasonably good: according to Definition 2.6, $F_p(\mathcal{Q}') = p^n$.

The problem with naively using Definitions 2.5 and 2.6 is that they allow the fault tolerance to be derived from quorums that intersect few other quorums, and are are hardly ever used by the strategy. Any reasonable definition of fault tolerance for probabilistic quorum systems should require that the fault tolerance be derived from those quorums that intersect other quorums with high probability. To make this intuition precise, we need the following technical definition and lemma, leading to Definitions 3.7 and 3.8.

From here on, all probabilities and expectations are taken with respect to the strategy w, unless explicitly denoted otherwise.

Definition 3.4 *Let $\langle \mathcal{Q}, w\rangle$ be an ε-intersecting quorum system, and let $0 \le \delta \le 1$ be given. The set of* δ-high quality quorums *of $\langle \mathcal{Q}, w\rangle$ is*

$$\mathcal{R} = \{Q \in \mathcal{Q} : \mathbb{P}(Q \cap Q' \neq \varnothing) \geq 1 - \delta\},$$

where $Q' \in \mathcal{Q}$ is chosen according to w.

The following lemma shows that in an ε-intersecting quorum system, most of the weight lies on the δ-high quality quorums.

Lemma 3.5 $\mathbb{P}(Q \in \mathcal{R}) \geq 1 - \frac{\varepsilon}{\delta}$.

Proof: From Definition 3.1,

$$\varepsilon \;\geq\; \mathbb{P}(Q \cap Q' = \varnothing) = \sum_{Q \in \mathcal{Q}} w(Q) \sum_{Q' : Q \cap Q' = \varnothing} w(Q') \;\geq\; \sum_{Q \notin \mathcal{R}} w(Q) \sum_{Q' : Q \cap Q' = \varnothing} w(Q').$$

For any fixed $Q \notin \mathcal{R}$, $\sum_{Q' : Q \cap Q' = \varnothing} w(Q') = \mathbb{P}(Q \cap Q' = \varnothing) > \delta$ by Definition 3.4. Thus,

$$\frac{\varepsilon}{\delta} \;\geq\; \sum_{Q \notin \mathcal{R}} w(Q) = 1 - \mathbb{P}(Q \in \mathcal{R}). \qquad \square$$

Consequently, by choosing δ so that both δ and ε/δ are small, the δ-high quality quorums are "high quality" in two respects: they intersect other chosen quorums with high probability (by definition), and they constitute the quorums that are selected with high probability (by Lemma 3.5). A reasonable choice of δ to render both δ and ε/δ small when ε is small is $\delta = \sqrt{\varepsilon}$. Henceforth, we adopt this convention and refer to the $\sqrt{\varepsilon}$-high quality quorums as simply the high quality quorums:

Definition 3.6 *Let $\langle \mathcal{Q}, w\rangle$ be an ε-intersecting quorum system. Then the* high quality quorums *of $\langle \mathcal{Q}, w\rangle$ are the $\sqrt{\varepsilon}$-high quality quorums of $\langle \mathcal{Q}, w\rangle$.*

We are now prepared to state our definitions for fault tolerance and failure probability. The difference between these definitions and Definitions 2.5 and 2.6 is that here we consider the system to be disabled if all the high quality quorums are hit.

Definition 3.7 *Let $\langle \mathcal{Q}, w\rangle$ be an ε-intersecting quorum system. Let $\mathcal{R}$ be the set of high quality quorums of $\langle \mathcal{Q}, w\rangle$, and let $\mathcal{S} = \{S : S \cap Q \neq \varnothing$ for all $Q \in \mathcal{R}\}$. Then the* fault tolerance *$A(\langle \mathcal{Q}, w\rangle)$ is $\min_{S\in\mathcal{S}} |S|$.*

Definition 3.8 *Let $\langle \mathcal{Q}, w\rangle$ be an ε-intersecting quorum system, and let $\mathcal{R}$ be the set of high quality quorums of $\langle \mathcal{Q}, w\rangle$. The* failure probability *$F_p(\langle \mathcal{Q}, w\rangle)$ is the probability that every $Q \in \mathcal{R}$ contains at least one crashed server, under the assumption that each server in U crashes independently with probability p.*

These definitions are consistent with Definitions 2.5 and 2.6 for strict quorum systems: In any strict quorum system $\mathcal{Q}$, all the quorums are high quality quorums by definition, irrespective of the access strategy used. Hence, $A(\langle \mathcal{Q}, w\rangle) = A(\mathcal{Q})$ for all strategies w. For a probabilistic quorum system $\langle \mathcal{Q}, w\rangle$, $A(\langle \mathcal{Q}, w\rangle) \leq A(\mathcal{Q})$ and $F_p(\langle \mathcal{Q}, w\rangle) \geq F_p(\mathcal{Q})$, which stands to reason since $A(\langle \mathcal{Q}, w\rangle)$ and $F_p(\langle \mathcal{Q}, w\rangle)$ depend on w. The reader can verify that unlike the strict $A(\mathcal{Q})$ and $F_p(\mathcal{Q})$, the probabilistic measures $A(\langle \mathcal{Q}, w\rangle)$ and $F_p(\langle \mathcal{Q}, w\rangle)$ cannot be artificially inflated by adding "hidden" servers and quorums.

As for finding a live quorum, all the approaches mentioned in Section 2.4 may be used in the probabilistic setting. As we shall see in the sequel, all our probabilistic constructions are based on a threshold (i.e., a quorum is any collection of q servers for some q). For such constructions, the various quorum-finding approaches translate into extremely simple strategies, which only involve the number of servers accessed in each round. In terms of Bazzi's cost-of-failures measure, the cost of any threshold construction is 1, which is optimal. Nonetheless, Proposition 4.9 of [PW96] proves that for threshold-based quorum systems (probabilistic or not), in the worst case the client will need to probe *all* the servers before being able to determine whether a live quorum exists or not. However, this negative result says nothing about the average number of probes required. This is a topic for future research.

3.3 A lower bound on the load

In this section we state and prove a lower bound on the load of ε-intersecting quorum systems. This lower bound is close to the lower bound for strict quorum systems, and thus indicates that we should not look to ε-intersecting quorums as a technique to circumvent the lower bound for strict ones.

Theorem 3.9 *Let $\langle \mathcal{Q}, w\rangle$ be a ε-*intersecting quorum system*, and let the random variable $|Q|$ be the size of a quorum chosen according to w. Then*

$$L(\langle \mathcal{Q}, w\rangle) \geq \max\left\{\frac{\mathbb{E}[|Q|]}{n}, \frac{(1-\sqrt{\varepsilon})^2}{\mathbb{E}[|Q|]}\right\}.$$

Theorem 3.9 is similar to the bounds shown in [NW98] for strict quorum systems. The main differences are that here we have a specific strategy w so we can work with the expected quorum size (rather than the minimal quorum size), and that we need to account for the small probability of quorums not intersecting each other. We prove Theorem 3.9 via the following two Lemmas.

Lemma 3.10 *Let $\mathcal{Q}$ be a set system and let w be a strategy for $\mathcal{Q}$. Then $L_w(\mathcal{Q}) \geq \frac{\mathbb{E}[|Q|]}{n}$.*

Proof: By summing the total load induced by w on all the elements of U we obtain

$$n \cdot L_w(\mathcal{Q}) \geq \sum_{u \in U} l_w(u) = \sum_{u \in U} \sum_{Q \ni u} w(Q) = \sum_{Q \in \mathcal{Q}} w(Q)|Q| = \mathbb{E}[|Q|]. \qquad \square$$

Lemma 3.11 *Let $\langle \mathcal{Q}, w \rangle$ and $|Q|$ be as in Theorem 3.9. Then $L_w(\mathcal{Q}) \geq \frac{(1-\sqrt{\varepsilon})^2}{\mathbb{E}[|Q|]}$.*

Proof: Let $\mathcal{R}$ be the high quality system associated with $\langle \mathcal{Q}, w \rangle$ (as in Definition 3.4 with $\delta = \sqrt{\varepsilon}$). Define a restricted strategy w_r over $\mathcal{Q}$ by

$$w_r(Q') = \begin{cases} w(Q')/\mathbb{P}(Q \in \mathcal{R}), & \text{if } Q' \in \mathcal{R}, \\ 0, & \text{otherwise.} \end{cases}$$

