The Load and Availability of Byzantine Quorum Systems

Dahlia Malkhi
AT&T Labs – Research,
Florham Park, NJ 07932
dalia@research.att.com

Michael Reiter
AT&T Labs – Research,
Florham Park, NJ 07932
reiter@research.att.com

Avishai Wool
Bell Laboratories, Lucent Technologies,
Murray Hill, NJ 07974
yash@research.bell-labs.com

Abstract

Replicated services accessed via *quorums* enable each access to be performed at only a subset (quorum) of the servers, and achieve consistency across accesses by requiring any two quorums to intersect. Recently, b-masking quorum systems, whose intersections contain at least $2b+1$ servers, have been proposed to construct replicated services tolerant of b arbitrary (Byzantine) server failures. In this paper we consider a hybrid fault model allowing benign failures in addition to the Byzantine ones. We present four novel constructions for b-masking quorum systems in this model, each of which has optimal *load* (the probability of access of the busiest server) or optimal availability (probability of some quorum surviving failures). To show optimality we also prove lower bounds on the load and availability of any b-masking quorum system in this model.

1 Introduction

Quorum systems are well known tools for increasing the efficiency of replicated services, as well as their availability when servers may fail benignly. A quorum system is a set of subsets (quorums) of servers, every pair of which intersect. Quorum systems enable each client operation to be performed only at a quorum of the servers, while the intersection property makes it possible to preserve consistency among operations at the service.

Quorum systems work well for environments where servers may fail benignly. However, when servers may suffer arbitrary (Byzantine) failures, the intersection property does not suffice for maintaining consistency; two quorums may intersect in a subset containing *faulty* servers only, who may deviate arbitrarily and undetectably from their assigned protocol. Malkhi and Reiter thus introduced *masking quorums systems* [MR97], in which each pair of quorums intersects in sufficiently many servers to mask out the behavior of faulty servers. More precisely, a *b-masking quorum system* is one in which any two quorums intersect in $2b+1$ servers, which suffices to ensure consistency in the system if at most b servers suffer Byzantine failures.

In this paper we develop four new constructions for b-masking quorum systems. For the first time in this context, we distinguish between masking Byzantine faults and surviving a possibly larger number of benign faults. Our systems remain available in the face of any f crashes, where f may be significantly larger than b (such a system is called f-resilient). In addition, our constructions demonstrate optimality (ignoring constants) in two widely accepted measures of quorum systems, namely *load* and *crash probability*. The load ($\mathcal{L}$), a measure of best-case performance of the quorum system, is the probability with which the busiest server is accessed under the best possible strategy for accessing quorums. The crash probability (F_p) is the probability, assuming that each server crashes with probability p, that all quorums in the system will contain at least one crashed server (and thus will be unavailable). The crash probability is an even more refined measure of availability than f, as a good system will tolerate many failure configurations with more than f crashes. In proving optimality of our constructions, we prove new lower bounds for the load and crash probability of masking quorum systems.

The techniques for achieving our constructions are of interest in themselves. Two of the constructions are achieved using a *boosting* technique, which can transform any regular (i.e., benign fault-tolerant) quorum system into a masking quorum system of an appropriately larger system. Thus, it makes all known quorum constructions available for Byzantine environments (of appropriate sizes). In the analysis of one of our best systems we employ strong results from percolation theory.

The properties of our quorum systems are summarized in Table 1. Three of our systems—M-Grid, boostFPP, and M-Path—are the first systems to demonstrate optimal load for b-masking quorum systems. M-Path does so simultaneously with optimal crash probability for its resilience f, which the fourth (RT) also demonstrates.

The rest of this paper is structured as follows. We review related work and preliminary definitions in Sections 2 and 3, respectively. In Section 4 we prove bounds on the load and crash failure probability for b-masking quorum systems and introduce quorum composition. In Sections 5–7 we describe our new constructions. We conclude in Section 8.

2 Related work

Our work borrows from extensive prior work in benignly fault-tolerant quorum systems (e.g., [Gif79, Tho79, Mae85, GB85, Her86, BG87, ET89, AE91, CAA92, NW94, PW97b]). The notion of availability we use here (crash probability) is well known in reliability theory [BP75] and has been applied extensively in the analysis of quorum systems (cf. [BG87, PW95, PW97a] and the references therein). The load of a quorum system was first defined and analyzed in [NW94], which proved a lower bound of $\Omega(\frac{1}{\sqrt{n}})$ on the load of any quorum system (and, a fortiori, any masking quorum system) over n servers. In proving load-optimality of our constructions, we generalize this lower bound to $\Omega(\sqrt{\frac{b}{n}})$ for b-masking quorum systems.

Grids, which form the basis for our M-Grid construction, were proposed in [Mae85, CAA92, KRS93, MR97]. The technique of quorum composition, which we use in our RT and boostFPP constructions, has been studied in [MP92, NM92, Nei92] under various names such as "coterie join" and "re-

System	$b <$	f	$\mathcal{L}$	F_p
Threshold [MR97]	$n/4$	$O(n-b)$	$1/2+O(b/n)$	$\exp(-\Omega(f))$ *
Grid [MR97]	$\sqrt{n}/3$	$O(\sqrt{n}-b)$	$O(b/\sqrt{n})$	$\underset{n\to\infty}{\longrightarrow} 1$
M-Grid	$\sqrt{n}/2$	$O(\sqrt{n}-\sqrt{b})$	$O(\sqrt{b/n})$ †	$\underset{n\to\infty}{\longrightarrow} 1$
RT‡	$O(n^{\alpha})$	$O(b)$	$n^{-\beta}$	$\exp(-\Omega(f))$ *
boostFPP	$n/4$	$O(\sqrt{bn})$	$O(\sqrt{b/n})$ †	$\exp(-\Omega(b-\log(n/b)))$
M-Path	$(1-o(1))\sqrt{n}$	$O(\sqrt{n}-\sqrt{b})$	$O(\sqrt{b/n})$ †	$\exp(-\Omega(f))$ *

† Optimal for b-masking systems.

* Optimal for f-resilient systems.

‡ E.g., $\alpha = 0.5$ and $\beta = 0.21$ for RT(4, 3).

Table 1: Constructions in this paper (n = number of servers).

cursive majority". Our M-Path construction generalizes the system of [WB92], coupled with the analysis of the Paths construction of [NW94], and the recent system of [Baz96].

Several constructions of masking quorum systems were given in [MR97] for a variety of failure models. For the model we consider here—i.e., any b servers may experience Byzantine failures—that work gave two constructions. The properties of these constructions are included in the preceding table for comparison with our results.

Hybrid failure models have been considered in other works (e.g., [LR93, LR94, RB94]).

3 Preliminaries

We assume a *universe* U of servers, $|U| = n$, over which our quorum systems will be constructed. Servers that obey their specifications are *correct*. A *faulty* server, however, may deviate from its specification arbitrarily. We assume that up to b servers may fail arbitrarily and that $4b < n$, since this is necessary for a b-masking quorum system to exist [MR97]. Beginning in Section 3.2.2, we will also distinguish benign (crash) failures as a particular failure of interest, and in general there may be more than b such failures.

