

Minimizing Content Staleness in Dynamo-Style Replicated Storage Systems

Jing Zhong, Roy D. Yates and Emina Soljanin

Dept. of ECE, Rutgers University, {jing.zhong, ryates, emina.soljanin}@rutgers.edu

Abstract—Consistency in data storage systems requires any read operation to return the most recent written version of the content. In replicated storage systems, consistency comes at the price of delay due to large-scale write and read operations. Many applications with low latency requirements tolerate data staleness in order to provide high availability and low operation latency. Using age of information as the staleness metric, we examine a data updating system in which real-time content updates are replicated and stored in a Dynamo-style quorum-based distributed system. A source sends updates to all the nodes in the system and waits for acknowledgements from the earliest subset of nodes, known as a write quorum. An interested client fetches the update from another set of nodes, defined as a read quorum. We analyze the staleness-delay tradeoff in replicated storage by varying the write quorum size. With a larger write quorum, an instantaneous read is more likely to get the latest update written by the source. However, the age of the content written to the system is more likely to become stale as the write quorum size increases. For shifted exponential distributed write delay, we derive the age optimized write quorum size that balances the likelihood of reading the latest update and the freshness of the latest update written by the source.

I. INTRODUCTION

In modern distributed storage systems, data is often replicated across multiple machines or datacenters to support fault tolerance due to server failures, and provide high availability by delivering data through replica servers. In order to overcome asynchrony in distributed storage systems, *quorum*-based algorithms [1]–[4] are well studied and widely used in practice to ensure write and read consistency of replicated data. In a quorum system, either a write or read of the data goes to a subset of nodes. More specifically, the data source or writer sends replicas of the data to all the nodes but only waits for the response from a subset of nodes known as the *write quorum* $\mathcal{W}$. The client or reader fetches the data from a possibly different subset of nodes, which is called the *read quorum* $\mathcal{R}$. In order to guarantee strict consistency that every read operation returns the most recent written content, a traditional quorum system requires the write quorum $\mathcal{W}$ and read quorum $\mathcal{R}$ to overlap in at least one element. This is known as a *strict quorum*. When a quorum is randomly selected by the writer and reader, a strict quorum requires that the write quorum size w and read quorum size r satisfy $w + r > n$, where n is the number of servers/nodes in the system.

However, as the write or read operation to a set of nodes experiences varying random delays, consistency of the storage system comes at the price of delay. A wide range of applications have crucial delay requirements, e.g., every 100ms of extra

latency cost Amazon 1% in sales, and an extra 0.5s delay in search results cuts Google’s traffic by 20% [5]. It has been shown that a non-strict or partial quorum of reduced size is widely used in practice because of the latency benefit despite a minor loss in consistency [6]. Amazon’s Dynamo database [7], and a variety of subsequent database implementations such as Apache Cassandra [8], use a non-strict quorum as the data replication mechanism in order to maintain a balance between consistency and latency.

Since strict consistency is not guaranteed in partial quorum systems, the level of consistency is quantified by data *staleness*. The definition of data staleness falls into two categories: 1) staleness in time [9] [10] and 2) staleness in data version [6]. In [9], a read is considered stale if the value returned was written more than δ time units before the most recent write, where δ is a pre-determined threshold. In a slightly different time-based staleness definition [10], the data is considered fresh if it was generated no more than δ time units ago or it’s the most recent written data in the system. On the other hand, [6] measures the staleness by how many versions the value returned by a read lags behind the most recent write.

In this work, we characterize data staleness from a strictly time-based perspective. We examine data monitoring and gathering systems in which real-time content updates generated by a source are stored in a quorum-based distributed database, and a client connects to the database and requests the most recent data update. Since the stored data in a node is desired to be as recent as possible, a write operation to a quorum of nodes can be seen as an information update by the source. In these applications, the freshness/staleness of the information updates is measured by an “Age of Information” (AoI) timeliness metric [11]–[15]. If a client reads data by fetching from a set of nodes at some time t , and the data has a version time-stamped $u(t)$, then the *age* of the data at the client is $t - u(t)$. We note that the age in time differs from other staleness metrics in distributed storage systems, since we take the age of the most recent written content into account. We start by first considering the baseline problem: in a distributed storage system with n nodes, how does the size of the write quorum w and read quorum r affect the average age of the content returned by a read?