The expected chosen quorum size with respect to w_r obeys

$$\mathbb{E}_{w_r}[|Q|] = \sum_{Q' \in \mathcal{Q}} w_r(Q') \cdot |Q'| = \sum_{Q' \in \mathcal{R}} \frac{w(Q')}{\mathbb{P}(Q \in \mathcal{R})} |Q'| \leq \frac{1}{\mathbb{P}(Q \in \mathcal{R})} \sum_{Q' \in \mathcal{Q}} w(Q') \cdot |Q'| = \frac{\mathbb{E}[|Q|]}{\mathbb{P}(Q \in \mathcal{R})},$$

and hence by Lemma 3.5,

$$\mathbb{E}[|Q|] \geq (1 - \sqrt{\varepsilon}) \mathbb{E}_{w_r}[|Q|]. \tag{1}$$

Now fix some $\widehat{Q} \in \mathcal{R}$ with $|\widehat{Q}| \leq \mathbb{E}_{w_r}[|Q|]$. (Such a set must exist by the definition of $\mathbb{E}_{w_r}$). Summing the load induced by w on the elements of $\widehat{Q}$ we have

$$\begin{aligned} |\widehat{Q}| \cdot L_w(\mathcal{Q}) &\geq \sum_{u \in \widehat{Q}} l_w(u) = \sum_{u \in \widehat{Q}} \sum_{Q \ni u} w(Q) = \sum_{Q \in \mathcal{Q}} w(Q) |Q \cap \widehat{Q}| \\ &\geq \sum_{Q : Q \cap \widehat{Q} \neq \varnothing} w(Q) = \mathbb{P}(Q \cap \widehat{Q} \neq \varnothing) \geq 1 - \sqrt{\varepsilon} \end{aligned} \tag{2}$$

by definition since $\widehat{Q} \in \mathcal{R}$. Now by (2), (1), and the definition of $\widehat{Q}$, we get

$$L_w(\mathcal{Q}) \geq \frac{1 - \sqrt{\varepsilon}}{|\widehat{Q}|} \geq \frac{1 - \sqrt{\varepsilon}}{\mathbb{E}_{w_r}[|Q|]} \geq \frac{(1 - \sqrt{\varepsilon})^2}{\mathbb{E}[|Q|]}. \qquad \square$$

Theorem 3.9 follows directly from Lemmas 3.10 and 3.11.

Corollary 3.12 $L(\langle \mathcal{Q}, w \rangle) \geq \frac{(1-\sqrt{\varepsilon})}{\sqrt{n}}$.

Proof: Immediate from Theorem 3.9. $\square$

3.4 An ε-intersecting quorum system construction

We now demonstrate an ε-intersecting quorum system $\langle \mathcal{Q}, w\rangle$ with $O(1/\sqrt{n})$ load and $\Omega(n)$ fault tolerance that meets any required ε for sufficiently large n. The construction is very simple: Given a universe of n servers, the quorums are all the sets of size $\ell\sqrt{n}$; the strategy chooses a quorum uniformly at random. The constant ℓ is chosen to make ε sufficiently small. Intuitively, it is easy to see that this should work—the expected, and most probable, size of the intersection of two such quorums is ℓ^2, so by making ℓ sufficiently large, it should be possible to reduce the probability ε that the intersection of two quorums is empty to any desired level. This is similar to the well-known birthday paradox (see [CLR89]): Given two quorums, the probability that any given element in one quorum is also in the second quorum is quite small ($\frac{\ell}{\sqrt{n}}$), but the probability that *some* element appears in both quorums is quite high (at least $1 - e^{-\ell^2}$, as we prove below).

Definition 3.13 *Let U be a universe of size n. Then $R(n, q)$ is the system $\langle \mathcal{Q}, w\rangle$ defined by $\mathcal{Q} = \{Q \subseteq U : |Q| = q\}$ with the uniform strategy $w(Q) = \frac{1}{|\mathcal{Q}|}$ for all quorums $Q \in \mathcal{Q}$.*

We consider $R(n, \ell\sqrt{n})$ and show that the probability of choosing at random two quorums that do not intersect can be made sufficiently small by appropriate choice of ℓ. We use the following combinatorial fact.

Proposition 3.14 *For integers n, c, and i, $\frac{\binom{n-c}{c-i}}{\binom{n}{c}} \le \left(\frac{c}{n}\right)^i \left(\frac{n-c}{n-i}\right)^{c-i}$.*

Lemma 3.15 *Let Q and Q' be quorums of size $\ell\sqrt{n}$ each chosen uniformly at random. Then $\mathbb{P}(Q \cap Q' = \varnothing) < e^{-\ell^2}$.*

Proof:

$$
\begin{aligned}
\mathbb{P}(Q \cap Q' = \varnothing) &= \frac{\binom{n-\ell\sqrt{n}}{\ell\sqrt{n}}}{\binom{n}{\ell\sqrt{n}}} \le \left(\frac{n-\ell\sqrt{n}}{n}\right)^{\ell\sqrt{n}} \\
&\le e^{-\frac{\ell\sqrt{n}}{n}\ell\sqrt{n}} = e^{-\ell^2},
\end{aligned}
$$

where the first inequality follows from Proposition 3.14. □

It is immediate from Lemma 3.15 that $R(n, \ell\sqrt{n})$ is an ε-intersecting quorum system:

Theorem 3.16 *$R(n, \ell\sqrt{n})$ is an $(e^{-\ell^2})$-intersecting quorum system.*

Quality measures Since every element is in $\binom{n-1}{\ell\sqrt{n}-1}$ quorums, the load $L(R(n, \ell\sqrt{n}))$ is $\frac{\ell}{\sqrt{n}} = O(1/\sqrt{n})$. $R(n, \ell\sqrt{n})$ is a symmetrical construction with a uniform access strategy, and hence all of its members are high quality quorums. Because only $\ell\sqrt{n}$ servers need be available in order for some (high quality) quorum to be available, the fault tolerance $A(R(n, \ell\sqrt{n})) = n - \ell\sqrt{n} + 1 = \Omega(n)$. The failure probability of $R(n, \ell\sqrt{n})$ is also exceptionally good. Let p denote the independent failure probability of servers. For the system to fail, at least $n - \ell\sqrt{n} + 1$ servers must fail. Using Chernoff's bound, this probability is at most

$$
\begin{aligned}
F_p(R(n, \ell\sqrt{n})) &= \mathbb{P}(\#\text{fail} > n - \ell\sqrt{n}) \\
&\le e^{-2n\left(1-\frac{\ell}{\sqrt{n}}-p\right)^2} \\
&= e^{-\Omega(n)}
\end{aligned}
$$

for all $p \leq 1 - \frac{\ell}{\sqrt{n}}$. Peleg and Wool showed that the failure probability of any strict quorum system whose fault tolerance is f is at most $e^{-\Omega(f)}$ [PW95]. Furthermore, they showed that for $p > \frac{1}{2}$, the failure probability is at least p. Therefore, if $p \leq 1 - \frac{\ell}{\sqrt{n}}$, the failure probability of $R(n, \ell\sqrt{n})$ is asymptotically optimal, and if $\frac{1}{2} \leq p \leq 1 - \frac{\ell}{\sqrt{n}}$, this probability is provably better than that of any strict quorum system.

Section 6 provides concrete examples of $R(n, \ell\sqrt{n})$ for various values of n and ℓ, compared in all three measures against concrete examples of strict quorum systems.

4 (b, ε)-dissemination quorum systems

To achieve consistency when servers can fail arbitrarily, it is not sufficient that two quorums have a nonempty intersection. This is because two quorums may intersect in a set containing *faulty* servers only, which may deviate arbitrarily and un-detectably from their assigned protocol. Malkhi and Reiter [MR98a] defined *(strict) dissemination* quorum systems that can be used to construct Byzantine fault-tolerant replicated services that store self-verifying data. Data is "self-verifying" if servers cannot alter data un-detectably, e.g., because clients digitally sign it. For such data a faulty server is constrained to return some value that was previously written, or be detected as faulty. In this case, it is sufficient to require that the intersection of every two quorums contains at least one non-faulty server, since this guarantees a correct, up-to-date value will be present and can be recognized. That is, the intersection of every two quorums should be of size at least $b + 1$, where b is the maximum number of Byzantine faults.

Here we modify dissemination quorum systems to a probabilistic setting. To achieve *probable* consistency in a Byzantine environment, it is not sufficient that two quorums should have a probably nonempty intersection, since again two quorums may intersect in a set consisting of faulty servers. Instead we use the following definition.

Definition 4.1 *Let $\mathcal{Q}$ be a quorum system, let w be an access strategy for $\mathcal{Q}$, and let $0 < \varepsilon < 1$ and an integer $b > 0$ be given. Then $\langle \mathcal{Q}, w \rangle$ is a (b, ε)-*dissemination quorum system *if $A(\langle \mathcal{Q}, w \rangle) \geq b$ and $\mathbb{P}(Q \cap Q' \not\subseteq B) \geq 1 - \varepsilon$ for all $B \subseteq U$ such that $|B| = b$.*

(b, ε)-dissemination quorum systems can be used to implement Byzantine fault-tolerant services for the same types of data that strict dissemination quorum systems can, using the same access protocol [MR98a]. Specifically, the read operation becomes:

Read: For a client to read x, it

1. chooses a quorum Q according to the strategy w,
2. queries each server in Q to obtain a set of value/timestamp pairs $V = \{\langle v_u, t_u \rangle\}_{u \in Q}$,
3. computes the set V' consisting of elements from V that are verifiable,
4. chooses the pair $\langle v, t \rangle$ in V' with the highest timestamp, and
5. chooses v as the result of the read operation.