3.1 Quorum systems

Definition 3.1 *A* quorum system $\mathcal{Q} \subseteq 2^U$ *is a collection of subsets of* U*, each pair of which intersect. Each* $Q \in \mathcal{Q}$ *is called a* quorum.

We use the following notation. The cardinality of the smallest quorum in $\mathcal{Q}$ is denoted by $c(\mathcal{Q}) = \min\{|Q| : Q \in \mathcal{Q}\}$. The size of the smallest intersection between any two quorums is denoted by $\mathcal{IS}(\mathcal{Q}) = \min\{|Q \cap R| : Q, R \in \mathcal{Q}\}$. The degree of an element $i \in U$ in a quorum system $\mathcal{Q}$ is the number of quorums that contain i: $\deg(i) = |\{Q \in \mathcal{Q} : i \in Q\}|$.

Definition 3.2 *A quorum system* $\mathcal{Q}$ *is* (s,d)-fair *if* $|Q| = s$ *for all* $Q \in \mathcal{Q}$ *and* $\deg(i) = d$ *for all* $i \in U$*.* $\mathcal{Q}$ *is called* fair *if it is* (s,d)*-fair for some* s *and* d*.*

Definition 3.3 *A set* T *is a* transversal *of a quorum system* $\mathcal{Q}$ *if* $T \cap Q \neq \varnothing$ *for every* $Q \in \mathcal{Q}$*. The cardinality of the smallest transversal is denoted by* $\mathcal{MT}(\mathcal{Q}) = \min\{|T| :$ T is a transversal of $\mathcal{Q}\}$.

Regular quorum systems, with $\mathcal{IS}(\mathcal{Q}) = 1$, are insufficient to guarantee consistency in case of Byzantine failures. Malkhi and Reiter [MR97] defined several varieties of quorum systems for Byzantine environments, which are suitable for different types of services. In this paper we focus on *masking quorum systems.*

Definition 3.4 [MR97] *The* resiliency *f of a quorum system* $\mathcal{Q}$ *is the largest* k *such that for every set* $K \subseteq U$, $|K| = k$, *there exists* $Q \in \mathcal{Q}$ *such that* $K \cap Q = \varnothing$.

Remark: The resiliency of any quorum system $\mathcal{Q}$ is $f = \mathcal{MT}(\mathcal{Q}) - 1$.

Definition 3.5 [MR97] *A quorum system* $\mathcal{Q}$ *is a* b-masking quorum system *if it is resilient to* $f \geq b$ *failures, and obeys the following* consistency *requirement:*

$$\forall Q_1, Q_2 \in \mathcal{Q} : |Q_1 \cap Q_2| \geq 2b+1.$$

Remark: Informally, if we view the service as a shared variable which is updated and read by the clients, then the resiliency requirement of Definition 3.4 ensures that no set of $b \leq f$ faulty servers will be able to block update operations (e.g., by causing every update transaction to abort). The consistency requirement of Definition 3.5 ensures that read operations can mask out any faulty behavior of up to b servers, provided that the algorithm given in [MR97] is used.

Lemma 3.6 *Let* $\mathcal{Q}$ *be a quorum system. Then* $\mathcal{Q}$ *is* b*-masking if both the following conditions hold:*

1. $\mathcal{MT}(\mathcal{Q}) \geq b+1$,
2. $\mathcal{IS}(\mathcal{Q}) \geq 2b+1$.

Proof: Assume that $\mathcal{MT}(\mathcal{Q}) \geq b+1$. To see that $\mathcal{Q}$ is resilient to b failures, note that if there exists some K such that $K \cap Q \neq \varnothing$ for all $Q \in \mathcal{Q}$, then K is a transversal. By the minimality we have $|K| \geq b+1$, and we are done. Condition 2 is obvious. ☐

Corollary 3.7 *Let* $\mathcal{Q}$ *be a quorum system, and let* $b = \min\{\mathcal{MT}(\mathcal{Q}) - 1, \frac{\mathcal{IS}(\mathcal{Q})-1}{2}\}$*. Then* $\mathcal{Q}$ *is* b*-masking.*

3.2 Measures

The goal of using quorum systems is to increase the availability of replicated services and decrease their access costs. A natural question is how well any particular quorum system achieves these goals, and moreover, how well it compares with other quorum systems. Several measures will be of interest to us.

3.2.1 Load

A measure of the inherent performance of a quorum system is its *load*. Naor and Wool define the load of a quorum system as the frequency of accessing the busiest server using the best possible strategy [NW94]. More precisely, given a quorum system $\mathcal{Q}$, an *access strategy* w is a probability distribution on the elements of $\mathcal{Q}$; i.e., $\sum_{Q\in\mathcal{Q}} w(Q) = 1$. The value $w(Q) \geq 0$ is the frequency of choosing quorum Q when the service is accessed. The load is then defined as follows:

Definition 3.8 *Let a strategy w be given for a quorum 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 a quorum system $\mathcal{Q}$ is $\mathcal{L}_w(\mathcal{Q}) = \max_{u\in U}\{l_w(u)\}$. The* system load *on a quorum system $\mathcal{Q}$ is $\mathcal{L}(\mathcal{Q}) = \min_w\{\mathcal{L}_w(\mathcal{Q})\}$, where the minimum is taken over all strategies.*

We reiterate that the 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 the load is a property of a quorum system, and not of the protocol using it.

Recall that $c(\mathcal{Q})$ denotes the cardinality of the smallest quorum in $\mathcal{Q}$. The following result will be useful to us in the sequel (recall Definition 3.2).

Proposition 3.9 [NW94] *Let $\mathcal{Q}$ be a fair quorum system. Then $\mathcal{L}(\mathcal{Q}) = c(\mathcal{Q})/n$.*

3.2.2 Availability

By definition a b-masking quorum system can mask up to b arbitrary (Byzantine) failures. However, such a system may be resilient to more *benign* failures. By benign failures we mean any failures that render a server unresponsive, which we refer to as *crashes* to distinguish them from Byzantine failures.

The resiliency f of a quorum system provides one measure of how many crash failures a quorum system is *guaranteed* to survive, and indeed this measure has been used in the past to differentiate among quorum systems [BG86]. However, it is possible that an f-resilient quorum system, though vulnerable to a few failure configurations of $f+1$ failures, can survive many configurations of more than f failures. One way to measure this property of a quorum system is to assume that each server crashes independently with probability p and then to determine the probability F_p that some quorum survives with no faulty members. This is known as *crash probability* and is formally defined as follows:

Definition 3.10 *Assume that each server in the system crashes independently with probability p. For every quorum $Q \in \mathcal{Q}$ let $\mathcal{E}_Q$ be the event that Q is* hit, *i.e., at least one element $i \in Q$ has crashed. Let $crash(\mathcal{Q})$ be the event that all the quorums $Q \in \mathcal{Q}$ were hit, i.e., $crash(\mathcal{Q}) = \bigwedge_{Q\in\mathcal{Q}} \mathcal{E}_Q$. Then the system crash probability is $F_p(\mathcal{Q}) = \mathbb{P}(crash(\mathcal{Q}))$.*

We would like F_p to be as small as possible. A desirable asymptotic behavior of F_p is that $F_p \to 0$ when $n \to \infty$ for *all* $p < 1/2$, and such an F_p is called Condorcet (after the Condorcet Jury Theorem [Con]).

