Probabilistic Quorum Systems

Dahlia Malkhi Michael Reiter Rebecca Wright

AT&T Labs—Research
180 Park Avenue, Florham Park, NJ 07932-0971
{dalia,reiter,rwright}@research.att.com

Abstract

Services replicated using a quorum system allow operations to be performed at only a subset (quorum) of the servers, and ensure consistency among operations by requiring that any two quorums intersect. In this paper we explore the consequences of requiring this intersection property to hold only with very high probability. We show that doing so can offer dramatic improvements in the performance and availability of the service, both for services tolerant of benign server failures and services tolerant of arbitrary (Byzantine) ones. We also prove a lower bound on the performance that can be achieved with this technique.

1 Introduction

Quorums are tools for increasing the availability and efficiency of replicated services. A *quorum system* is a set of subsets of servers, every pair of which intersect. Intuitively, the intersection property guarantees that if a "write" operation is performed at one quorum, and later a "read" operation 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—as in the State Machine Approach [Sch90]—using quorums reduces the load on servers and increases service availability despite server crashes.

Quorum systems have been extensively studied and measured (cf., [Gif79, Tho79, Mae85, GB85, Her86, BG87, ET89, CAA90, AE91, NW94, PW95a, PW95b]). Three measures of a quorum system will be of particular interest in this paper: load [NW94], fault tolerance [BG87], and failure probability (see [BG87, PW95b]). The *load* of a quorum system is a measure of its efficiency. Intuitively, the load is the rate at which the busiest server will be accessed. The *fault tolerance*, also called the *availability*, of a system is the number of servers that can fail without disabling the system. A related measure is *failure probability*, the probability that the system is disabled. (Load, fault tolerance, and failure probability will be defined precisely in Section 2.) The fault tolerance of any quorum system is bounded by half of the number of servers. Moreover, as we show in Section 3, there is a tradeoff between low load and good fault-tolerance (and failure probability), and in fact it is impossible to simultaneously achieve both optimally.

To break these limitations, in this paper we relax the intersection property of a quorum system so that "quorums" chosen according to a specified strategy intersect only with very high probability. We accordingly call these *probabilistic quorum systems*, and henceforth refer to systems that satisfy the original definition of quorums as *strict*. Probabilistic quorum systems admit the possibility, albeit small, that two operations will be performed at non-intersecting quorums, in which case consistency of the system may suffer.

We show, however, that even a small relaxation of consistency can yield dramatic improvements in the fault tolerance and failure probability of the system, while the load remains essentially unchanged. Probabilistic quorum systems are thus most suitable for use when availability of operations despite the presence of faults is more important than certain consistency. This might be the case if the cost of inconsistent operations is high but not irrecoverable, or if obtaining the most up-to-date information is desirable but not critical, while having no information may have heavier penalties. For example, probabilistic quorum systems could be useful wherever quick access to an answer that is likely to be correct can greatly improve efficiency in the normal case, and the cost of dealing with incorrect answers when they do occur is not too high. Lampson [Lam83] describes this kind of mechanism as *hints*, and describes several systems that use such hints [LS79, MW77, Smi81]. More recently, hints have been used in mobile systems to find more direct routes to the current location of a mobile device [JP96, CP96].

1.1 Related Work

Though ours is the first work to study probabilistic quorum systems as such, the use of replicated variables to give probably correct results has proved useful in other contexts. Two examples of this are used to efficiently simulate a PRAM using an asynchronous system [KPRR92, AR92]. Specifically, Kedem *et al.* [KPRR92] use a replicated variable 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 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.

Malkhi et al. use essentially a hybrid construction of quorums, combining randomized and deterministic choice of members, to solve the problem of secure reliable multicast in a large network with many components [MMR97]. Their work focuses on a protocol that enforces random choice of members by involving a set of deterministically chosen processes, whose size is constant, in every operation. Because of this, if any member of this set fails, the probabilistic "quorums" become inaccessible, in which case their protocol reverts to

Appears in PODC '97: Sixteenth Annual ACM Symposium on Principles of Distributed Computing

strict quorums.

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. Our techniques are consequently very different from those used in these previous works. A possible direction of future work is to determine whether our techniques could be useful in the context of PRAM simulation.

1.2 Our Results

We begin by exploring the limits of probabilistic quorum systems. In particular, we show a lower bound on the load of probabilistic quorum systems that is within a small constant fraction of the bound for strict systems. Thus, probabilistic quorum systems cannot yield substantial improvements on load in general.