The timestamps are assumed to be included as part of the self-verifying data. It follows that this protocol approximates a multi-reader, single-write safe variable when used with verifiable data in a Byzantine environment.

Theorem 4.2 *Consider a multi-reader, single-writer variable over verifiable data replicated using the above access protocol with a (b, ε)-dissemination quorum system. If a read operation is not concurrent with any write operation and at most b Byzantine failures occur, then with probability at least $1 - \varepsilon$ the read returns the value written by the last preceding write operation.*

Proof: As in the proof of Theorem 3.2, the last write operation prior to the read operation has the highest timestamp of any write operation that precedes the read. Moreover, with probability at least $1 - \varepsilon$, the quorum Q' picked in this write operation and the quorum Q picked in the current read operation satisfy $Q \cap Q' \not\subseteq B$ where B is the set of actually faulty servers. So, with probability at least $1 - \varepsilon$, this value/timestamp pair appears in V'. Further, by the verifiability of the data, only values from correct servers appear in V'. It follows that the correct value will be returned by the read. □

Note that since any (b, ε)-dissemination quorum system is also an ε-intersecting quorum system, the lower bound of Theorem 3.9 applies. Nonetheless, we show that relaxing quorum intersection in a Byzantine environment can provide dramatic improvements in both load and availability over strict dissemination quorum systems. Specifically, we show that our ε-intersecting quorum construction $R(n, \ell\sqrt{n})$ provides the following in this environment. First, it breaks the lower bound on the load of strict dissemination quorum systems of $\Omega(\sqrt{b/n})$ and achieves a load of $O(1/\sqrt{n})$. Second, its resilience level b can be increased to any constant fraction of the system, thus breaking the $\lfloor \frac{n-1}{3} \rfloor$ upper bound on the resilience of strict dissemination quorum systems, while retaining asymptotically optimal load. Third, it maintains an outstanding failure probability (for crash failures) for sufficiently large universes even for $p > 1/2$, thus beating the best failure probability of any strict quorum system. For convenience of the exposition, we first present a construction whose resilience is $b = \frac{n}{3}$, and later modify it for arbitrarily large b.

4.1 A (b, ε)-dissemination quorum system for $b = \frac{n}{3}$

In this section, we show that $R(n, \ell\sqrt{n})$, as defined in Section 3.4, can be used as a (b, ε)-dissemination quorum system for a Byzantine threshold $b = \frac{n}{3}$, the resilience bound for strict dissemination quorum systems [MR98a]. $R(n, \ell\sqrt{n})$ exhibits much better load and fault tolerance (to crash failures) than strict quorum constructions for this value of b. Although the bound ε is different than for the case with no Byzantine server failures, we are still able to show that for an appropriate choice of the parameter ℓ, this construction ensures intersection with any desired probability for sufficiently large universes.

Lemma 4.3 *Let U be a universe of size n, let B be a subset of U of size b where $b = \frac{n}{3}$, and let Q and Q' be quorums of size $\ell\sqrt{n}$ each chosen uniformly at random. Then $\mathbb{P}((Q \cap Q') \subseteq B) \leq 2e^{-\frac{\ell^2}{6}}$.*

Proof:

$$
\begin{aligned}
&\mathbb{P}(Q \cap Q' \subseteq B) \qquad (3)\\
&\quad = \mathbb{P}(|Q \cap Q'| = |Q \cap Q' \cap B|)\\
&\quad = \sum_{i=0}^{\ell\sqrt{n}} \mathbb{P}((|Q \cap Q'| = i) \wedge (|Q \cap Q' \cap B| = i))
\end{aligned}
$$

$$
\begin{aligned}
&\le \sum_{i=0}^{\ell\sqrt{n}} \frac{\binom{\ell\sqrt{n}}{i}\binom{n-\ell\sqrt{n}}{\ell\sqrt{n}-i}}{\binom{n}{\ell\sqrt{n}}}\left(\frac{1}{3}\right)^i && (4)\\
&\le \sum_{i=0}^{\ell\sqrt{n}} \binom{\ell\sqrt{n}}{i}\left(\frac{\ell\sqrt{n}}{n}\right)^i\left(\frac{n-\ell\sqrt{n}}{n-i}\right)^{\ell\sqrt{n}-i}\left(\frac{1}{3}\right)^i && (5)\\
&\le \sum_{i=0}^{\frac{\ell\sqrt{n}}{6}} \frac{(\ell^2)^i}{i!}\cdot e^{-\frac{(\ell\sqrt{n}-i)^2}{n-i}}\cdot 3^{-i} \;+\; \sum_{i=\frac{\ell\sqrt{n}}{6}+1}^{\ell\sqrt{n}} 3^{-i} && (6)\\
&\le \sum_{i=0}^{\frac{\ell\sqrt{n}}{6}} \frac{(\frac{\ell^2}{3})^i}{i!}\cdot e^{-\ell^2\left(\frac{5}{6}\right)^2} \;+\; \sum_{i=\frac{\ell\sqrt{n}}{6}+1}^{\ell\sqrt{n}} 3^{-i} && (7)\\
&\le e^{-\ell^2\left(\frac{5}{6}\right)^2}\cdot e^{\frac{\ell^2}{3}} \;+\; 3^{-\frac{\ell\sqrt{n}}{6}} && (8)\\
&\le 2e^{-\frac{\ell^2}{6}} && (9)
\end{aligned}
$$

Let $c = \ell\sqrt{n}$. Then (4) holds because $\mathbb{P}((|Q\cap Q'\cap B| = i) : (|Q\cap Q'| = i)) = \frac{\frac{1}{2}\binom{\frac{n}{3}}{i}\binom{n-i}{c-i}\binom{n-c}{c-i}}{\frac{1}{2}\binom{n}{c}\binom{c}{i}\binom{n-c}{c-i}} = \frac{(\frac{n}{3})!(n-i)!}{(\frac{n}{3}-i)!n!} \le \left(\frac{1}{3}\right)^i$; (5) is by Proposition 3.14; (6) is because for the first part of the sum: $\binom{c}{i}\left(\frac{c}{n}\right)^i \le \frac{c^i}{i!}\frac{c^i}{n^i} = \frac{(\ell^2)^i}{i!}$ and $1+x \le e^x$, for the second: $\binom{c}{i}\left(\frac{c}{n}\right)^i\left(\frac{n-c}{n-i}\right)^{c-i} \le 1$; (7) holds since $e^{-\frac{(c-i)^2}{n-i}} \le e^{-\frac{(c-\frac{c}{6})^2}{n}} = e^{-\ell^2\left(\frac{5}{6}\right)^2}$ for $i \le \frac{c}{6}$; (8) is because $\sum_{i\ge 0}\frac{(\frac{\ell^2}{3})^i}{i!} = e^{\frac{\ell^2}{3}}$; and (9) is because $e < 3$ and $\ell \le \sqrt{n}$. □

Thus we have proved the following result:

Theorem 4.4 *$R(n, \ell\sqrt{n})$, where $\ell < \frac{2\sqrt{n}}{3}$, is a $(\frac{n}{3}, 2e^{-\frac{\ell^2}{6}})$-dissemination quorum system.*

Quality measures Load, fault tolerance, and failure probability do not depend on b or ε. (Recall that fault tolerance and failure probability relate to crash failures, while b is the number of Byzantine failures tolerated.) Hence, we have as before that the load $L(R(n, \ell\sqrt{n}))$ is $\frac{\ell}{\sqrt{n}}$, the fault tolerance $A(R(n, \ell\sqrt{n}))$ is $n - \ell\sqrt{n} + 1$, and the failure probability $F_p(R(n, \ell\sqrt{n}))$ is at most $e^{-2n\gamma^2}$, where $\gamma = 1 - \frac{\ell}{\sqrt{n}} - p$, for $p < 1 - \frac{\ell}{\sqrt{n}}$.

4.2 A (b, ε)-dissemination quorum system for $b = \alpha n$

Surprisingly, the same technique can be used to overcome *any* fraction α of Byzantine failures. In this case, the parameter ℓ needed to achieve a particular value of ε depends on the fraction α of servers that may simultaneously fail. Since our construction works, with appropriate choice of parameters, for $b = \alpha n$ for any constant fraction α of the servers, it is significantly more versatile than constructions of strict dissemination quorum systems, where an upper bound of $b = \lfloor\frac{n-1}{3}\rfloor$ limits the resilience. We present the result here for $\frac{1}{3} < \alpha < 1$, as the case $0 < \alpha \le \frac{1}{3}$ was already covered in the previous section. (A similar result holds for $0 < \alpha < 1$, but yields a more complicated ε).