4 Building Blocks

In this section, we prove several theorems which will be our basic tools in the sequel. First we prove lower bounds on the load and availability of b-masking quorum systems, against which we measure all our new constructions. Then we prove the properties of a quorum composition technique, which we later use extensively.

4.1 The load and availability of masking quorum systems

We begin by establishing a lower bound on the load of b-masking quorum systems, thus tightening the lower bound on general quorum systems [NW94] as presented in [MR97].

Theorem 4.1 *Let $\mathcal{Q}$ be a b-masking quorum system. Then $\mathcal{L}(\mathcal{Q}) \geq max\{\frac{2b+1}{c(\mathcal{Q})}, \frac{c(\mathcal{Q})}{n}\}$.*

Proof: Let w be any strategy for the quorum system $\mathcal{Q}$, and fix $Q_1 \in \mathcal{Q}$ such that $|Q_1| = c(\mathcal{Q})$. Summing the loads induced by w on all the elements of Q_1, and using the fact that any two quorums have at least $2b+1$ elements in common, we obtain:

$$\begin{aligned}\sum_{u\in Q_1} l_w(u) &= \sum_{u\in Q_1}\sum_{Q_i\ni u} w(Q_i) = \sum_{Q_i}\sum_{u\in(Q_1\cap Q_i)} w(Q_i)\\ &\geq \sum_{Q_i}(2b+1)w(Q_i) = 2b+1.\end{aligned}$$

Therefore, there exists some element in Q_1 that suffers a load of at least $\frac{2b+1}{|Q_1|}$.

Similarly, summing the total load induced by w on all of the elements of the universe, and using the minimality of $c(\mathcal{Q})$, we get:

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

Therefore, there exists some element in U that suffers a load of at least $\frac{c(\mathcal{Q})}{n}$. □

Corollary 4.2 *Let $\mathcal{Q}$ be a b-masking quorum system. Then $\mathcal{L}(\mathcal{Q}) \geq \sqrt{\frac{2b+1}{n}}$, and equality holds if $c(\mathcal{Q}) = \sqrt{(2b+1)n}$.* □

Remark: Corollary 4.2 shows that the threshold construction of [MR97] in fact has optimal load when $b = \Omega(n)$. E.g., when $b \approx n/4$ the obtained load is ≈ 0.75, but for such systems we can only hope for a constant load of $\approx 1/\sqrt{2} = 0.707$. However the load of the threshold construction is always $\geq 1/2$, which is far from optimal for smaller values of b.

On the other hand, the grid-based construction of [MR97] does not have optimal load. It has quorums of size $O(b\sqrt{n})$ and load of roughly $2b/\sqrt{n}$. In the sequel we show systems which significantly improve this: some of our new constructions have quorums of size $O(\sqrt{bn})$ and optimal load.

Our next propositions show lower bounds on the crash probability F_p in terms of $\mathcal{MT}(\mathcal{Q})$ and b.

Proposition 4.3 *Let $\mathcal{Q}$ be a quorum system. Then $F_p(\mathcal{Q}) \geq p^{\mathcal{MT}(\mathcal{Q})} = p^{f+1}$ for any $p \in [0,1]$.*

Proof: Consider a minimal transversal T with $|T| = \mathcal{MT}(\mathcal{Q})$. If all the elements of T crash then every quorum contains a crashed element, so $F_p(\mathcal{Q}) \geq p^{\mathcal{MT}(\mathcal{Q})}$. □

Proposition 4.4 *Let $\mathcal{Q}$ be a b-masking quorum system. Then $F_p(\mathcal{Q}) \geq p^{c(\mathcal{Q})-2b}$ for any $p \in [0,1]$.*

Proof: Let $Q \in \mathcal{Q}$ be a minimal quorum with $|Q| = c(\mathcal{Q})$, and consider $Z \subset Q$, $|Z| = 2b$. Since $\mathcal{Q}$ is b-masking then $|R \cap Q| \geq 2b+1$ for any $R \in \mathcal{Q}$, and so $|(Q \setminus Z) \cap R| \geq 1$ and $Q \setminus Z$ is a transversal. Therefore $\mathcal{MT}(\mathcal{Q}) \leq c(\mathcal{Q}) - 2b$, which we plug into Proposition 4.3. □

The next proposition is less general than Proposition 4.4, however it is applicable for most of our constructions and it gives a much tighter bound.

Proposition 4.5 *Let $\mathcal{Q}$ be a b-masking quorum system such that $\mathcal{MT}(\mathcal{Q}) \leq (\mathcal{IS}(\mathcal{Q})+1)/2$. Then $F_p(\mathcal{Q}) \geq p^{b+1}$ for any $p \in [0,1]$.*

Proof: If $\mathcal{MT}(\mathcal{Q}) \leq (\mathcal{IS}(\mathcal{Q})+1)/2$ then from Corollary 3.7 we have that $b+1 = \mathcal{MT}(\mathcal{Q})$, which again we plug into Proposition 4.3. □

4.2 Quorum system composition

Quorum system composition is a well known technique for building new systems out of existing components. We compose a quorum system $\mathcal{S}$ over another system $\mathcal{R}$ by replacing each element of $\mathcal{S}$ with a distinct copy of $\mathcal{R}$. In other words, when element i is used in a quorum $S \in \mathcal{S}$ we replace it with a complete quorum from the i'th copy of $\mathcal{R}$. Using the terminology of reliability theory, the system $\mathcal{S} \circ \mathcal{R}$ has a *modular decomposition* where each module is a copy of $\mathcal{R}$. Formally:

Definition 4.6 *Let $\mathcal{S}$ and $\mathcal{R}$ be two quorum systems, over universes of sizes n_S and n_R, respectively. Let $\mathcal{R}_1, \ldots, \mathcal{R}_{n_S}$ be n_S copies of $\mathcal{R}$ over disjoint universes. Then the* composition *of $\mathcal{S}$ over $\mathcal{R}$ is*

$$\mathcal{S} \circ \mathcal{R} = \left\{ \bigcup R_i : S \in \mathcal{S}, R_i \in \mathcal{R}_i \text{ for all } i \in S \right\}.$$

The next theorem summarizes the properties of quorum composition.