In contrast, we show that probabilistic quorum systems *can* yield substantial improvements on load when high fault-tolerance is also needed. Strict quorum systems over n servers that achieve the optimal load of $\frac{1}{\sqrt{n}}$ can tolerate at most $\sqrt{n}$ faults, and more generally suffer from an inherent tradeoff between load and fault-tolerance, where improving one must come at the expense of the other. We show that this limitation does not hold for probabilistic quorum systems. Specifically, we explore probabilistic quorum systems for the case where up to a constant fraction of the servers fail, for any constant smaller than 1. We construct a probabilistic quorum system tolerant of such failures and that has a load of only $O(\frac{1}{\sqrt{n}})$. More precisely, we provide a generic construction whose load is $\frac{\ell}{\sqrt{n}}$, for any chosen parameter $\ell \leq \sqrt{n}$, such that the achieved consistency guarantee (probability of quorum intersection) is at least $1 - e^{-\ell^2}$. Thus, using probabilistic techniques, we break the tradeoff between low load and high fault tolerance, achieving optimal load with essentially limitless resiliency. In addition, our construction has failure probability better than any strict quorum system.

Relaxing consistency can also provide dramatic improvements in an environment in which servers may experience Byzantine failures. The intersection property of quorums does not suffice for maintaining consistency in this model, since two quorums may intersect in a set containing *faulty* servers only, who may deviate arbitrarily and undetectably from their assigned protocol. Therefore, stronger requirements are necessary in order to use quorums in Byzantine environments. For such environments, Malkhi and Reiter defined *(strict) dissemination quorum systems* [MR97] to support replicated servers that store *self-verifying* data, i.e., data that servers can suppress but not undetectably alter (e.g., digitally signed data). Briefly, in a t-dissemination quorum system, any two quorums intersect in $t+1$ servers. Dissemination quorums systems can be constructed only for $t \leq \lfloor \frac{n-1}{3} \rfloor$ arbitrarily faulty servers, and the load of a t-dissemination quorum system is at least $\sqrt{\frac{t+1}{n}}$. We define a *probabilistic dissemination quorum system* in an analogous way to the definition above, where a probabilistic consistency property replaces the dissemination consistency one. Once again, we are able to construct a probabilistic dissemination quorum system resilient to the Byzantine failure of any constant fraction of the system and with outstanding failure probability, for sufficiently large universes, whose load is $O(\frac{1}{\sqrt{n}})$. For large n, this construction provides considerable advantage over strict dissemination quorum system constructions.

The contributions of this paper can be summarized as follows.

- the introduction of *probabilistic quorum systems*
- a lower bound on the load of probabilistic quorum systems that is within a small constant fraction of the bound for strict quorum systems.
- a generic probabilistic quorum system construction that achieves asymptotically optimal load and fault tolerance, with arbitrarily high consistency.
- a modification of the construction to work for the case of Byzantine server failures.

The rest of this paper is structured as follows. We review the basic definitions of quorum systems and ways of measuring them in Section 2. Section 3 defines probabilistic quorum systems, proves a lower bound on the load of any such quorum system, and presents a construction of one that exhibits very good load, fault tolerance and failure probability. Section 4 introduces probabilistic dissemination quorum systems and provides a construction tolerant of the Byzantine failure of any constant fraction of the servers. We conclude in Section 5.

2 Preliminary definitions

In this section, we define precisely the concepts introduced in Section 1. Assume a *universe* U of servers, $|U| = n$.

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

Definition: A *(strict) quorum system* $\mathcal{Q}$ over a universe U is a set system over U such that for every $Q_1, Q_2 \in \mathcal{Q}$, $Q_1 \cap Q_2 \neq \emptyset$. Each $Q \in \mathcal{Q}$ is called a *quorum.* □

As discussed in Section 1, quorum systems are generally insufficient to guarantee consistency in case of Byzantine server failures. A *t-dissemination quorum system* increases quorum overlap to $t+1$ servers, which suffices to mask faulty server behavior for some types of data [MR97].[1]

Definition: A quorum system $\mathcal{Q}$ is a *t-dissemination quorum system* if for every $Q_1, Q_2 \in \mathcal{Q}$, $|Q_1 \cap Q_2| \geq t+1$. □

Intuitively, 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: An *access strategy* (or just *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 $\sum_{Q \in \mathcal{Q}} w(Q) = 1$. □

In this paper we consider several measures of quorum systems, including the load, fault tolerance, and failure probability of the system.

The load of a quorum system, defined in [NW94], captures the probability of accessing the busiest server in the best case. Load is a measure of efficiency; all other things equal, systems with lower load can process more requests than those with higher load.

[1] The original definition of [MR97] treats dissemination quorum systems more generally than we do here. The simplified definition presented here suffices for our purposes.

Definition: Let w be a strategy for a set system $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ over a universe U. For an element $u \in U$, 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. $\square$

Load is a best-case definition (of a worst-behavior property). 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.