II. SYSTEM MODEL AND METRIC

A. Staleness in Dynamo-style Systems

In a Dynamo-style replicated quorum system, a write/read request by a user will be replicated and sent to *all* the nodes in the system. The write operation is considered completed

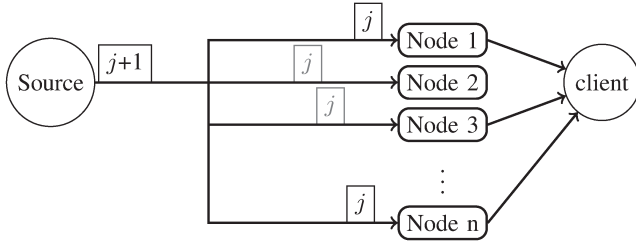


Fig. 1: Dynamo-style distributed storage: the source sequentially writes content updates to multiple nodes with random write delays. The next write $j + 1$ is initiated right after update j is written to w out of n nodes. The client reads the content through a random set of r nodes and selects the freshest version.

only if the user receives at least w acknowledgements from the n nodes in the system, where w is the pre-determined write quorum size. Similarly, the read operation is successful only if the user gets r responses from the system, where r is the read quorum size. In this paper, we consider a Dynamo-style replicated quorum system with a writing source and a reading client as shown in Fig. 1. The source writes a time-stamped data item by replicating it to n different nodes in the system with independent random write delays. An update takes time X_i to be written to node i . We refer to X_i as the write delay for node i . We assume that the X_i are i.i.d. shifted exponential (λ, c) random variables. Consequently, each X_i has CDF

$$F_X(x) = 1 - e^{-\lambda(x-c)}, \quad x \geq c. \quad (1)$$

The constant time shift $c > 0$ captures the delay produced by the update generation and assembly process. On the other hand, c can also represent a propagation delay on top of an exponential network delay if the source and database are geographically separated. When freshest received data at time t at node i is time-stamped time $u_i(t)$, the age of information or simply the *age* at node i , is the random process $\Delta_i(t) = t - u_i(t)$. When a data item reaches node i , $u_i(t)$ is advanced to the timestamp of the new content and the node sends an acknowledgement to the source through a feedback channel that is assumed to be instantaneous. Assuming that the data item is a time-sensitive content update, the source then obeys a zero-wait policy and initiates a new write request as soon as the current write is completed. At the same time, the write operation to the remaining $n - w$ nodes are canceled.

To read data, a client connects to a random set of r nodes in the system, and selects the freshest content among all r nodes. In [6], the average read latency in Basho Riak [16], a commercial distributed database, is shown to be an order of magnitude smaller than the average write latency. Hence, we assume the read process is instantaneous with zero delay in this work. Under this model, the age at the read client at time t is defined as the minimum age over all the nodes in the read quorum $\mathcal{R}$, i.e.

$$\Delta(t) = \min_{i \in \mathcal{R}} \Delta_i(t). \quad (2)$$

Since the write delays are i.i.d. for each node i and content

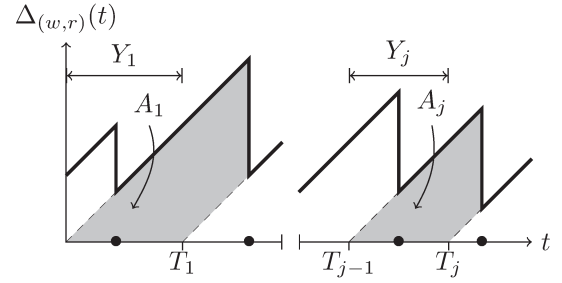


Fig. 2: Sample path of the age $\Delta_{(w,r)}(t)$ with strict quorum $w + r > n$. Update delivery instances are marked by $\bullet$.

update j , the $\Delta_i(t)$ processes at each node are statistically identical. The age processes for different read quorum $\mathcal{R}$ are also statistically identical since $\mathcal{R}$ is randomly chosen. The time average of age process at the client is then given by

$$\Delta = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{t=0}^{\tau} \Delta(t) dt. \quad (3)$$

The mathematical model we consider is also relevant to other status updating systems, e.g., multicast with HARQ, where single source transmits coded update packets to multiple clients [17]. In related work [18], update messages are replicated and sent to the receiver through multiple servers; given a general packet arrival process and memoryless packet service times, it was shown that Last-Generated First-Serve scheduling policy is age-optimal. In [19], a pull-based updating system is considered, in which the arriving source updates are sent to multiple servers and the interested user fetches the update by sending replicated requests to all the servers. Similar to this work, it was shown there exists an optimal number of responses k from n servers for the user to wait for. The problem considered in this work differs from [19] by allowing the source to control when to submit an update based on the delivery feedback.

B. Order Statistics

We first introduce the notion of order statistics for i.i.d. random variables which plays a key role in our analysis. We denote the k -th order statistic of the random variables $X_1, \dots, X_n$, i.e., the k -th smallest variable, as $X_{k:n}$.