Theorem 4.7 *Let $\mathcal{S}$ and $\mathcal{R}$ be two quorum systems, and let $\mathcal{Q} = \mathcal{S} \circ \mathcal{R}$. Then*

- *The universe size is $n_Q = n_S n_R$.*
- *The minimal quorum size is $c(\mathcal{Q}) = c(\mathcal{S})c(\mathcal{R})$.*
- *The minimal intersection size is $\mathcal{IS}(\mathcal{Q}) = \mathcal{IS}(\mathcal{S})\mathcal{IS}(\mathcal{R})$.*
- *The minimal transversal size is* $$\mathcal{MT}(\mathcal{Q}) = \mathcal{MT}(\mathcal{S})\mathcal{MT}(\mathcal{R}).$$
- *The load is $\mathcal{L}(\mathcal{Q}) = \mathcal{L}(\mathcal{S})\mathcal{L}(\mathcal{R})$.*
- *Denote the crash probability functions of $\mathcal{S}$ and $\mathcal{R}$ by $s(p) = F_p(\mathcal{S})$ and $r(p) = F_p(\mathcal{R})$. Then $F_p(\mathcal{Q}) = s(r(p))$.*

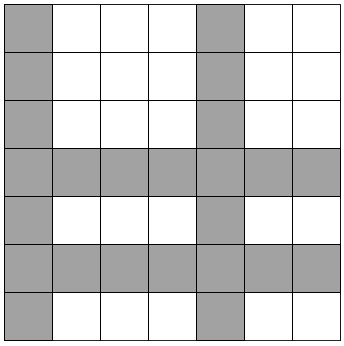

Figure 1: The multi-grid construction, $n = 7 \times 7, b = 3$, with one quorum shaded.

Proof Sketch: The behavior of the combinatorial parameters $c(\mathcal{Q})$, $\mathcal{IS}(\mathcal{Q})$ and $\mathcal{MT}(\mathcal{Q})$ is obvious. The strategy which achieves the optimal load for $\mathcal{Q}$ is the following: pick a quorum $S \in \mathcal{S}$ using the optimal strategy for $\mathcal{S}$. Then for each element i in this S, pick a quorum $R_i \in \mathcal{R}_i$ using the optimal strategy for (the i'th copy of) $\mathcal{R}$. It is easy to see that the composite strategy is indeed optimal. The behavior of $F_p(\mathcal{Q})$ is standard in reliability theory (cf. [BP75]). □

The multiplicative behavior of the combinatorial parameters in composing quorum systems provides a powerful tool for "boosting" existing constructions into larger systems with possibly improved characteristics. Below, we use quorum composition in two cases, and demonstrate that this technique yields improved constructions over their basic building blocks, for appropriately larger system sizes. In particular, in Section 6 we show a composition that allows us to transform any regular quorum construction into a (larger) b-masking quorum system.

5 Simple systems

In this section we show two types of constructions, the multi-grid (denoted M-Grid) and the recursive threshold (RT). These systems significantly improve upon the original constructions of [MR97], however both are still suboptimal in some parameter: M-Grid has optimal load but can mask only up to $b = O(\sqrt{n})$ failures and has poor crash probability; and RT can mask up to $b = O(n)$ failures and has near optimal crash probability, but has suboptimal load.

In Sections 6 and 7 we present systems which are superior to the M-Grid and RT. Nonetheless, we feel that the simplicity of the M-Grid and RT systems, and the fact that they are suitable for very small universe sizes, are what makes them appealing.

5.1 A multi-grid

We begin with the M-Grid system, which achieves an optimal load among b-masking quorum systems, where $b \leq (\sqrt{n}-1)/2$. The idea of the construction is as follows. Arrange the elements in a $\sqrt{n} \times \sqrt{n}$ grid. A quorum in a multi-grid consists of any choice of $\sqrt{b+1}$ rows and $\sqrt{b+1}$ columns, as shown in Figure 1. Formally, denote the rows and columns of the grid by R_i and C_i, respectively, where $1 \leq i \leq \sqrt{n}$. Then, the quorum system is

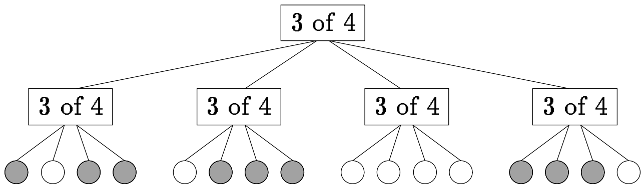

Figure 2: An RT(4,3) system of depth $h = 2$, with one quorum shaded.

$$\text{M-Grid}(b) = \left\{ \bigcup_{j\in J} C_j \cup \bigcup_{i\in I} R_i : J, I \subseteq \{1\dots\sqrt{n}\}, |J| = |I| = \sqrt{b+1} \right\}.$$

Proposition 5.1 *The multi-grid* M-Grid(b) *is a b-masking quorum system for $b \le (\sqrt{n}-1)/2$.*

Proof: Consider two quorums $R, S \in$ M-Grid(b). If they have either a row or a column in common, then $|R \cap S| \ge \sqrt{n} \ge 2b+1$ and we are done. Otherwise the intersection of S's columns with R's rows is disjoint from the intersection of R's columns with S's rows, so $|R\cap S| \ge 2\sqrt{b+1}\sqrt{b+1} > 2b+1$. Therefore consistency holds.

Resiliency holds since $f = \mathcal{MT}(\text{M-Grid}(b)) - 1 = \sqrt{n} - \sqrt{b+1} \ge b$. Therefore $\mathcal{MT}(\text{M-Grid}(b)) \ge b+1$, and Lemma 3.6 finishes the proof. $\Box$

Proposition 5.2 $\mathcal{L}(\text{M-Grid}(b)) \approx 2\sqrt{\frac{b+1}{n}}$.

Proof: Since M-Grid(b) is fair we can use Proposition 3.9 to get $\mathcal{L}(\text{M-Grid}(b)) = c(\text{M-Grid}(b))/n$. $\Box$

Remark: The load of M-Grid(b) is within a factor of $\sqrt{2}$ from the optimal load which can be achieved for $b \approx \sqrt{n}/2$.

A disadvantage of the M-Grid system is its poor asymptotic crash probability. If crashes occur with some constant probability p then any configuration of crashes with at least one crash per row disables the system. Therefore, as shown by [KC91, Woo96],

$$F_p(\text{M-Grid}) \ge (1-(1-p)^{\sqrt{n}})^{\sqrt{n}} \underset{n\to\infty}{\longrightarrow} 1.$$

5.2 Recursive threshold systems

A recursive threshold system RT(k,ℓ) of depth h is built by taking a simple building block, which is an ℓ-of-k threshold system (with $k > \ell > k/2$), and recursively composing it over itself to depth h. In the sequel, we often omit the depth parameter h when it has no effect on the discussion. The RT systems generalize the recursive majority constructions of [MP92], the HQS system of [Kum91] is an RT$(3,2)$ system, and in fact the threshold system of [MR97] can be viewed as a trivial RT$(4b+1, 3b+1)$ system with depth $h=1$. As an example throughout this section we will use the RT$(4,3)$ system, depicted in Figure 2.