Fault tolerance and failure probability capture the resiliency of the service 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: For a set system $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ define $\mathcal{S} = \{S \mid S \cap Q_i \neq \emptyset \text{ for all } 1 \leq i \leq m\}$. The *fault tolerance* of $\mathcal{Q}$ is $A(\mathcal{Q}) = \min_{S \in \mathcal{S}} |S|$. $\square$

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.

The failure probability $F_p(\mathcal{Q})$ of a quorum system is the probability that there exists a quorum containing no faulty servers, assuming that servers fail independently with probability p.

Definition: Assume that each server in U fails with probability p, and that server failures are independent. The *failure probability* $F_p(\mathcal{Q})$ of $\mathcal{Q}$ is the probability that every $Q \in \mathcal{Q}$ contains at least one faulty server. $\square$

A good failure probability $F_p(\mathcal{Q})$ for a strict quorum system $\mathcal{Q}$ has $\lim_{n \to \infty} F_p(\mathcal{Q}) = 0$ when $p < \frac{1}{2}$ [NW94]. For $p = \frac{1}{2}$, there exist strict quorum constructions with $F_p(\mathcal{Q}) = \frac{1}{2}$, whereas for $p > \frac{1}{2}$, $F_p(\mathcal{Q})$ tends to 1 for all strict quorum systems.

3 Probabilistic quorum systems

In this section, we show that relaxing the consistency requirement for quorum systems to require only that any two quorums intersect with high probability can yield dramatic improvements in the fault tolerance of the system.

There is a tradeoff between load and fault tolerance in strict quorum systems. It is known that for any strict 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 $\mathcal{Q}$ [NW94]. In particular, this implies that for any strict quorum system $\mathcal{Q}$, $L(\mathcal{Q}) \geq \frac{1}{\sqrt{n}}$. Moreover, the intersection property implies that the failure of any full quorum in $\mathcal{Q}$ will disable all quorums (i.e., $A(\mathcal{Q}) \leq c(\mathcal{Q})$), and so by the aforementioned lower bound on load, $A(\mathcal{Q}) \leq nL(\mathcal{Q})$. It follows that any strict quorum system with optimal load of $\Theta(\frac{1}{\sqrt{n}})$ has fault tolerance of (only) $O(\sqrt{n})$.

We show that probabilistic quorums are not subject to this tradeoff by demonstrating a probabilistic quorum system over a universe of n elements that has a load of $O(\frac{1}{\sqrt{n}})$ and fault tolerance of $\Omega(n)$, with an increasing guarantee of consistency 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 any strict system.

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

Definition: Let $\mathcal{Q}$ be a set system, w an access strategy for $\mathcal{Q}$, and ϵ a constant, $0 < \epsilon < 1$. The tuple $\langle \mathcal{Q}, w, \epsilon \rangle$ is a *probabilistic quorum system* if

$$\sum_{Q, Q' : (Q \cap Q') \neq \emptyset} w(Q)w(Q') \;\geq\; 1 - \epsilon.$$

$\square$

Abusing terminology slightly, we still call elements of $\mathcal{Q}$ quorums, even though a probabilistic quorum system will not in general be a (strict) quorum system.

Several points are noteworthy with regards to this definition. First, a probabilistic quorum system is defined with respect to a specific guarantee level ϵ, and thus, there are different systems for different levels of consistency guarantee. Second, the definition contains an access strategy, which is chosen to achieve the desired level of guarantee. Other access strategies on the same set system may fail to achieve the required consistency level, as can be trivially demonstrated by a strategy that chooses each of two nonintersecting quorums with probability $1/2$. Thus, for a probabilistic quorum system to obtain the advertised probability of consistency when used in a protocol, the specified access strategy must be enforced. In addition, we have to adjust our definition of load accordingly.

Definition: If $\langle \mathcal{Q}, w, \epsilon \rangle$ is a probabilistic quorum system, then $L(\langle \mathcal{Q}, w, \epsilon \rangle) = L_w(\mathcal{Q})$. $\square$

Similarly, the definitions of fault tolerance and failure probability carry over as expected:

Definition: Let $\langle \mathcal{Q}, w, \epsilon \rangle$ be a probabilistic quorum system. Then the fault tolerance of $\langle \mathcal{Q}, w, \epsilon \rangle$ is $A(\langle \mathcal{Q}, w, \epsilon \rangle) = A(\mathcal{Q})$ and the failure probability of $\langle \mathcal{Q}, w, \epsilon \rangle$ is $F_p(\langle \mathcal{Q}, w, \epsilon \rangle) = F_p(\mathcal{Q})$. $\square$

3.1 A lower bound on load

We start by exploring the limits of the improvements over strict quorum systems that can be achieved by probabilistic quorum systems. Specifically, we show a lower bound on the load of probabilistic quorum systems. This lower bound is close to the lower bound for strict quorum systems, and thus indicates that we should not look to probabilistic quorums as a technique to circumvent the lower bound for strict ones.