Let $\frac{1}{3} < \alpha < 1$ and let $\varepsilon_\alpha = \frac{2}{1-\alpha}\alpha^{\ell^2\left(\frac{1-\sqrt{\alpha}}{2}\right)}$. An argument similar to Lemma 4.3 shows the following.

Lemma 4.5 *Let U be a universe of n servers, let B be a subset of U of size b where $b = \alpha n$ for some $\frac{1}{3} < \alpha < 1$, and let Q and Q' be quorums of size $\ell\sqrt{n}$ each chosen uniformly at random. Then $\mathbb{P}((Q \cap Q') \subseteq B) \leq \varepsilon_\alpha$.*

Theorem 4.6 *$R(n, \ell\sqrt{n})$, where $\ell < \sqrt{n}(1-\alpha)$, is a $(\alpha n, \varepsilon_\alpha)$-dissemination quorum system.*

Remarks:

- Since we assume that αn servers may fail, we must have $n - \ell\sqrt{n} > \alpha n$, or equivalently, $\ell < \sqrt{n}(1-\alpha)$. This limits the achievable intersection guarantee ε_α of $R(n, \ell\sqrt{n})$, for any particular system size n and Byzantine threshold αn.

- Note that $\mathcal{Q}$ and w do not directly depend on α. Hence, even if the fraction of Byzantine faults that may occur is not known, it is possible to use this construction, but the intersection parameter ε that is achieved will also be unknown. Furthermore, the construction has the desirable "graceful degradation" property that actual intersection probability will be better if fewer Byzantine faults actually occur.

5 (b, ε)-masking quorum systems

When Byzantine faults occur with data that is not self-verifying, it is necessary that correct servers be able to out-vote incorrect ones. Accordingly, a strict b-masking quorum system is defined to be one in which any two quorums intersect in at least $2b+1$ elements [MR98a]. As a result, when a client performs a read operation at some quorum Q, the value written in the last preceding write operation, say to Q', is returned by at least $b+1$ correct servers, namely servers in the set $(Q \cap Q') \setminus B$ where B is the set of faulty servers. Any other returned value is either an old value, which can be detected by its earlier timestamp, or a "made-up" value returned only by servers in B. So, if the client discards any values that were returned by b or fewer servers, and then chooses from the remaining values the one with the most recent timestamp, then the client is guaranteed to obtain the correct value [MR98a].

To formulate a probabilistic version of masking quorum systems, a natural place to start is the definition of an ε-intersecting quorum system. Mimicking that approach for masking quorum systems, we would require that any two selected quorums intersect in at least $2b+1$ elements with high probability. One advantage of such a definition is that there is no need to change the client access protocol: simply adopting the read and write protocols from [MR98a] would ensure that clients receive correct answers with high probability. However, this definition does not yield the performance benefits that the probabilistic approach did for regular quorum systems. In particular, it is not difficult to verify that the load for any such system with $b = \Theta(n)$ would be constant, which is poor.

The trouble with this definition is that it is stronger than necessary. If Q and Q' are the quorums used in a read and a previous write operation, respectively, and B is the set of faulty servers, then the definition requires $Q \cap Q' \setminus B$ to be so large that it is *impossible* for $Q \cap B$ to be of equal cardinality. For the correct answer to be "probably detectable" to a reading client, the set $Q \cap Q' \setminus B$ need only be of a size sufficiently large that it is *improbable* that $Q \cap B$ is of the same size or larger. To weaken this requirement, our definition of a (b, ε)-masking quorum system employs a threshold value k that is expected to be between $|Q \cap B|$ and $|Q \cap Q' \setminus B|$. Thus, a reading client that requires at least k

occurrences of a value in order to accept it as the outcome of the read operation will get the right value with high probability.

Definition 5.1 *Let $\mathcal{Q}$ be a set system over a universe U of size n, let w be an access strategy for $\mathcal{Q}$, and let $0 < \varepsilon < 1$ and integers $1 \leq k \leq n$ and $b > 0$ be given. The tuple $\langle \mathcal{Q}, w, k\rangle$ is a (b, ε)-*masking quorum system *if $A(\langle \mathcal{Q}, w\rangle) \geq b$ and*

$$\mathbb{P}(|Q \cap B| < k \ \wedge \ |Q \cap Q' \setminus B| \geq k) \quad = \sum_{\substack{Q,Q' \in \mathcal{Q} \\ |Q \cap B| < k \ \wedge \ |Q \cap Q' \setminus B| \geq k}} w(Q)w(Q') \quad \geq \quad 1 - \varepsilon,$$

for all $B \subseteq U$ such that $|B| = b$, where the probability is taken with respect to w.

We modify the access protocol as follows. Write operations are as before, but read operations now require a value that passes the threshold k:

Read: For a client to read x, it

1. chooses a quorum Q according to the strategy w,
2. queries each server in Q to obtain a set of value/timestamp pairs $V = \{\langle v_u, t_u\rangle\}_{u \in Q}$,
3. computes the set $V' = \{\langle v, t\rangle \ : \ \exists C \subseteq Q \ [|C| \geq k \ \wedge \ \forall u \in C \ [v_u = v \wedge t_u = t] \ ]\}$,
4. returns the pair $\langle v, t\rangle$ in V' with the highest timestamp, or $\perp$ if V' is empty.

Theorem 5.2 *Consider a multi-reader, single-writer variable replicated using the above access protocol with an (b, ε)-masking quorum system. If a read operation is not concurrent with any write operation and at most b Byzantine failures occur, then with probability at least $1 - \varepsilon$ the read returns the value written by the last preceding write operation.*

Proof: As in the proof of Theorem 3.2, the last write operation prior to the read operation has the highest timestamp of any write operation that precedes the read. Moreover, with probability at least $1 - \varepsilon$, the quorum Q' picked in this write operation and the quorum Q picked in the current read operation satisfy $|Q \cap B| < k \ \wedge \ |(Q \cap Q') \setminus B| \geq k$ where B is the set of actually faulty servers. So, with probability at least $1 - \varepsilon$, this value/timestamp pair appears in V' and thus the correct value will be returned by the read. □

Note that when an incorrect value is returned, it can either be an old or null value (if $|(Q \cap Q') \setminus B| < k$) or a value chosen by the faulty servers (if $|Q \cap B| \geq k$).

We define load, fault tolerance, and failure probability of (b, ε)-masking quorum systems in the standard manner:

Definition 5.3 *Let $\langle \mathcal{Q}, w, k\rangle$ be a (b, ε)-masking quorum system. Then*

- *The* load *of $\langle \mathcal{Q}, w, k\rangle$ is $L(\langle \mathcal{Q}, w, k\rangle) = L_w(\mathcal{Q})$.*
- *The* fault tolerance *of $\langle \mathcal{Q}, w, k\rangle$ is $A(\langle \mathcal{Q}, w, k\rangle) = A(\langle \mathcal{Q}, w\rangle)$.*
- *The* failure probability *of $\langle \mathcal{Q}, w, k\rangle$ is $F_p(\langle \mathcal{Q}, w, k\rangle) = F_p(\langle \mathcal{Q}, w\rangle)$.*

5.1 Lower bounds on the load

In addition to the general lower bound on load given in Theorem 3.9, which a fortiori holds in the case of (b, ε)-masking quorum systems, we present here a lower bound that depends on the number n of servers, on the threshold b of tolerated Byzantine faults, and on the error probability ε. This demonstrates a relationship between the number of faulty servers a system tolerates and the load it may achieve. Our main result in this section is the lower bound of Theorem 5.5. To prove this, we show that the expected quorum size must exceed b (up to a factor close to 1) in order to satisfy the intersection requirement, and then use "half" of Theorem 3.9.