Proposition 5.3 *An* RT(k,ℓ) *system of depth h is a fair quorum system, with $n = k^h$ elements, quorums of size $c(\text{RT}(k,\ell)) = \ell^h$, intersection size of $\mathcal{IS}(\text{RT}(k,\ell)) = (2\ell-k)^h$, and minimal transversals of size $\mathcal{MT}(\text{RT}(k,\ell)) = (k-\ell+1)^h$.*

Proof: The basic ℓ-of-k system is symmetric (and therefore fair), with $c(\ell\text{-of-}k) = \ell$, $\mathcal{MT}(\ell\text{-of-}k) = k-\ell+1$, and $\mathcal{IS}(\ell\text{-of-}k) = 2\ell-k$. The combinatorial parameters are computed by activating Theorem 4.7 h times, and the composition preserves the fairness. $\Box$

Plugging this into Corollary 3.7 we obtain

Corollary 5.4 *An* RT(k,ℓ) *system (of any depth) is a b-masking quorum system for*

$$b = \min\{(n^{\log_k(2\ell-k)}-1)/2, n^{\log_k(k-\ell+1)}-1\}. \qquad \Box$$

In the 3-of-4 example we have $\mathcal{IS}(\text{3-of-4}) = \mathcal{MT}(\text{3-of-4}) = 2$ and $c(\text{3-of-4}) = 3$. Therefore for the whole system (to depth $\log_4 n$) we get $c(\text{RT}(4,3)) = n^{\log_4 3} = n^{0.79}$, with $\mathcal{IS}(\text{RT}(4,3)) = \mathcal{MT}(\text{RT}(4,3)) = \sqrt{n}$ and thus $b = (\sqrt{n}-1)/2$. Note that the basic 3-of-4 system is not even 1-masking since intersections of size 2 are too small, however already from $h=2$ (i.e., $n=16$) we obtain a masking system.

Proposition 5.5 *The load* $\mathcal{L}(\text{RT}(k,\ell)) = n^{-(1-\log_k \ell)}$.

Proof: Since RT(k,ℓ) is fair we can use Proposition 3.9 to get $\mathcal{L}(\text{RT}(k,\ell)) = c(\text{RT}(k,\ell))/n$. $\Box$

Remark: In general the load is suboptimal for this construction. For instance, in the RT$(4,3)$ system we obtain $\mathcal{L}(\text{RT}(4,3)) = n^{-0.21}$. However for $b = (\sqrt{n}-1)/2$ we could hope for a load of $\sqrt{(2b+1)/n} = n^{-0.25}$.

Proposition 5.6 *There exists a unique critical probability $0 < p_c < 1/2$ for which*

$$\lim_{h\to\infty} F_p(\text{RT}(k,\ell) \text{ of depth } h) = \begin{cases} 0, & p < p_c, \\ 1, & p > p_c. \end{cases}$$

Proof: Let $g(p)$ be the crash probability function of the ℓ-of-k system and let $F(h) = F_p(\text{RT}(k,\ell)$ of depth $h)$ denote the crash probability for the RT(k,ℓ) system of depth h. Then $F(h)$ obeys the recurrence

$$F(h) = \begin{cases} g(F(h-1)), & h \ge 1, \\ p, & h = 0. \end{cases} \tag{1}$$

Now $g(p)$ is a reliability function, and therefore it is "S-shaped" (see [BP75]). This implies that there exists a unique critical probability $0 < p_c < 1$ for which $g(p_c) = p_c$, such that $g(p) < p$ when $p < p_c$ and $g(p) > p$ when $p > p_c$ (and [PW95] shows that for quorum systems such as RT in fact $p_c < 1/2$). Therefore if $p < p_c$ then repeated applications of recurrence (1) would decrease $F(h)$ arbitrarily close to 0, and when $p > p_c$ the limit is 1. $\Box$

Proposition 5.7 *If $p < 1/\binom{k}{\ell-1}$ and $\ell < k$ then $F_p(\text{RT}(k,\ell)) < \exp(-\Omega(n^{\log_k(k-\ell+1)}))$, which is optimal for systems with resiliency $f = n^{\log_k(k-\ell+1)}$.*

Proof: Let $g(p)$ and $F(h)$ be as in the proof of Proposition 5.6. Any configuration of at least $k-\ell+1$ crashes disables the ℓ-of-k system, so

$$g(p) = \sum_{j=k-\ell+1}^{k} \binom{k}{j} p^j (1-p)^{k-j}.$$

By Lemma A.2 (see Appendix) we have that

$$g(p) \le \binom{k}{\ell-1} p^{k-\ell+1}.$$

Plugging this into (1) gives that

$$\begin{aligned} F(h) &\leq \binom{k}{\ell-1}^{1+(k-\ell+1)+\cdots+(k-\ell+1)^{h-1}} p^{(k-\ell+1)^h} \\ &< \left[\binom{k}{\ell-1}p\right]^{(k-\ell+1)^h}. \end{aligned}$$

If $p < 1/\binom{k}{\ell-1}$ then the last expression decays to zero with h, so $F_p(\mathrm{RT}(k,\ell)) < \exp(-\Omega(n^{\log_k(k-\ell+1)}))$.

The lower bound of Proposition 4.3 shows that

$$F_p(\mathrm{RT}(k,\ell)) \geq p^{n^{\log_k(k-\ell+1)}},$$

so our analysis is tight. □

For the RT(4, 3) system a direct calculation shows that $g(p) = 6p^2 - 8p^3 + 3p^4$ and $p_c = 0.2324$. Therefore Proposition 5.6 guarantees that when the element crash probability is in the range $p < 0.2324$ then $F_p \to 0$ when $n \to \infty$. Furthermore, when $p < 1/6$ then Proposition 5.7 shows that the decay is rapid, with $F_p(\mathrm{RT}(4,3)) < (6p)^{\sqrt{n}}$, which is optimal.

6 Boosted finite projective planes

In this section we introduce a family of b-masking quorum systems, the *boosted finite projective planes*, which we denote by boostFPP. A boostFPP system is a composition of a finite projective plane (FPP) over a threshold system (Thresh).

The first component of a boostFPP system is a finite projective plane of order q (a good reference on FPP's is [Hal86]). It is known that FPP's exist for $q = p^r$ when p is prime. Such an FPP has $n_F = q^2 + q + 1$ elements, and quorums of size $c(\mathrm{FPP}) = q + 1$. This is a regular quorum system, i.e., it has intersections of size $\mathcal{IS}(\mathrm{FPP}) = 1$. The minimal transversals of an FPP are of size $\mathcal{MT}(\mathrm{FPP}) = q+1$ (in fact the only transversals of this size are the quorums themselves). The load of FPP was analyzed in [NW94] and shown to be $\mathcal{L}(\mathrm{FPP}) = \frac{q+1}{n_F} \approx 1/\sqrt{n_F}$, which is optimal for regular quorum systems.

The second component of a boostFPP is a Thresh system, with $n_T = 4b+1$ elements and a threshold of $3b+1$. This is a b-masking quorum system in itself, with $\mathcal{IS}(\mathrm{Thresh}) = 2b+1$, $\mathcal{MT}(\mathrm{Thresh}) = b + 1$ and a load of $\mathcal{L}(\mathrm{Thresh}) \approx 3/4$.