In order to state and prove our lower bound, we make use of the following notation. Given a probabilistic quorum system $\langle \mathcal{Q}, w, \epsilon \rangle$, we denote

$$\mathcal{P} = \left\{ Q \in \mathcal{Q} : \sum_{Q' : Q' \cap Q \neq \emptyset} w(Q') \geq 1 - \sqrt{\epsilon} \right\}$$

Thus, $Q \notin \mathcal{P}$ when $\sum_{Q' : Q' \cap Q = \emptyset} w(Q') > \sqrt{\epsilon}$. Note that $\mathcal{P}$ is not empty because probabilistic consistency requirement implies that the total probability of choosing pairs Q, Q' such that $Q \cap Q' = \emptyset$ is at most ϵ. Thus,

$$\begin{aligned}\epsilon &\geq \sum_{Q} w(Q) \sum_{Q' : Q \cap Q' = \emptyset} w(Q') \\ &\geq \sum_{Q \notin \mathcal{P}} w(Q) \sum_{Q' : Q \cap Q' = \emptyset} w(Q') \\ &\geq \sum_{Q \notin \mathcal{P}} w(Q)\sqrt{\epsilon}\end{aligned}$$

or, equivalently, $\sqrt{\epsilon} \geq \sum_{Q_i \notin \mathcal{P}} w(Q_i)$. It then follows that $\sum_{Q_i \in \mathcal{P}} w(Q_i) \geq 1 - \sqrt{\epsilon}$. Finally, we let $c(\mathcal{P})$ denote the size of the smallest quorum in $\mathcal{P}$.

Theorem 3.1 *If $\langle \mathcal{Q}, w, \epsilon \rangle$ is a probabilistic quorum system, then $L_w(\mathcal{Q}) \geq (1-\sqrt{\epsilon}) \max\{\frac{1}{c(\mathcal{P})}, \frac{c(\mathcal{P})}{n}\}$. In particular, $L_w(\mathcal{Q}) \geq (1-\sqrt{\epsilon})\frac{1}{\sqrt{n}}$.*

Proof: Fix $Q \in \mathcal{P}$ such that $|Q| = c(\mathcal{P})$. Summing the loads induced by w on all the elements of Q we obtain:

$$\begin{aligned}\sum_{u \in Q} l_w(u) &= \sum_{u \in Q} \sum_{Q_i \ni u} w(Q_i) \\ &= \sum_{Q_i \in \mathcal{Q}} \sum_{u \in (Q \cap Q_i)} w(Q_i) \\ &\geq \sum_{Q_i : Q_i \cap Q \neq \emptyset} w(Q_i) \\ &\geq 1 - \sqrt{\epsilon}\end{aligned}$$

Therefore, some element in Q suffers a load of at least $\frac{1-\sqrt{\epsilon}}{c(\mathcal{P})}$, so $L_w(\mathcal{Q}) \geq \frac{1-\sqrt{\epsilon}}{c(\mathcal{P})}$.

To prove the second part, we sum the total load induced by w on all of the elements of the universe:

$$\begin{aligned}\sum_{u \in U} l_w(u) &= \sum_{u \in U} \sum_{Q_i \ni u} w(Q_i) \\ &= \sum_{Q_i \in \mathcal{Q}} \sum_{u \in Q_i} w(Q_i) \\ &= \sum_{Q_i \in \mathcal{Q}} |Q_i| w(Q_i) \\ &\geq \sum_{Q_i \in \mathcal{P}} c(\mathcal{P}) w(Q_i) \\ &\geq (1-\sqrt{\epsilon}) c(\mathcal{P})\end{aligned}$$

It follows that some element in U suffers a load of at least $\frac{(1-\sqrt{\epsilon})c(\mathcal{P})}{n}$, so $L_w(\mathcal{Q}) \geq \frac{(1-\sqrt{\epsilon})c(\mathcal{P})}{n}$. ■

3.2 A probabilistic quorum construction

We now demonstrate a probabilistic quorum system $\mathcal{Q}$ with $O(\frac{1}{\sqrt{n}})$ load and $\Omega(n)$ fault tolerance, that meets any required level of consistency guarantee for sufficiently large universes. The construction is very simple: Given a universe of n servers, the quorums are all the sets of size $\ell\sqrt{n}$, where the constant ℓ is chosen to make the probability that two random quorums intersect sufficiently high. 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 to any desired level the probability that the intersection of two quorums is empty. This is somewhat similar to the well-known birthday paradox [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 shall prove below).

Definition: Let U be a universe of size n. $W(n,\ell)$, $\ell \geq 1$, is the system $\langle \mathcal{Q}, w, \epsilon \rangle$ defined by $\mathcal{Q} = \{Q \subseteq U : |Q| = \ell\sqrt{n}\}$; $\forall Q \in \mathcal{Q}, w(Q) = \frac{1}{|\mathcal{Q}|}$; and $\epsilon = e^{-\ell^2}$. □

The probability of choosing at random two quorums that do not intersect can be made sufficiently small by appropriate choice of ℓ. We will need the following combinatorial fact.