Lemma 5.4 *Let $\langle \mathcal{Q}, w, k\rangle$ be a (b, ε)-masking quorum system. Then*

$$\mathbb{P}(|Q| > b) \geq \frac{1-2\varepsilon}{1-\varepsilon}.$$

Proof: Fix some $\widehat{Q} \in \mathcal{Q}$ with $|\widehat{Q}| \leq b$. (If no such $\widehat{Q}$ exists, then we are done). Then there exists some set $B_{\widehat{Q}}$ of size b such that $\widehat{Q} \subseteq B_{\widehat{Q}}$. By Definition 5.1,

$$1-\varepsilon \;\leq\; \mathbb{P}(|Q \cap B_{\widehat{Q}}| < k \wedge |(Q \cap Q') \setminus B_{\widehat{Q}}| \geq k) \;\leq\; \mathbb{P}(|Q \cap B_{\widehat{Q}}| < k) \;\leq\; \mathbb{P}(|Q \cap \widehat{Q}| < k).$$

Since $\widehat{Q}$ was chosen arbitrarily subject to the restriction that $|\widehat{Q}| \leq b$, we have the following bound on the conditional probability:

$$\mathbb{P}\left(|Q \cap \widehat{Q}| < k \;\middle|\; |\widehat{Q}| \leq b\right) \geq 1-\varepsilon. \tag{10}$$

Now using (10) and Definition 5.1, for any B with $|B| = b$ we have

$$\begin{aligned}
\varepsilon \;&\geq\; \mathbb{P}(|Q \cap B| \geq k \;\vee\; |Q \cap Q' \setminus B| < k) \\
&\geq\; \mathbb{P}(|Q \cap Q'| < k) \\
&\geq\; \mathbb{P}(|Q \cap Q'| < k \;\wedge\; |Q| \leq b) \\
&=\; \mathbb{P}\left(|Q \cap Q'| < k \;\middle|\; |Q| \leq b\right) \cdot \mathbb{P}(|Q| \leq b) \;\;\geq\;\; (1-\varepsilon) \cdot \mathbb{P}(|Q| \leq b)
\end{aligned}$$

Thus $\mathbb{P}(|Q| \leq b) \leq \varepsilon/(1-\varepsilon)$ and $\mathbb{P}(|Q| > b) \geq (1-2\varepsilon)/(1-\varepsilon)$. □

Theorem 5.5 *If $\langle \mathcal{Q}, w, k\rangle$ is a (b, ε)-masking quorum system, then*

$$L(\langle \mathcal{Q}, w, k\rangle) > \left(\frac{1-2\varepsilon}{1-\varepsilon}\right)\frac{b}{n}.$$

Proof: By Lemma 5.4,

$$\mathbb{E}[|Q|] \;>\; 0 \cdot \mathbb{P}(|Q| \leq b) + b \cdot \mathbb{P}(|Q| > b) \;\geq\; b\left(\frac{1-2\varepsilon}{1-\varepsilon}\right).$$

The theorem then follows from Lemma 3.10. □

5.2 A (b, ε)-masking quorum system construction

In this section, we give a family of (b, ε)-masking quorum system constructions. To do this, we modify our basic $R(n, q)$ construction by introducing the appropriate threshold k. The constructions in the resulting family differ in the number b of faulty servers that they tolerate and in the probability ε of reading an incorrect value. They all fit the following template:

Definition 5.6 *Let U be a universe of size n and let $0 \leq k \leq n$. Then $R_k(n, q)$ is the system $\langle \mathcal{Q}, w, k \rangle$ such that $\mathcal{Q} = \{Q \subseteq U : |Q| = q\}$ and $w(Q) = \frac{1}{|\mathcal{Q}|} = \frac{1}{\binom{n}{q}}$.*

In the following sections, leading to the precise statement of Theorem 5.10, we take $k = q^2/2n$ and prove that $R_k(n, q)$ is indeed a (b, ε)-masking quorum system for large b and vanishingly small ε. In order to do this, we need to specify the relationship between the quorum size q and the number b of tolerated Byzantine faults. We quantify this relationship by defining

$$\ell = q/b,$$

and we will specify the ratio ℓ later. We are further required to specify an error probability ε so that when two quorums Q, Q' are chosen according to w then

$$\mathbb{P}(|Q \cap B| < k \wedge |Q \cap Q' \setminus B| \geq k) \geq 1 - \varepsilon \tag{11}$$

for any set B of faulty servers with $|B| = b$.

5.3 On the choice of the threshold k

We start by computing two expectations which will be central in the analysis, and by showing that our choice of $k = q^2/2n$ is a good one. For any fixed set of faulty servers B we define two random variables:

1. $X = |Q \cap B|$, and
2. $Y = |Q \cap Q' \setminus B|$.

Then (11) can be written as

$$\mathbb{P}(X \geq k \ \vee \ Y < k) \leq \varepsilon. \tag{12}$$

We observe that for any fixed set of faulty servers B, X is a *hypergeometric* random variable $X \sim H(q/\ell, n, q)$. This is because the q members of Q are sampled without replacement from a universe of size n containing $b = q/\ell$ faulty servers. Therefore

$$\mathbb{E}[X] = \frac{q^2}{\ell n} \tag{13}$$

(for instance, see [Fel67, p. 233]).

We can compute the expectation of Y directly. Fix the set B. For each element $u \notin B$ define an indicator random variable I_u such that $I_u = 1$ if $u \in Q \cap Q' \setminus B$ and $I_u = 0$ otherwise. For such u we have $\mathbb{P}(I_u = 1) = q^2/n^2$ since Q and Q' are chosen independently. By linearity of expectation,

$$\mathbb{E}[Y] = \sum_{u \in U \setminus B} \mathbb{E}[I_u] = \sum_{u \in U \setminus B} \mathbb{P}(I_u = 1) = (n - b)\frac{q^2}{n^2} = \frac{q^2}{n}\left(1 - \frac{q}{\ell n}\right). \tag{14}$$

Taking (13) and (14) we see that the threshold k needs to obey

$$\frac{q^2}{\ell n} < k < \frac{q^2}{n}\left(1 - \frac{q}{\ell n}\right),$$

for otherwise (12) shows that we cannot hope for an error probability ε that decreases with n. Our choice of $k = \frac{q^2}{2n}$ fits the requirement for all $\ell > 2$.

5.4 Bounding the error probability ε

In order to bound ε, we analyze the probabilities of the two events $X \geq k$ (too many faulty servers accessed) and $Y < k$ (too few up-to-date servers accessed), for $k = q^2/2n$.

Lemma 5.7 *Let X, k and $\ell > 2$ be as in Section 5.3 and define*

$$\rho_1(\ell) = \begin{cases} (\ell/2 - 1)^2/4\ell & \text{if } 2 < \ell \leq 4e, \\ 1/3 & \text{if } \ell > 4e. \end{cases}$$

Then

$$\mathbb{P}(X \geq k) \leq \exp\left(- \rho_1(\ell)\frac{q^2}{n}\right).$$

For the proof, we need to bound the tail of the hypergeometric distribution. In [Hoe63], Hoeffding derives such a bound as a special case of more general results (see also [Chv79]). However, we obtain a tighter bound by comparing X with the sum of *independent* Bernoulli random variables. We use the following result, which is a consequence of [Hoe63, Theorem 4].

Proposition 5.8 *Let $\hat{x}_i \sim B(\frac{q}{\ell n})$ be independent Bernoulli random variables for $i = 1, \ldots, q$, and let $\widehat{X} = \sum_{i=1}^{q} \hat{x}_i$. Then $\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \mathbb{P}(\widehat{X} - \mathbb{E}[\widehat{X}] \geq t)$ for any $t > 0$.*

Remark: The expectations $\mathbb{E}[X]$ and $\mathbb{E}[\widehat{X}]$ are equal; however, the variances are not: $V[X] < V[\widehat{X}]$.

Proof of Lemma 5.7: We use Chernoff's bound to bound the deviation of $\widehat{X}$ from its expectation. That is:

$$\begin{aligned} \mathbb{P}(\widehat{X} > (1+\gamma)\mathbb{E}[\widehat{X}]) &\leq e^{-\mathbb{E}[\widehat{X}]\gamma^2/4} && \text{when } 0 < \gamma \leq 2e-1, \text{ and} \\ \mathbb{P}(\widehat{X} > (1+\gamma)\mathbb{E}[\widehat{X}]) &\leq 2^{-(1+\gamma)\mathbb{E}[\widehat{X}]} && \text{when } \gamma > 2e-1, \end{aligned}$$

(cf. [MR95, p. 72]). In our case, $\mathbb{E}[\widehat{X}] = \mathbb{E}[X] = q^2/\ell n$ by (13). Setting $k = (1+\gamma)\mathbb{E}[\widehat{X}]$ and substituting $k = q^2/2n$ yields $\gamma = \ell/2 - 1$. Then using Proposition 5.8 and the above Chernoff bounds for the appropriate ranges of ℓ we have:

$$\begin{aligned} \mathbb{P}(X \geq k) \;\leq\; \mathbb{P}\left(\widehat{X} > (1+\gamma)\mathbb{E}[\widehat{X}]\right) &\leq \exp\left(-\frac{q^2}{\ell n}\frac{(\ell/2-1)^2}{4}\right) && \text{when } 2 < \ell \leq 4e, \text{ and} \\ \mathbb{P}(X \geq k) &\leq 2^{-\frac{\ell}{2}\frac{q^2}{\ell n}} < \exp\left(-\frac{q^2}{3n}\right) && \text{when } \ell > 4e. \end{aligned}$$

□

We now turn to bounding $\mathbb{P}(Y < k)$:

Lemma 5.9 *Let Y, k and $\ell > 2$ be as in Section 5.3, and define*

$$\rho_2(\ell) = \frac{(\ell-2)^2}{8\ell(\ell-1)}.$$

Then

$$\mathbb{P}(Y < k) \le \exp\left(-\rho_2(\ell)\frac{q^2}{n}\right).$$

Proof: Consider a random variable $Z \sim H(q-b, n, q)$. It is not hard to show that when $k < \mathbb{E}[Z]$, we have

$$\mathbb{P}(Y < k) \le \mathbb{P}(Z < k) \tag{15}$$

and so it suffices to show the result for $\mathbb{P}(Z < k)$. Intuitively, Z can be thought of as $|Q \cap Q' \setminus B|$ in the case where $B \subseteq Q'$, i.e., when there are the fewest possible correct servers in the write quorum Q'. Thus, (15) is unsurprising.