Proposition 6.1 *Let* $\mathrm{boostFPP}(q,b) = \mathrm{FPP}(q) \circ \mathrm{Thresh}(3b + 1 \text{ of } 4b + 1)$. *Then the composed system has* $n = (4b + 1)(q^2 + q + 1)$ *elements, with quorums of size* $c(\mathrm{boostFPP}(q,b)) = (3b + 1)(q + 1)$, *intersections of size* $\mathcal{IS}(\mathrm{boostFPP}(q,b)) = 2b + 1$ *and minimal transversals of size* $\mathcal{MT}(\mathrm{boostFPP}(q,b)) = (b + 1)(q + 1)$. *Therefore* boostFPP$(q,b)$ *is a b-masking quorum system.*

Proof: We obtain the combinatorial parameters by plugging the values of the component systems into Theorem 4.7. By Corollary 3.7 we have that the system can mask $\min\{(b + 1)(q + 1) - 1, b\} = b$ failures. □

Proposition 6.2 $\mathcal{L}(\mathrm{boostFPP}(q,b)) \approx \frac{3}{4q}$, *which is optimal for b-masking quorum systems with $n \approx 4bq^2$ elements.*

Proof: boostFPP(q,b) is a fair quorum system since both its components are fair, so by Proposition 3.9 we have

$$\begin{aligned} \mathcal{L}(\mathrm{boostFPP}(q,b)) &= \frac{c(\mathrm{boostFPP}(q,b))}{n} \\ &= \frac{(3b+1)(q+1)}{(4b+1)(q^2+q+1)} \approx \frac{3}{4q}. \end{aligned}$$

On the other hand, for b-masking systems with $n \approx 4bq^2$ elements the lower bound of Theorem 4.1 gives

$$\mathcal{L}(\mathrm{boostFPP}(q,b)) \geq \sqrt{\frac{2b}{n}} \approx \frac{1}{\sqrt{2}q}. \qquad \Box$$

Note that the optimality of the load holds for *any* choice of q and b. Therefore when the number of servers (or elements) increases, the boostFPP(q,b) system can scale up using different policies while maintaining load optimality. There are two extremal policies:

1. Fix q and increase b; then the system can mask more failures when new servers are added, however the load on the servers does not decrease.
2. Fix b and increase q; then the load decreases when new servers are added, but the number of failures that the system can mask remains unchanged.

It is important to note that systems of arbitrarily high resiliency can be constructed using the first policy since b can be chosen independently of q. In particular, we can choose $b = q^a$ for any $a > 0$. Then the resulting system has $n \approx 4bq^2 = 4b^{\frac{a+2}{a}}$, and so $b \approx \left(\frac{n}{4}\right)^{\frac{a}{a+2}}$, thus asymptotically approaching the resiliency upper bound of $\frac{n}{4}$.

Finally we analyze the crash probability of boostFPP. The following proposition shows that boostFPP has good availability as long as $p < 1/4$.

Proposition 6.3 *If* $p < 1/4$ *then* $F_p(\mathrm{boostFPP}(q,b)) \leq \exp(-\Omega(b - \log q))$.

Proof: We start by estimating $F_p(\mathrm{Thresh})$. Let *#crashed* denote the number of crashed elements in a universe of size $4b + 1$. Let $\gamma = \frac{b+1}{4b+1} - p$, thus $0 < \gamma < 1$ when $p < 1/4$. Then using the Chernoff bound we obtain

$$\begin{aligned} F_p(\mathrm{Thresh}) &= \mathbb{P}(\#crashed \geq b + 1) \\ &= \mathbb{P}(\#crashed \geq (p + \gamma)(4b + 1)) \\ &\leq e^{-2(4b+1)\gamma^2} \approx e^{-b(1-4p)^2/2}. \end{aligned} \tag{2}$$

Next we estimate $F_p(\mathrm{FPP})$. Let $Q_0 \in \mathrm{FPP}$ be some quorum. Then

$$F_p(\mathrm{FPP}) = 1 - \mathbb{P}(\exists Q \in \mathrm{FPP} : Q \text{ is alive}) \leq$$
$$1 - \mathbb{P}(Q_0 \text{ is alive}) = 1 - (1-p)^{q+1} \leq (q+1)p. \tag{3}$$

Using Theorem 4.7 we plug (2) into (3) to obtain

$$F_p(\mathrm{boostFPP}(q,b)) \leq (q+1)e^{-b(1-4p)^2/2} = e^{-\Omega(b-\log q)}. \qquad \Box$$

Remarks:

- In general the crash probability is not optimal; since $\mathcal{MT}(\mathrm{boostFPP}(q,b)) \approx bq$ then the lower bound of Proposition 4.3 shows we could hope for a crash probability of $\exp(-\Omega(bq))$. Nevertheless if q is constant then F_p is asymptotically optimal, and if $b \gg q$ then the gap between the upper and lower bounds is small.

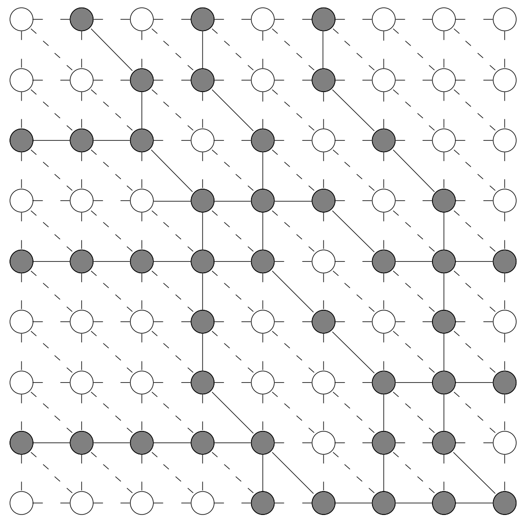

Figure 3: A multi-path construction on a 9×9 grid, $b = 4$, with one quorum shaded.

- The final estimate we get for $F_p(\text{boostFPP})$ seems poor, as the bound is *higher* than the crash probability of the Thresh components. However this is not an artifact of over-estimates in our analysis. Rather, it is a result of the property that the crash probability of FPP is higher than p, and in fact $F_p(\text{FPP}) \to 1$ as shown by [RST92, Woo96]. In this light it is not surprising that boostFPP does not have an optimal crash probability.
- The requirement $p < 1/4$ is essential for this system; if $p > 1/4$ then in fact $F_p(\text{boostFPP}) \to 1$ as $n \to \infty$.

7 The multi-path system

Here we introduce the construction we call the Multi-Path system, denoted by M-Path. The elements of this system are the vertices of a triangulated square $\sqrt{n} \times \sqrt{n}$ grid, and a quorum consists of $\sqrt{2b+1}$ disjoint paths from the left side to the right side of the grid (LR paths) and $\sqrt{2b+1}$ disjoint top-bottom (TB) paths (see Figure 3).

The M-Path system has several characteristics similar to the basic M-Grid system of Section 5, namely an ability to mask $b = O(\sqrt{n})$ failures, and optimal load. Its major advantage is that it also has an optimal crash probability F_p. Moreover, it is the only construction we have for which $F_p \to 0$ as $n \to \infty$ when the individual crash probability p is arbitrarily close to $1/2$. We are able to prove this behavior of F_p using results from Percolation theory [Kes82, Gri89].
Remark: The system we present here is based on a triangular lattice, with elements corresponding to vertices, as in [WB92, Baz96]. We have also constructed a second system which is based on the square lattice with elements corresponding to the edges, as in [NW94]. The properties of this second system are almost identical to those of M-Path, so we omit it.