Proposition 3.2 *For non-negative integers n, c, and i, $\frac{\binom{n-c}{c-i}}{\binom{n}{c}} \leq \left(\frac{c}{n}\right)^i \left(\frac{n-c}{n-i}\right)^{c-i}$.*

Lemma 3.3 *Let Q_1 and Q_2 be quorums of size $\ell\sqrt{n}$ each chosen uniformly at random. Then $\Pr[Q_1 \cap Q_2 = \emptyset] < e^{-\ell^2}$.*

Proof:

$$\begin{aligned}\Pr[Q_1 \cap \mathcal{Q}_2 = \emptyset] &= \frac{\binom{n-\ell\sqrt{n}}{\ell\sqrt{n}}}{\binom{n}{\ell\sqrt{n}}} \leq \left(\frac{n-\ell\sqrt{n}}{n}\right)^{\ell\sqrt{n}} \\ &\leq e^{-\frac{\ell\sqrt{n}}{n}\ell\sqrt{n}} = e^{-\ell^2}\end{aligned}$$

The first inequality follows from Proposition 3.2. ■

It is immediate from Lemma 3.3 that $W(n,\ell)$ is a probabilistic quorum system.

Theorem 3.4 *$W(n,\ell)$ is a probabilistic quorum system.*

Since every element is in $\binom{n-1}{\ell\sqrt{n}-1}$ quorums, the load $L(W(n,\ell))$ is $\frac{\ell}{\sqrt{n}} = O(\frac{1}{\sqrt{n}})$. Because only $\ell\sqrt{n}$ servers need be available in order for some quorum to be available, the fault tolerance $A(W(n,\ell)) = n - \ell\sqrt{n} + 1 = \Omega(n)$. The failure probability of $W(n,\ell)$ is 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

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

for all $p \leq 1 - \frac{\ell}{\sqrt{n}}$. Peleg and Wool showed that the failure probability of any quorum system whose fault tolerance is f is at least $e^{-\Omega(f)}$ [PW95b]. Therefore, for any $p \leq 1 - \frac{\ell}{\sqrt{n}}$, the failure probability of $W(n,l)$ is asymptotically optimal. Moreover, if $\frac{1}{2} \leq p \leq 1 - \frac{\ell}{\sqrt{n}}$, this probability is provably better than any strict quorum system.

Figure 1 demonstrates the dramatic improvement in failure probability achieved by $W(n,\ell)$ over **majority** and **singleton** (the strict quorum systems that are the two extremes in terms of failure probabilities [BG87, PW95b]). The figure plots the failure probability of **majority** and **singleton** against $W(n,\ell)$, for $n=100$ and $n=900$, respectively. The first construction plotted is $W(100,2)$, giving a probabilistic consistency guarantee of at least $1-e^{-4}\approx 0.982$, and the second one is $W(900,4)$, providing a guarantee level of $1-e^{-16}\approx 0.99999887$. As shown, $W(100,2)$ has marginal failure probability (<0.1) for server failure probabilities p up to 0.74, and $W(900,4)$ achieves similar failure probability for $p<0.83$.

4 Probabilistic dissemination quorum systems

To achieve consistency in a Byzantine environment, it is not sufficient that two quorums should have a nonempty intersection. This is because two quorums may intersect in a set containing *faulty* servers only, which may deviate arbitrarily and undetectably from their assigned protocol. Malkhi and Reiter [MR97] defined *(strict) dissemination quorum systems* that can be used to construct Byzantine-fault-tolerant replicated services that store certain types of data.

Similarly, 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 containing faulty servers only. We define *probabilistic dissemination quorum systems*, where the strict dissemination quorum system consistency requirement is replaced by a probabilistic one. We show that relaxing consistency can provide dramatic improvements in this setting, as well. As with crash failures, we are able to construct a probabilistic dissemination quorum system resilient to the Byzantine failure of any constant fraction of the system and with outstanding failure probability, for sufficiently large universes, whose load is $O(\frac{1}{\sqrt{n}})$. Indeed, the fault tolerance can be increased to *any* constant fraction of n for sufficiently large n while retaining asymptotically optimal load. For large n, this construction provides considerable advantage over strict dissemination quorum system constructions.