To bound $\mathbb{P}(Z < k)$, first note that

$$\mathbb{E}[Z] = \frac{q(q-b)}{n} = \frac{q^2}{n}\left(\frac{\ell-1}{\ell}\right), \tag{16}$$

which is somewhat smaller than $\mathbb{E}[Y]$ shown in (14). Now we repeat the argument of Lemma 5.7. Namely, we rely on Hoeffding's result to compare Z with a sum $\widehat{Z}$ of independent variables (sampling with replacement) which has $\mathbb{E}[\widehat{Z}] = \mathbb{E}[Z]$. Then we use the following Chernoff bound on the lower tail.

$$\mathbb{P}(\widehat{Z} < (1-\delta)\mathbb{E}[\widehat{Z}]) \le e^{-\mathbb{E}[Z]\delta^2/2} \text{ when } 0 \le \delta \le 1.$$

In our case $\mathbb{E}[\widehat{Z}] = \frac{q^2}{n}\left(\frac{\ell-1}{\ell}\right)$ by (16). Setting $k = (1-\delta)\mathbb{E}[\widehat{Z}]$ and substituting $k = \frac{q^2}{2n}$ yields $\delta = \frac{\ell-2}{2(\ell-1)}$ (and note that $0 \le \delta \le 1$ when $\ell > 2$). Using (15) and the above Chernoff bounds we get

$$\mathbb{P}(Y < k) \le \mathbb{P}(Z < k) \le \exp\left(-\frac{q^2}{n}\left(\frac{\ell-1}{2\ell}\right)\left(\frac{\ell-2}{2(\ell-1)}\right)^2\right) = \exp\left(-\frac{q^2(\ell-2)^2}{n8\ell(\ell-1)}\right). \qquad \square$$

With all the groundwork prepared we can now state a fully-qualified theorem regarding $R_k(n, q)$.

Theorem 5.10 *Let $2 < \ell < n/b$, define $\rho_1(\ell)$, $\rho_2(\ell)$ as in Lemmas 5.7 and 5.9, let $k = \frac{q^2}{2n}$, and let $\varepsilon = 2\exp\left(-\frac{q^2}{n}\min\{\rho_1(\ell), \rho_2(\ell)\}\right)$. Then $R_k(n, \ell b)$ is a (b, ε)-masking quorum system.*

Proof: We need to show

$$\mathbb{P}(|Q \cap B| < k \wedge |Q \cap Q' \setminus B| \ge k) \ge 1 - \varepsilon.$$

Hence, by the union bound, it is sufficient to show

$$\mathbb{P}(|Q \cap B| \ge k) + \mathbb{P}(|Q \cap Q' \setminus B| < k) \le \varepsilon,$$

which follows from Lemmas 5.7 and 5.9. $\square$

Remarks:

- Theorem 5.10 shows that for any quorum size q satisfying $q = \omega(\sqrt{n})$, the error probability ε vanishes as $n \to \infty$ for all $b < q/2$.
- The factors ρ_1 and ρ_2 are quite manageable. For example, when $\ell = 3$ we have $\varepsilon \leq 2e^{-q^2/48n}$, and when $\ell = 20$ we have $\varepsilon \leq 2e^{-q^2/10n}$.
- The choice of the threshold $k = q^2/2n$ was somewhat arbitrary. We have also analyzed the case where k is chosen so as to balance the bounds on $\mathbb{P}(X \geq k)$ and $\mathbb{P}(Y < k)$. The resulting computations (omitted) are more complicated, and yield marginally better factors ρ_1, ρ_2.

5.5 Properties of the system

As shown in Theorem 5.10, $R_k(n,q)$ works for $q = \ell b$ for any b and ε, provided that $2 < \ell < n/b$. It also does well in our performance measures in most cases. Note that if ℓ is chosen to be larger, then the error probability ε is lower. In particular, it is desirable to choose $\ell = \omega(\sqrt{n}/b)$ so that ε vanishes as n grows. In contrast, if ℓ is chosen to be smaller, then the quorums are smaller, and so the load and failure probability are better. We show that if b is not too large, it is possible to balance these conflicting requirements on ℓ.

Load $R_k(n,q)$ has load $q/n = \ell b/n$, which is within a factor of $\ell\left(\frac{1-\varepsilon}{1-2\varepsilon}\right)$ of the bound in Theorem 5.5. Because we can choose ℓ to be a constant when $b = \omega(\sqrt{n})$, our construction is asymptotically load-optimal (as a function of b and n) for $b = \omega(\sqrt{n})$. However for smaller b's, if we choose ℓ so that $q = \omega(\sqrt{n})$ in order to obtain a vanishingly small error probability, then our construction does not meet the lower bound on load when $b = O(\sqrt{n})$. Closing this gap is left as an open problem.

It is also interesting to note that the load of $R_k(n, \ell b)$ beats the load of any strict masking quorum system in the case when both $b = \omega(\sqrt{n})$ and $b = o(n)$. More precisely, the load of any strict masking quorum system is $\Omega(\sqrt{b/n})$ [MRW00]. So, if $b = \Theta(\sqrt{n})$, then choosing $\ell = n^{1/5}$, for example, yields a load of $O(n^{-0.3})$, which beats the lower bound of $\Omega(n^{-0.25})$ for any strict masking quorum system. If $b = \omega(\sqrt{n})$ and $b = o(n)$, then we can choose ℓ to be a constant, yielding a load of $O(b/n) = o(\sqrt{b/n})$.

Failure probability The $R_k(n,q)$ system is uniform, so by symmetry all of its quorums are high quality. Therefore, it has optimal fault tolerance, since for the system to fail, at least $n - q + 1 = \Theta(n)$ servers must crash. Using Chernoff's bound, the probability of this occurring is

$$F_p(R_k(n,q)) \;=\; \mathbb{P}(\#\text{fail} > n-q) \;\leq\; e^{-2n\left(1-\frac{q}{n}-p\right)^2} \;=\; e^{-\Omega(n)}$$

for all $p < 1 - \frac{q}{n}$. This bound shows that our construction cannot be outperformed (asymptotically) with respect to failure probability by any strict quorum system, and if $\frac{1}{2} < p < 1 - \frac{q}{n}$ then the failure probability of our construction is provably better than that of any strict quorum system [PW95]. The extent of our construction's improvement in failure probability over strict quorum systems is best demonstrated by precise failure probability comparisons, rather than by asymptotic bounds. These comparisons are given in Section 6.

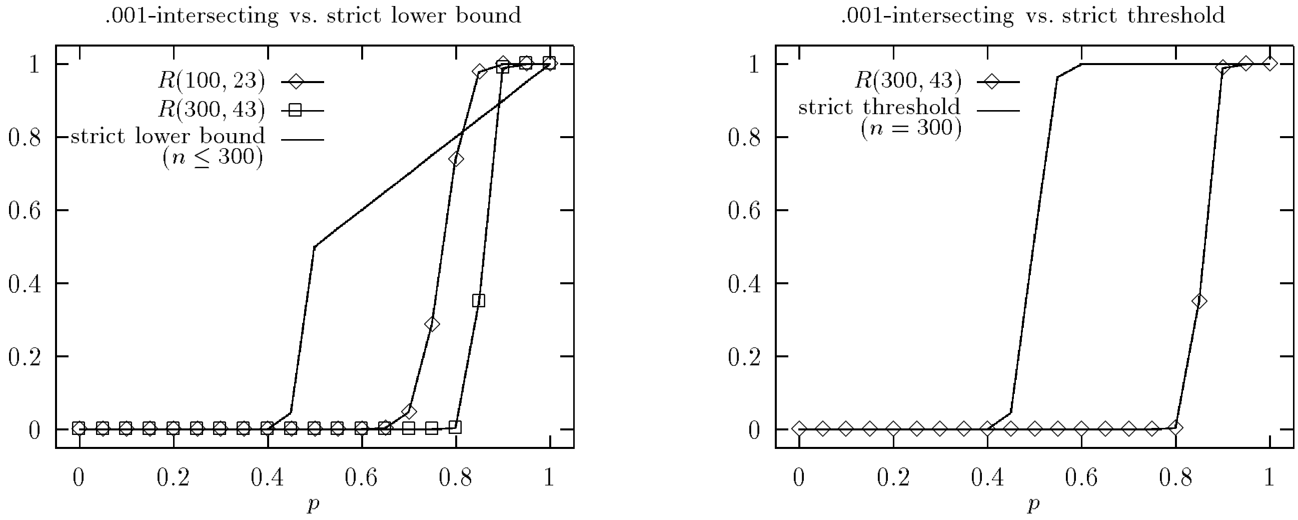

Figure 1: Failure probabilities of probabilistic quorum systems.