Proposition 7.1 M-Path(b) *has minimal quorums of size* $c(\text{M-Path}) \leq 2\sqrt{n(2b+1)}$, *minimal intersections of size* $\mathcal{IS}(\text{M-Path}) \geq 2b+1$, *and minimal transversals of size* $\mathcal{MT}(\text{M-Path}) = \sqrt{n} - \sqrt{2b+1} + 1$. *Therefore* M-Path *is a b-masking quorum system for* $b \leq \sqrt{n} - \sqrt{2}n^{1/4}$.

Proof: Let $Q_1, Q_2 \in \text{M-Path}(b)$. Then the $\sqrt{2b+1}$ LR paths of Q_1 intersect the $\sqrt{2b+1}$ TB paths of Q_2 in $\geq 2b+1$ elements, since the LR and TB paths are disjoint. As in the M-Grid system we have that $\mathcal{MT}(\text{M-Path}(b)) = \sqrt{n} - \sqrt{2b+1} + 1$, so when $b \leq \sqrt{n} - \sqrt{2}n^{1/4}$ it follows that $\mathcal{MT}(\text{M-Path}(b)) \geq b+1$ and we are done. $\square$

Proposition 7.2 $\mathcal{L}(\text{M-Path}(b)) \leq 2\sqrt{\frac{2b+1}{n}}$, *which is optimal.*

Proof: The strategy only uses straight line LR and TB paths. It picks $\sqrt{2b+1}$ of the $\sqrt{n}$ rows uniformly at random and likewise for the columns. Clearly the load equals the probability of accessing some element in position i, j, which is

$$\begin{aligned}\mathcal{L}(\text{M-Path}) &\leq \mathbb{P}(\text{row } i \text{ chosen}) + \mathbb{P}(\text{column } j \text{ chosen}) \\ &\leq 2\binom{\sqrt{n}-1}{\sqrt{2b+1}-1} \Big/ \binom{\sqrt{n}}{\sqrt{2b+1}} \\ &= 2\sqrt{2b+1}/\sqrt{n}.\end{aligned}$$

By Corollary 4.2 this is optimal. $\square$

Proposition 7.3 $F_p(\text{M-Path}(b)) \leq \exp(-\Omega(\sqrt{n} - \sqrt{b}))$ *for any* $p < 1/2$, *which is optimal for systems with resiliency* $f = O(\sqrt{n} - \sqrt{b})$.

Proof: We use the notation $\mathbb{P}_p(\mathcal{E})$ to denote the probability of event $\mathcal{E}$ defined on the grid when the individual crash probability is p. A path is called "open" if all its elements are alive.

Let LR be the event "there exists an open LR path in the grid", and let LR_k be the event "there exist k open LR paths". A failure configuration in M-Path(b) is one in which either $\sqrt{2b+1}$ open LR paths or $\sqrt{2b+1}$ open TB paths do not exist. By symmetry we have that

$$F_p(\text{M-Path}(b)) \leq 2\mathbb{P}_p(\overline{LR_{\sqrt{2b+1}}}) = 2(1 - \mathbb{P}_p(LR_{\sqrt{2b+1}})). \quad (4)$$

Fix some p' such that $p < p' < 1/2$. Then by Theorem B.3 (see Appendix) we have that

$$1 - \mathbb{P}_p(LR_{\sqrt{2b+1}}) \leq \left(\frac{1-p}{p'-p}\right)^{\sqrt{2b+1}-1} [1 - \mathbb{P}_{p'}(LR)]. \quad (5)$$

Plugging the bound on $\mathbb{P}_{p'}(LR)$ from Theorem B.1 into (5) and (4) yields

$$\begin{aligned}F_p(\text{M-Path}(b)) &\leq 2\left(\frac{1-p}{p'-p}\right)^{\sqrt{2b+1}-1} e^{-\psi(p')\sqrt{n}} \\ &= 2e^{-\psi(p')\sqrt{n} + (\sqrt{2b+1}-1)\ln\left(\frac{1-p}{p'-p}\right)}\end{aligned}$$

for some function $\psi(p') > 0$. Now $\sqrt{2b+1} = O(n^{1/4})$, so for large enough n we can certainly write

$$F_p(\text{M-Path}(b)) \leq \exp(-\Omega(\sqrt{n} - \sqrt{b})).$$

This is optimal by Proposition 4.3. $\square$

8 Conclusion

We have presented four novel constructions of b-masking quorum systems. For the first time in this context, we considered the resiliency of such systems to crash failures in addition to their tolerance of (possibly fewer) Byzantine failures. Each of our constructions is optimal in either its load or its crash probability. Moreover, one of our constructions, namely M-Paths, is optimal in both measures. Two of our constructions are achieved using a novel boosting technique that makes all known benign fault-tolerant quorum constructions available for Byzantine environments (of appropriate sizes). In proving optimality of our constructions, we also contribute lower bounds on the load and crash probability of any b-masking quorum system.

Acknowledgments

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

References

[ACC+83] M. Aizenman, J. T. Chayes, L. Chayes, J. Fröhlich, and L. Russo. On a sharp transition from area law to perimeter law in a system of random surfaces. *Comm. Mathematical Physics*, 92:19–69, 1983.

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

[Baz96] R. Bazzi. Planar quorums. In *Proc. 10'th Inter. Workshop on Dist. Algorithms (WDAG)*, Bologna, Italy, October 1996.

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

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

[BP75] R. E. Barlow and F. Proschan. *Statistical Theory of Reliability and Life Testing.* Holt, Rinehart and Winston, Inc., 1975.

[CAA92] S. Y. Cheung, M. H. Ammar, and M. Ahamad. The grid protocol: A high performance scheme for maintaining replicated data. *IEEE Trans. Knowledge and Data Eng.*, 4(6):582–592, 1992.

[Con] N. Condorcet. Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralite des voix. Paris, 1785.

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

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

[Gif79] D. K. Gifford. Weighted voting for replicated data. In *Proc. 7th Symp. Oper. Sys. Princip.*, pages 150–159, 1979.

[Gri89] G. R. Grimmett. *Percolation.* Springer-Verlag, 1989.

[Hal86] M. Hall. *Combinatorial Theory.* John Wiley & Sons, 2nd edition, 1986.

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

[KC91] A. Kumar and S. Y. Cheung. A high availability $\sqrt{n}$ hierarchical grid algorithm for replicated data. *Inf. Proc. Letters*, 40:311–316, 1991.

[Kes80] H. Kesten. The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$. *Comm. Mathematical Physics*, 71:41–59, 1980.

[Kes82] H. Kesten. *Percolation Theory for Mathematicians.* Birkhäuser, Boston, 1982.

[KRS93] A. Kumar, M. Rabinovich, and R. K. Sinha. A performance study of general grid structures for replicated data. In *Proc. 13th Inter. Conf. Dist. Comp. Sys.*, pages 178–185, 1993.

[Kum91] A. Kumar. Hierarchical quorum consensus: A new algorithm for managing replicated data. *IEEE Trans. Comput.*, 40(9):996–1004, 1991.

[LR93] P. Lincoln and J. Rushby. A formally verified algorithm for interactive consistency under a hybrid fault model. In *Proc. 23rd IEEE Symp. Fault-Tolerant Computing (FTCS)*, pages 402–411, June 1993.