Definition: Let $\mathcal{Q}$ be a quorum system, w an access strategy for $\mathcal{Q}$, and ϵ a constant, $0<\epsilon<1$. The tuple $\langle\mathcal{Q},w,\epsilon\rangle$ is a *probabilistic t-dissemination quorum system* if for all $B\subseteq U$ such that $|B|=t$,

$$\sum_{Q,Q':Q\cap Q'\not\subseteq B} w(Q)w(Q') \;\geq\; 1-\epsilon.$$

$\square$

Probabilistic dissemination quorum systems can be used to implement Byzantine fault-tolerant services for the same types of data that strict ones can, using identical protocols to access them (see [MR97]). Note that, given a t-dissemination probabilistic quorum system $\mathcal{Q}$, t is the number of Byzantine failures that can be tolerated, while $A(\mathcal{Q})$ is the number of crash failures that can be tolerated. Since servers that fail arbitrarily can always opt to send no messages, $A(\mathcal{Q})\geq t$.

4.1 A probabilistic $\frac{n}{3}$-dissemination quorum construction

In this section we present a probabilistic t-dissemination quorum construction for $t=\frac{n}{3}$, the resiliency bound for strict dissemination quorum systems [MR97]. Our construction exhibits much better load and fault tolerance than strict constructions for this resiliency. We use a construction similar to $W(n,\ell)$, and show that for an appropriate choice of the parameter ℓ, this construction ensures consistency with any desired probability for sufficiently large universes.

Definition: Let U be a universe of size n. $W_{\frac{1}{3}}(n,\ell)$, $\ell\geq 1$, is the system $\langle\mathcal{Q},w,\epsilon\rangle$ defined by $\mathcal{Q}=\{Q\subseteq U:|Q|=\ell\sqrt{n}\}$; $\forall Q\in\mathcal{Q}, w(Q)=\frac{1}{|\mathcal{Q}|}$; and $\epsilon=2e^{-\frac{\ell^2}{6}}$. $\square$

Lemma 4.1 *Let U be a universe of size n, let B be a subset of U of size t where $t=\frac{n}{3}$, and let Q_1 and Q_2 be quorums of size $\ell\sqrt{n}$ each chosen uniformly at random. Then* $\Pr[(Q_1\cap Q_2)\subseteq B]\leq 2e^{-\frac{\ell^2}{6}}$.

Proof:

$$\begin{aligned}
&\Pr[Q_1\cap Q_2\subseteq B] && (1)\\
&\quad= \Pr[|Q_1\cap Q_2|=|Q_1\cap Q_2\cap B|]\\
&\quad= \sum_{i=0}^{\ell\sqrt{n}}\Pr[(|Q_1\cap Q_2|=i)\wedge(|Q_1\cap Q_2\cap B|=i)]\\
&\quad\leq \sum_{i=0}^{\ell\sqrt{n}}\frac{\binom{\ell\sqrt{n}}{i}\binom{n-\ell\sqrt{n}}{\ell\sqrt{n}-i}}{\binom{n}{\ell\sqrt{n}}}\left(\frac{1}{3}\right)^i && (2)\\
&\quad\leq \sum_{i=0}^{\ell\sqrt{n}}\binom{\ell\sqrt{n}}{i}\left(\frac{\ell\sqrt{n}}{n}\right)^i\left(\frac{n-\ell\sqrt{n}}{n-i}\right)^{\ell\sqrt{n}-i}\left(\frac{1}{3}\right)^i && (3)\\
&\quad\leq \sum_{i=0}^{\frac{\ell\sqrt{n}}{6}}\frac{(\ell^2)^i}{i!}\cdot e^{-\frac{(\ell\sqrt{n}-i)^2}{n-i}}\cdot 3^{-i}+\sum_{i=\frac{\ell\sqrt{n}}{6}+1}^{\ell\sqrt{n}}3^{-i} && (4)\\
&\quad\leq \sum_{i=0}^{\frac{\ell\sqrt{n}}{6}}\frac{\left(\frac{\ell^2}{3}\right)^i}{i!}\cdot e^{-\ell^2\left(\frac{5}{6}\right)^2}+\sum_{i=\frac{\ell\sqrt{n}}{6}+1}^{\ell\sqrt{n}}3^{-i} && (5)\\
&\quad\leq e^{-\ell^2\left(\frac{5}{6}\right)^2}\cdot e^{\frac{\ell^2}{3}}+3^{-\frac{\ell\sqrt{n}}{6}} && (6)\\
&\quad\leq 2e^{-\frac{\ell^2}{6}} && (7)
\end{aligned}$$

