Probabilistic Byzantine Quorum Systems

Dahlia Malkhi* Michael Reiter* Avishai Wool† Rebecca N. Wright*

Overview

Quorums are tools for increasing the availability and efficiency of replicated data. A *quorum system* is a set of subsets (*quorums*) of servers such that every pair of quorums intersect. Recently, *probabilistic* quorum systems, in which any two chosen quorums intersect with (only) high probability, have been introduced to break availability and efficiency tradeoffs inherent in non-probabilistic (*strict*) quorum systems. However, this prior work addressed only crash failures and limited forms of Byzantine failures.

In this paper, we define *probabilistic masking quorum systems* that can be used to mask, with high probability, Byzantine server failures in their full generality. We also present a general construction for a probabilistic masking quorum system. Compared with strict quorum systems that can mask Byzantine server failures, our approach yields dramatically better data availability in the face of crash failures, and does so while simultaneously decreasing the load on servers.

Probabilistic masking quorum systems

To recall the definition of strict masking quorum systems, let b denote the bound on the number of Byzantine server failures that may occur. Then a strict b-masking quorum system is one in which any two quorums intersect in at least $2b+1$ elements. 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. 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.

Our definition of a probabilistic masking quorum system relaxes this constraint on intersection size. In general, a probabilistic quorum system consists of a set of subsets (quorums) of servers and a *strategy* w for accessing quorums, i.e., a probability distribution on quorums that captures the probability that each quorum is accessed in each client operation. Were we to adapt the prior probabilistic quorum system definition to the masking case in the straightforward way, we would require that any two quorums selected according to w intersect in at least $2b+1$ elements with high probability. However, this definition does not yield the same performance benefits as the probabilistic approach does for benign-fault-tolerant quorum systems. For example, the *load*—i.e., the probability with which the busiest server is accessed in each client operation—for any such system with $b = \Theta(n)$ would be constant, which is poor.

The trouble with the above definition is that it is stronger than necessary, as it requires that with high probability, the set $Q \cap Q' \setminus B$ be so large that it is *impossible* that $Q \cap B$ is 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. Accordingly, our definition of a probabilistic masking quorum system employs a threshold value k that we expect to be greater than $|Q \cap B|$ but less than $|Q \cap Q' \setminus B|$. Thus, a 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 1 *Let $\mathcal{Q}$ be a set of subsets of servers, w be a strategy for $\mathcal{Q}$, and let an integer k and $0 < \varepsilon < 1$ be given. The tuple $\langle \mathcal{Q}, w, k\rangle$ is a* probabilistic (b, ε)-masking quorum system *if for all $B \subseteq U$ such that $|B| = b$,*

$$\mathbf{P}(|Q \cap B| < k \ \wedge\ |Q \cap Q' \setminus B| \geq k) \quad \geq \quad 1 - \varepsilon,$$

where the probability is taken over selections of Q and Q' according to w.

We show two lower bounds on the load of any probabilistic (b, ε)-masking quorum system, namely $\frac{1-\sqrt{\varepsilon}}{\sqrt{n}}$ and $\left(\frac{1-2\varepsilon}{1-\varepsilon}\right)\frac{b}{n}$.

A probabilistic masking quorum system construction

A construction for a probabilistic masking quorum system that works for any $b < n/2$ in a system with n servers is as follows: $\mathcal{Q} = \{Q : |Q| = q = \ell b\}$ where $2 < \ell < n/b$; $w(Q) = 1/|\mathcal{Q}|$ for all $Q \in \mathcal{Q}$; and $k = q^2/2n$. Then $\langle \mathcal{Q}, w, k\rangle$ is a (b, ε)-masking quorum system for $\varepsilon = 2\exp(-\Omega(q^2/n))$. In particular, if $q = \omega(\sqrt{n})$, then $\varepsilon \to 0$ as $n \to \infty$.

This construction has load $q/n = \ell b/n$, which is within a factor of $\ell\left(\frac{1-\varepsilon}{1-2\varepsilon}\right)$ of the lower bound for probabilistic masking quorum systems. And, because we can choose ℓ to be a constant when $b = \omega(\sqrt{n})$, it is asymptotically load-optimal (as a function of b and n) for $b = \omega(\sqrt{n})$. This construction also has optimal availability, since some quorum will be available even if up to $n - q = \Theta(n)$ servers crash.

*AT&T Labs—Research, Florham Park, New Jersey, USA; {dalia,reiter,rwright}@research.att.com

†Bell Labs, Lucent Technologies, Murray Hill, New Jersey, USA; yash@research.bell-labs.com

In *Proceedings of Seventeenth Annual ACM Symposium on Principles of Distributed Computing*, Puerto Vallarta, Mexico, May 1998.