[LR94] P. Lincoln and J. Rushby. Formal verification of an interactive consistency algorithm for the Draper FTP architecture under a hybrid fault model. In *Proc. 9th IEEE Conference on Computer Assurance (COMPASS '94)*, pages 107–120, June 1994.

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

[Men86] M. V. Menshikov. Coincidence of critical points in percolation problems. *Soviet Mathematics Doklady*, 33:856–859, 1986.

[MP92] Y. Marcus and D. Peleg. Construction methods for quorum systems. Technical Report CS92-33, The Weizmann Institute of Science, Rehovot, Israel, 1992.

[MR97] D. Malkhi and M. Reiter. Byzantine quorum systems. In *Proc. 29th ACM Symp. Theory of Computing*, May 1997. To appear.

[Nei92] M. L. Neilsen. *Quorum Structures in Distributed Systems.* PhD thesis, Dept. Computing and Information Sciences, Kansas State University, 1992.

[NM92] M. L. Neilsen and M. Mizuno. Coterie join algorithm. *IEEE Trans. Par. Dist. Sys.*, 3(5):582–590, 1992.

[NW94] M. Naor and A. Wool. The load, capacity and availability of quorum systems. In *Proc. 35th IEEE Symp. Foundations of Comp. Sci. (FOCS)*, pages 214–225, 1994. To appear in SIAM J. Computing.

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

[PW97a] D. Peleg and A. Wool. The availability of crumbling wall quorum systems. *Discrete Applied Math.*, 74(1):69–83, April 1997.

[PW97b] D. Peleg and A. Wool. Crumbling walls: A class of practical and efficient quorum systems. *Distributed Computing*, 10(2):87–98, 1997.

[RB94] M. K. Reiter and K. P. Birman. How to securely replicate services. *ACM Trans. Prog. Lang. Sys.*, 16(3):986–1009, 1994.

[RST92] S. Rangarajan, S. Setia, and S. K. Tripathi. A fault-tolerant algorithm for replicated data management. In *Proc. 8th IEEE Int. Conf. Data Engineering*, pages 230–237, 1992.

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

[WB92] C. Wu and G. G. Belford. The triangular lattice protocol: A highly fault tolerant protocol for replicated data. In *Proc. 11th IEEE Symp. on Reliable Dist. Sys.*, pages 66–73, 1992.

[Woo96] A. Wool. *Quorum Systems for Distributed Control Protocols*. PhD thesis, Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot, Israel, 1996.

Appendix

A Combinatorial Lemmas

Lemma A.1 *Let $0 \le i, d \le k$ be integers. Then*

$$\frac{\binom{k}{d+i}}{\binom{k}{d}} \le \binom{k-d}{i}.$$

Proof:

$$\begin{aligned}\frac{\binom{k}{d+i}}{\binom{k}{d}} &= \frac{k!d!(k-d)!}{(d+i)!(k-d-i)!k!} = \frac{(k-d)!}{(k-d-i)!}\frac{d!}{(d+i)!} \\ &\le \frac{(k-d)!}{(k-d-i)!i!} = \binom{k-d}{i} \qquad \square\end{aligned}$$

Lemma A.2 *Let $0 \le d \le k$ be integers and let $p \in [0,1]$. Then*

$$\sum_{j=d}^{k} \binom{k}{j} p^j (1-p)^{k-j} \le \binom{k}{d} p^d.$$

Proof:

$$\sum_{j=d}^{k} \binom{k}{j} p^j (1-p)^{k-j} = \binom{k}{d} p^d \sum_{j=d}^{k} \frac{\binom{k}{j}}{\binom{k}{d}} p^{j-d}(1-p)^{k-j},$$

so it suffices to show that the last sum is ≤ 1. But using Lemma A.1 we get

$$\begin{aligned}&\sum_{j=d}^{k} \frac{\binom{k}{j}}{\binom{k}{d}} p^{j-d}(1-p)^{k-j} = \sum_{i=0}^{k-d} \frac{\binom{k}{d+i}}{\binom{k}{d}} p^{i}(1-p)^{k-d-i} \\ \le\ &\sum_{i=0}^{k-d} \binom{k-d}{i} p^i (1-p)^{k-d-i} = [p + (1-p)]^{k-d} = 1 \qquad \square\end{aligned}$$

B Theorems of Percolation Theory

In this section we list the definitions and results that are used in our analysis of the M-Path system, following [Kes82, Gri89].

The percolation model we are interested in is as follows. Let $\mathbb{Z}$ be the graph of the (infinite) triangle lattice in the plane. Assume that a vertex is *closed* with probability p and *open* with probability $1-p$, independently of other vertices. This model is known as site percolation on the triangle lattice. Another natural model, which plays a minor role in our work, is the bond percolation model. In it the *edges* are closed with probability p.

A key idea in percolation theory is that there exists a *critical probability*, p_c, such that graphs with $p < p_c$ exhibit qualitatively different properties than graphs with $p > p_c$. For example, $\mathbb{Z}$ with $p < p_c$ has a single connected (open) component of infinite size. When $p > p_c$ there is no such component. For site percolation on the triangle $p_c = 1/2$ [Kes80].

The following theorem shows that when the probability p for a closed vertex is below the critical probability, the probability of having long open paths tends to 1 exponentially fast. Recall that LR is the event "there exists an open LR path in the $\sqrt{n} \times \sqrt{n}$ grid". Then [Men86] (see also [Gri89] p. 287) implies

Theorem B.1 *If $p < 1/2$ then $\mathbb{P}_p(LR) \ge 1 - e^{-\psi(p)\sqrt{n}}$, for some $\psi(p) > 0$ independent of n.*

Remark: The dependence of ψ on p is such that $\psi(p) \to 0$ when $p \to 1/2$. However for $p \le 1/4$ we can use the estimate of Bazzi [Baz96] which shows that $\mathbb{P}_p(LR) \ge 1 - O(e^{-0.4\sqrt{n}})$.

Definition B.2 *Let $\mathcal{E}$ be an event defined in the percolation model. Then the* interior *of $\mathcal{E}$ with depth r, denoted $I_r(\mathcal{E})$, is the set of all configurations in $\mathcal{E}$ which are still in $\mathcal{E}$ even if we perturb the states of up to r vertices.*

We may think of $I_r(\mathcal{E})$ as the event that $\mathcal{E}$ occurs and is 'stable' with respect to changes in the states of r or fewer vertices. The definition is useful to us in the following situation. If LR is the event "there exists an open left-right path in a rectangle D", then it follows that $I_r(LR)$ is the event "there are at least $r+1$ disjoint open left-right paths in D".

Theorem B.3 [ACC+83] *Let $\mathcal{E}$ be an increasing event and let r be a positive integer. Then*

$$1 - \mathbb{P}_p(I_r(\mathcal{E})) \le \left(\frac{1-p}{p'-p}\right)^r [1 - \mathbb{P}_{p'}(\mathcal{E})]$$

whenever $0 \le p < p' \le 1$.

The theorem amounts to the assertion that if $\mathcal{E}$ is likely to occur when the crash probability is p', then $I_r(\mathcal{E})$ is likely to occur when the crash probability p is smaller than p'.