Let $c=\ell\sqrt{n}$. Then (2) holds because $\Pr[(|Q_1\cap Q_2\cap B|=i)\mid(|Q_1\cap Q_2|=i)]=\frac{\frac{1}{2}\binom{\frac{n}{3}}{i}\binom{n-i}{c-i}\binom{n-c}{c-i}}{\frac{1}{2}\binom{n}{c}\binom{c}{i}\binom{n-c}{c-i}}=\frac{\left(\frac{n}{3}\right)!(n-i)!}{\left(\frac{n}{3}-i\right)!n!}\leq\left(\frac{1}{3}\right)^i$; (3) is by Proposition 3.2; (4) is because for the first part of the sum: $\binom{c}{i}\left(\frac{c}{n}\right)^i\leq\frac{c^i}{i!}\frac{c^i}{n^i}=\frac{(l^2)^i}{i!}$ and $1+x\leq e^x$, for the second: $\binom{c}{i}\left(\frac{c}{n}\right)^i\left(\frac{n-c}{n-i}\right)^{c-i}\leq 1$; (5) holds since $e^{-\frac{(c-i)^2}{n-i}}\leq e^{-\frac{(c-\frac{c}{6})^2}{n}}=e^{-\ell^2\left(\frac{5}{6}\right)^2}$ for $i\leq\frac{c}{6}$; (6) is since $\sum_{i\geq 0}\frac{\frac{\ell^2}{3}}{i!}=e^{\frac{\ell^2}{3}}$; and (7) is because $e<3$ and $\ell\leq\sqrt{n}$. $\blacksquare$

Theorem 4.2 *$W_{\frac{1}{3}}(n,\ell)$ is a probabilistic $\frac{n}{3}$-dissemination quorum system.*

As with $W(n,\ell)$, the load $L(W_{\frac{1}{3}}(n,\ell))$ is $\frac{\ell}{\sqrt{n}}$, the fault tolerance is $A(W_{\frac{1}{3}}(n,\ell))=n-\ell\sqrt{n}+1$, and the failure probability is $F_p(W_{\frac{1}{3}}(n,\ell))\leq e^{-2n\gamma^2}$, where $\gamma=1-\frac{\ell}{\sqrt{n}}-p$, for $p<1-\frac{\ell}{\sqrt{n}}$.

4.2 A probabilistic αn-dissemination quorum construction

Surprisingly, the same technique can be used to overcome *any* fraction α of Byzantine failures. The construction in this section is essentially identical to $W_{\frac{1}{3}}(n,\ell)$, with the distinction that ϵ is chosen to depend on the fraction α of servers that

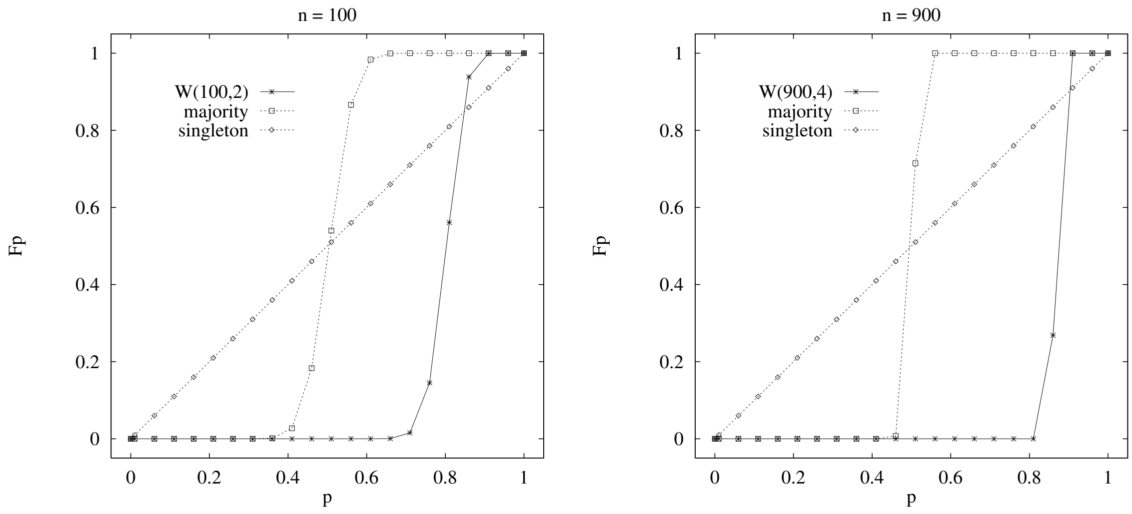

Figure 1: Comparison of failure probabilities for $W(n,\ell)$, majority, and singleton for $n = 100$ (with $\ell = 2$) and $n = 900$ (with $\ell = 4$).

may simultaneously fail. The choice of the appropriate parameter ℓ depends also on α. Since our construction works, with appropriate choice of parameters, with $t = \alpha n$ 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 $t = \frac{n}{3}$ limits the fault tolerance. We present the construction here for $\frac{1}{3} < \alpha < 1$, as the case $0 < \alpha \le \frac{1}{3}$ was already covered. (A similar result holds for $0 < \alpha < 1$, but ϵ is more complicated in this case).