6 Concrete comparisons

In order to ascertain how our probabilistic quorum systems compare to strict quorum systems for particular system sizes, in this section we perform such comparisons, focusing specifically on the measures of failure probability, fault tolerance, and quorum size.

Figures 1–3 compare the failure probabilities of our probabilistic quorum systems against the failure probabilities that can be achieved by strict quorum systems. The graphs on the left side of the figures plot the failure probability of our constructions for system sizes $n = 100$ and $n = 300$ against the lower bound on failure probability for any strict quorum system with $n \leq 300$ [BG87, PW95].[3] These graphs show that our constructions can beat this lower bound—and hence the failure probability of any strict quorum system—for many values of p in settings where $b = \sqrt{n}$.

These comparisons are further pronounced on the right side of Figures 1–3, where our constructions are directly compared to the corresponding "threshold" strict quorum constructions for intersection (quorums are of size $\lceil \frac{n+1}{2} \rceil$ in Figure 1), b-dissemination (quorums are of size $\lceil \frac{n+b+1}{2} \rceil$ in Figure 2), or b-masking (quorums are of size $\lceil \frac{n+2b+1}{2} \rceil$ in Figure 2). Like our probabilistic constructions, each of these strict quorum constructions has asymptotic failure probability of $e^{-\Omega(n)}$. However, the figures show that our probabilistic constructions decisively beat them for concrete system sizes. We remind the reader that these gains in failure probability come at the cost of risking intersection with some probability ε. Each of the probabilistic systems depicted in Figures 1–3 guarantees $\varepsilon \leq .001$.

The other measures that we consider in this section are quorum size and fault tolerance. A smaller quorum size generally indicates lower load and better efficiency in accessing the service, and accordingly the quorum size is an important measure in practice. In Tables 2–4 we detail the quorum size and fault

[3] The lower bound is formed as the minimum of two curves: the failure probabilities of the simple majority system when $p < 1/2$ and a singleton server when $p \geq 1/2$. That is, for large values of p, the most available strict quorum system is one with a single server. Relaxing the intersection property of quorums to hold only probabilistically does not change this fact, but it significantly raises the "crossover value" of p at which a singleton begins to provide the best availability.

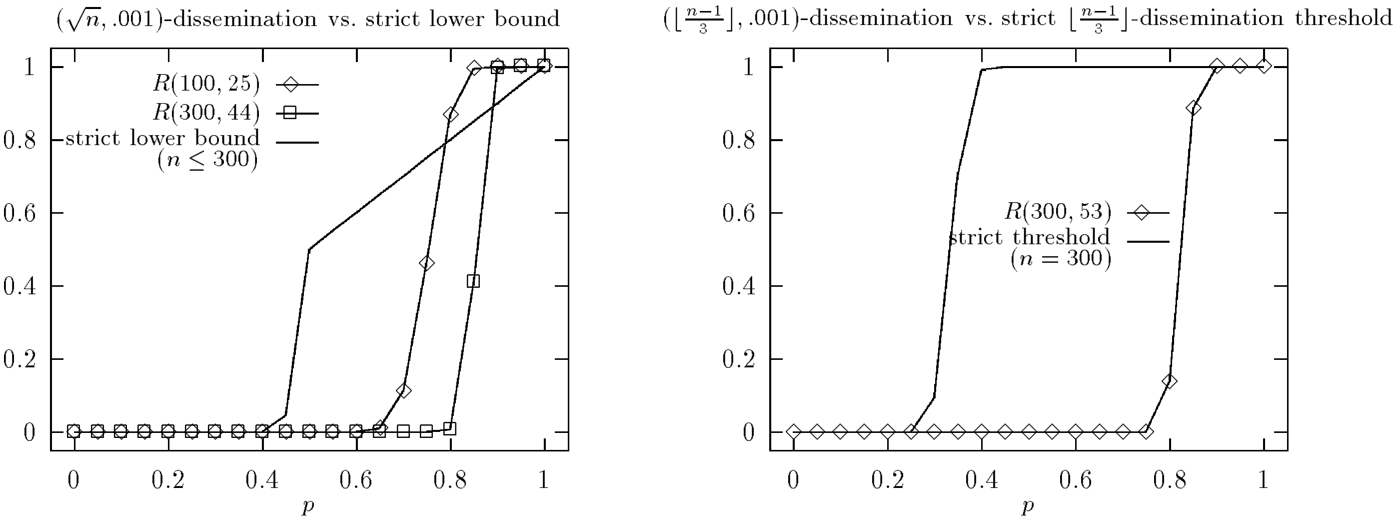

Figure 2: Failure probabilities of probabilistic dissemination quorum systems.

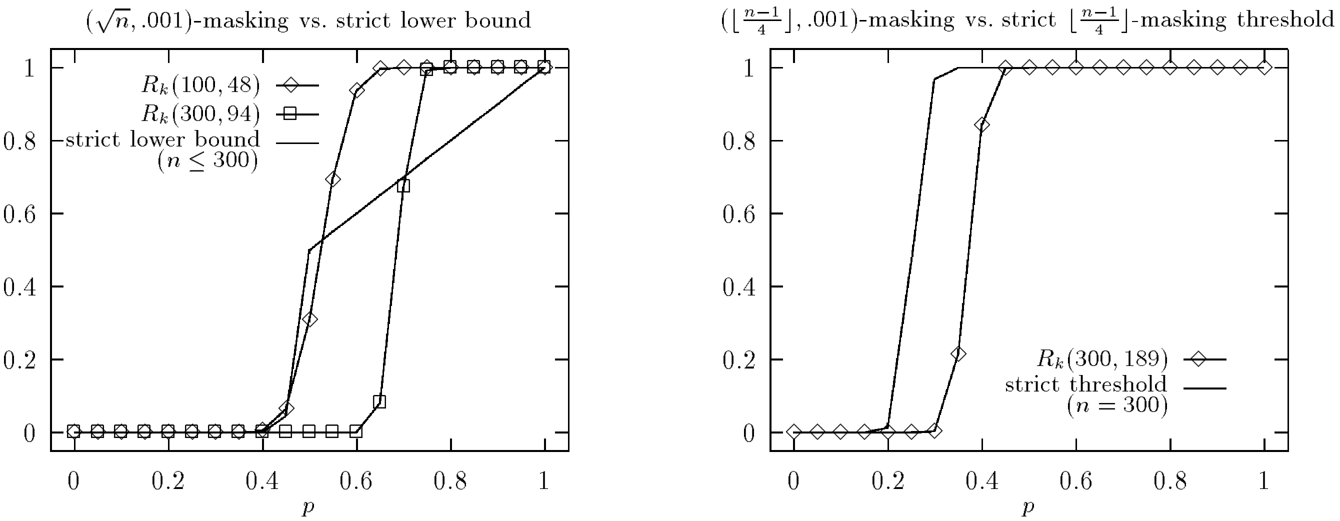

Figure 3: Failure probabilities of probabilistic masking quorum systems.

n	ℓ	**ε-intersecting** quorum size	fault tolerance	**threshold** quorum size	fault tolerance	**grid** quorum size	fault tolerance
25	1.80	9	17	13	13	9	5
100	2.20	22	79	51	51	19	10
225	2.40	36	190	113	113	29	15
400	2.45	49	352	201	201	39	20
625	2.48	62	564	313	313	49	25
900	2.50	75	826	451	451	59	30

Table 2: Properties of various quorum systems.

n	b	ℓ	**(b, ε)-dissemination** quorum size	fault tolerance	**threshold** quorum size	fault tolerance	**grid** quorum size	fault tolerance
25	2	2.20	11	15	14	12	16	5
100	4	2.40	24	77	53	48	36	10
225	7	2.47	37	189	166	60	56	15
400	9	2.50	50	351	205	196	111	20
625	12	2.52	63	563	319	307	141	25
900	14	2.57	77	824	458	443	171	30

Table 3: Properties of various dissemination quorum systems.

tolerance of probabilistic quorum systems for various system sizes and failure assumptions, contrasted against several types of strict quorum systems.

Table 2 shows the quorum sizes and fault tolerance of three quorum constructions: our ε-intersecting construction, the strict threshold quorum construction (quorums of size $\lceil \frac{n+1}{2} \rceil$), and grid quorums [Mae85] (servers are laid out in a $\sqrt{n} \times \sqrt{n}$ grid; each quorum consists of one row and one column). The table clearly demonstrates the superiority of an ε-intersecting quorum system in achieving excellent fault tolerance simultaneously with small quorum size (and hence, with good load). The advantage becomes more pronounced as the system size grows.