Definition: Let U be a universe of size n. $W(n, \ell, \alpha)$ where $\ell \ge 1$ and $\frac{1}{3} < \alpha < 1$ is the system $\langle \mathcal{Q}, w, \epsilon \rangle$ defined by $\mathcal{Q} = \{Q \subseteq U : |Q| = \ell\sqrt{n}\}$; $\forall Q \in \mathcal{Q}, w(Q) = \frac{1}{|\mathcal{Q}|}$; and $\epsilon = 2\alpha^{\ell^2(\frac{1-\sqrt{\alpha}}{2})} \frac{1}{1-\alpha}$. □

An argument similar to that in Section 4.1 shows the following lemma holds.

Lemma 4.3 *Let U be a universe of n servers, let B be a subset of U of size t where $t = \alpha n$ for some $\frac{1}{3} < \alpha < 1$, and let Q_1 and Q_2 be quorums of size $\ell\sqrt{n}$ each chosen uniformly at random. Then* $\Pr[(Q_1 \cap \mathcal{Q}_2) \subseteq B] \le 2\alpha^{\ell^2(\frac{1-\sqrt{\alpha}}{2})} \frac{1}{1-\alpha}$.

Theorem 4.4 *$W(n, \ell, \alpha)$ is a probabilistic αn-dissemination quorum system.*

Here, again the load is $L(W(n, \ell, \alpha)) = \frac{\ell}{\sqrt{n}}$. Since we assume that αn servers may fail, we must have $n - \ell\sqrt{n} > \alpha n$, or equivalently, $\ell < \sqrt{n}(1 - \alpha)$.

Note that $\mathcal{Q}$ and w do not 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 consistency parameter ϵ that is achieved will also be unknown. Furthermore, note that the construction has the desirable property that actual probability of consistency will be better if fewer Byzantine faults actually occur.

5 Conclusions

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 showed a generic construction of probabilistic quorum systems that have optimal load but far exceed the resiliency 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 resiliency for this model as well.

An obvious drawback of the probabilistic approach is the chance of inconsistency allowed in any construction. We have shown how this probability can be limited to any desired level of guarantee, for appropriate universe sizes. Nevertheless, the probabilistic constructions are best suited for applications that can tolerate some (marginal and known) fraction of inconsistency, and where availability may be more important than utmost consistency. Moreover, our probabilistic construction may be easily combined with some strict quorum constructions, e.g., the set of majorities, to produce "hybrid" constructions with the following guarantee: Among operations performed only on strict quorums, consistency is provided absolutely, whereas all other operations provide the appropriate probabilistic guarantee.

Acknowledgments

We are thankful to Oded Goldreich and the anonymous referees for many helpful comments on an earlier version of this paper.

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.

[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.

[CP96] R. Cáceres and V. Padmanabhan. Fast and scalable handoffs for wireless internetworks. In *Proceedings of the 2nd ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom '96)*, pages 56–66, November 1996.

[BG87] D. Barbara and H. Garcia-Molina. The reliability of voting mechanisms. In *IEEE Transactions on Computers*, 36(10):1197–1208, October 1987.

[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.

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

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

[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.

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

[JP96] D. Johnson and C. Perkins. Route optimization in Mobile IP, *Internet Draft*, Internet Engineering Task Force, February 1996.

[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.

[Lam83] B. Lampson. Hints for computer system design. *Operating Systems Review*, 17(5):33–48, 1983.

[LS79] B. Lampson and R. Sproul. An open operating system for a single-user machine. *Operating Systems Review*, 13(5):98–105, 1979.

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

[MMR97] D. Malkhi, M. Merritt and O. Rodeh. Secure multicast in a WAN. *The International Conference on Distributed Computing Systems* (ICDCS), Baltimore, May 1997, to be published.

[MR97] D. Malkhi and M. Reiter. Byzantine quorum systems. In *Proceedings of the 29th ACM Symposium on Theory of Computing (STOC)*, May 1997. To appear.

[MW77] M. McQuillan and D. Walden. The ARPA network design decisions. In *Proceedings of Computer Networks 1*, pages 243–289, August 1977.

[NW94] M. Naor and A. Wool. The load, capacity, and availability of quorum systems. In *Proceedings of the 35th IEEE Symposium on Foundations of Computer Science*, pages 214–225, November 1994.

[PW95a] D. Peleg and A. Wool. Crumbling walls: A class of high availability quorum systems. In *Proceedings of the 14th ACM Symposium on Principles of Distributed Computing*, pages 120–129, August 1995.

[PW95b] D. Peleg and A. Wool. The availability of quorum systems. *Information and Computation* 123(2):210–233, 1995.

[Sch90] F. B. Schneider. Implementing fault-tolerant services using the state machine approach: A tutorial. *ACM Computing Surveys*, 22(4):299–319, December 1990.

[Smi81] J. Smith. A study of branch prediction strategies. In *Proceedings of the 8th Symposium on Computer Architecture*, pages 135–148, May 1981.

[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.