Table 3 shows three dissemination constructions: our (b, ε)-dissemination quorum system, a strict threshold [MR98a] (quorums of size $\lceil \frac{n+b+1}{2} \rceil$) and a grid construction [MRW00] (each quorum is $\sqrt{\frac{b+1}{2}}$ rows and columns). In calculating these numbers, we assumed that $b = (\sqrt{n} - 1)/2$, as this is the largest b for which all the constructions in the table work. The corresponding masking constructions are shown in Table 4. All the probabilistic constructions we show achieve a consistency guarantee of 0.999 or higher ($\varepsilon \leq .001$) for the appropriate intersection requirement; ℓ was chosen as small as possible subject to this restriction.

		(b,ε)-**masking**			**threshold**		**grid**	
n	b	ℓ	quorum size	fault tolerance	quorum size	fault tolerance	quorum size	fault tolerance
25	2	3.00	15	11	15	11	16	5
100	4	3.80	38	63	55	46	51	10
225	7	4.27	64	162	120	106	81	15
400	9	4.70	94	307	210	191	144	20
625	12	4.92	123	503	325	301	184	25
900	14	5.07	152	749	465	436	224	30

Table 4: Properties of various masking quorum systems.

7 Conclusion

In this paper, we used a probabilistic approach in the construction of quorum systems and obtained a new class of set systems, called probabilistic quorum systems. We formally defined these systems, showed that they can be used to implement robust replicated data, and proved lower bounds on the load of any such system. We showed a generic construction of probabilistic quorum systems that have optimal load but far exceed the resilience of any known strict quorum system. With modified parameters, we were able to apply the general construction also to Byzantine environments, demonstrating a dramatic improvement in both load and availability for this model.

Acknowledgments

We thank Cynthia Dwork and the anonymous referees for their helpful comments.

References

[AE91] D. Agrawal and A. El Abbadi. An efficient and fault-tolerant solution for distributed mutual exclusion. *ACM Transactions on Computer Systems*, 9(1):1–20, 1991.

[AES97] D. Agrawal, A. El-Abbadi, and R. Steinke. Epidemic algorithms in replicated databases. In *Proc. 16th ACM SIGACT-SIGMOD Symp. Princip. of Database Systems (PODS)*, Tucson, Arizona, May 1997.

[AR92] Y. Aumann and M. Rabin. Clock construction in fully asynchronous parallel systems and PRAM simulation. In *Proceedings of the 33rd IEEE Symposium on Foundations of Computer Science*, pages 147–156, October 1992.

[BG86] D. Barbara and H. Garcia-Molina. The vulnerability of vote assignments. *ACM Transactions on Computer Systems*, 4(3):187–213, 1986.

[BG87] D. Barbara and H. Garcia-Molina. The reliability of vote mechanisms. *IEEE Transactions on Computers*, C-36:1197–1208, October 1987.

[BHG87] P. A. Bernstein, V. Hadzilacos, and N. Goodman. *Concurrency Control and Recovery in Database Systems.* Addison-Wesley, Reading, MA, 1987.

[Baz96] R. Bazzi. Planar quorums. In *Proceedings of the 10th International Workshop on Distributed Algorithms (WDAG)*, pages 251–568, Bologna, Italy, October 1996.

[Baz97] R. A. Bazzi. Synchronous Byzantine quorum systems. In *Proceedings of the 16th ACM Symposium on Principles of Distributed Computing*, pages 259–266, August 1997.

[Baz99] R. Bazzi. Nonblocking Asynchronous Byzantine Quorum Systems. In *Proceedings of the 13th International Workshop on Distributed Algorithms (DISC)*, Bratislava, Slovakia, 1999.

[CAA90] S. Y. Cheung, M. H. Ammar, and M. Ahamad. The grid protocol: A high performance scheme for maintaining replicated data. In *Proceedings of the 6th IEEE International Conference on Data Engineering*, pages 438–445, 1990.

[Chv79] V. Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25:285–287, 1979.

[CLR89] T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithms.* The MIT Press. 1989.

[DGH+87] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry. Epidemic algorithms for replicated database maintenance. In *Proceedings of the 6th ACM Symposium on Principles of Distributed Computing*, pages 1–12, 1987.

[ET89] A. El Abbadi and S. Toueg. Maintaining availability in partitioned replicated databases. *ACM Transactions on Database Systems*, 14(2):264–290, June 1989.

[Fel67] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley & Sons, New York, 3rd edition, 1967.

[GB85] H. Garcia-Molina and D. Barbara. How to assign votes in a distributed system. *Journal of the ACM*, 32(4):841–860, October 1985.

[Gif79] D. K. Gifford. Weighted voting for replicated data. In *Proceedings of the 7th Symposium on Operating Systems Principles*, pages 150–162, 1979.

[HL99] Z. J. Haas and B. Liang. Ad hoc mobility management with uniform quorum systems. *IEEE/ACM Transactions on Networking* 7(2):228–240, April 1999.

[HB95] M. Hayden and K. P. Birman. Achieving critical reliability with unreliable components and unreliable glue. Cornell University Technical Report, TR95-1493, March, 1995.

[Her86] M. Herlihy. A quorum-consensus replication method for abstract data types. *ACM Transactions on Computer Systems*, 4(1):32–53, February 1986.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[IS92] A. Israeli and A. Shaham. Optimal multi-write multi-reader atomic register. In *Proceedings of the 11th ACM Symposium on Principles of Distributed Computing*, pages 71–82, 1992.

[KB94] N. Krishnakumar and A. J. Bernstein. Bounded ignorance: A technique for increasing concurrency in a replicated system. *ACM Transactions on Database Systems*, 19(4):586–625, December 1994.

[KPRR92] Z. Kedem, K. Palem, M. Rabin, and A. Raghunathan. Efficient program transformations for resilient parallel computation via randomization. In *Proceedings of the 24th ACM Symposium on Theory of Computing*, pages 306–317, May 1992.

[Lam86] L. Lamport. On interprocess communication (part II: algorithms). *Distributed Computing* 1:86–101, 1986.

[Mae85] M. Maekawa. A $\sqrt{n}$ algorithm for mutual exclusion in decentralized systems. *ACM Transactions on Computer Systems*, 3(2):145–159, 1985.

[MMR99] D. Malkhi, Y. Mansour and M. Reiter. On Propagating Updates in a Byzantine Environment. In *Proceedings of the 18th IEEE Symposium on Reliable Distributed Systems*, pages 134–143, October 1999, EPFL, Lausanne, Switzerland.

[MMR97] D. Malkhi, M. Merritt and O. Rodeh. Secure multicast in a WAN. *Distributed Computing*, 13(1):19–28, 2000.

[MR98a] D. Malkhi and M. Reiter. Byzantine quorum systems. *Distributed Computing*, 11(4):203–213, 1998.

[MR98b] D. Malkhi and M. Reiter. Secure and scalable replication in Phalanx. In *Proceedings of the 17th IEEE Symposium on Reliable Distributed Systems*, pages 51–60, October 1998.

[MRW00] D. Malkhi, M. Reiter and A. Wool. The load and availability of Byzantine quorum systems. SIAM Journal of Computing 29(6):1889–1906, 2000.

[MRW97b] D. Malkhi, M. Reiter, and R. Wright. Probabilistic quorum systems. In *Proceedings of the 16th ACM Symposium on Principles Distributed Computing*, pages 267–273, August 1997.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[NW98] M. Naor and A. Wool. The load, capacity and availability of quorum systems. *SIAM Journal of Computing*, 27(2):423–447, April 1998.

[PW95] D. Peleg and A. Wool. The availability of quorum systems. *Information and Computation*, 123(2):210–223, 1995.

[PW96] D. Peleg and A. Wool. How to be an efficient snoop, or the probe complexity of quorum systems. In *Proceedings of the 15th ACM Symposium on Principles of Distributed Computing (PODC)*, pages 290–299, Philadelphia, 1996.

[PW97] D. Peleg and A. Wool. Crumbling walls: A class of practical and efficient quorum systems. *Distributed Computing*, 10(2):87–98, 1997.

[PL91] C. Pu and A. Leff. Replica control in distributed systems: An asynchronous approach. In *Proceedings of the 1991 ACM SIGMOD International Conference on Management of Data*, pages 377–386, 1991.

[SSS95] J. P. Schmidt, A. Siegel, and A. Srinivasan. Chernoff-Hoeffding bounds for applications with limited independence. *SIAM Journal of Discrete Mathematics*, 8, 1995.

[Tho79] R. H. Thomas. A majority consensus approach to concurrency control for multiple copy databases. *ACM Transactions on Database Systems*, 4(2):180–209, 1979.

[WA92] M. H. Wong and D. Agrawal. Tolerating bounded inconsistency for increasing concurrency in database systems. In *Proc. 11th ACM SIGACT-SIGMOD Symp. Princip. of Database Systems (PODS)*, pages 236–245, San Diego, CA, June 1992.