Lemma 1. [20] For shifted exponential random variable X with CDF $F_X(x) = 1 - e^{-\lambda(x-c)}, x \geq c$, the expectation and variance of the order statistics $X_{k:n}$ are given by

$$E[X_{k:n}] = c + \frac{1}{\lambda} (H_n - H_{n-k}) \quad (4)$$

$$\text{Var}[X_{k:n}] = \frac{1}{\lambda^2} (H_{n^2} - H_{(n-k)^2}), \quad (5)$$

where H_n and H_{n^2} are the generalized harmonic numbers defined as $H_n = \sum_{j=1}^n \frac{1}{j}$ and $H_{n^2} = \sum_{j=1}^n \frac{1}{j^2}$.

III. AGE ANALYSIS

Our objective is to obtain the average of the age (2) of the data content given by a read at any time t , which is determined by the minimum age over all nodes in the read quorum $\mathcal{R}$.

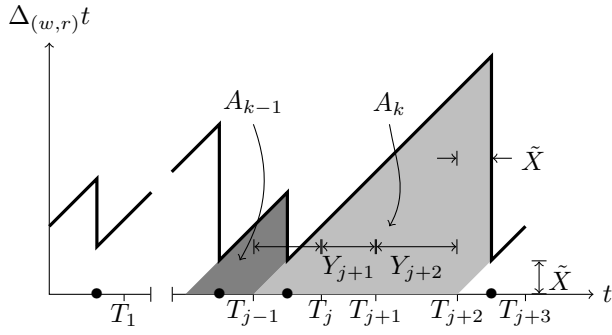


Fig. 3: Sample path of the age $\Delta_{(w,r)}(t)$ with non-strict quorum $w + r \leq n$: successful updates (at times marked by $\bullet$) occur in intervals 1, $j - 1$, j , and $j + 3$.

Since the read operation is instantaneous by assumption, this is equivalent to evaluating the time-averaged minimum age of any set of r nodes, where $r \in \{1, 2, \dots, n\}$.

With a write quorum of size $w \in \{1, 2, \dots, n\}$, the source considers the current write committed and initiates the next write operation only after the current update is successfully written to w out of n nodes. In this case, we denote the write delay for node i and content update j as X_{ij} , and the total write delay for update j as Y_j . Note that Y_j is statistically identical to $X_{w:n}$. If one node gets the content earlier than any of the other nodes, it has to wait for an idle period until that content is written to $w - 1$ other nodes. Otherwise, if update j is not written to a node i , the node waits for time Y_j until the source starts the next write, and we refer to this random waiting time as a *write interval* for w nodes.

For a strict quorum system with $r + w > n$, there is at least one overlapping node between the read quorum $\mathcal{R}$ and the write quorum $\mathcal{W}$, thus every update will be eventually written to at least one nodes in the read quorum $\mathcal{R}$. That is, if the client reads the data content right after the most recent committed write, it is guaranteed to receive the most up-to-date update committed by the source. On the other hand, it is also possible that a read operation returns the most recent write even before the commitment since the client selects the freshest content from all nodes in the read quorum.

Fig. 2 depicts a sample path of the minimum age of a random set of r nodes which satisfies the strict quorum requirement. The start and end time of the j -th write interval are marked as T_{j-1} and $T_j = T_{j-1} + Y_j$, respectively. For update j , let's denote the earliest node to complete the write process in the read quorum as node i . The instantaneous age $\Delta(t)$ drops to exactly the write delay for node i , denoted by X_{ij} , when the earliest node i receives the update j . This also implies at time T_j when update j is written to the write quorum $\mathcal{W}$, the age of the read quorum is $\Delta(T_i) = Y_j$.

For a non-strict or partial quorum system with $r + w \leq n$, an update by the source may not be successfully read by the client because the random sets $\mathcal{W}$ and $\mathcal{R}$ can be disjoint. Suppose an update is written to at least one node in the read quorum $\mathcal{R}$ during write interval j and the next successful write

to read quorum $\mathcal{R}$ is in write interval $j + M$. In this case, M is a geometric r.v. with probability mass function (PMF) $P_M(m) = (1 - p)^{m-1}p, m \geq 1$. Thus M has first and second moments

$$\mathbb{E}[M] = \frac{1}{p}, \quad \mathbb{E}[M^2] = \frac{2-p}{p^2}. \quad (6)$$

A fresh update fails to be written to the client during a write interval if and only if the write quorum $\mathcal{W}$ and read quorum $\mathcal{R}$ do not overlap. Let's denote the probability of a write failure as $q = 1 - p$, then

$$q = \Pr\{\mathcal{R} \cap \mathcal{W} = \emptyset\} = \begin{cases} 0, & r + w > n \\ \binom{n-w}{r} / \binom{n}{r}, & r + w \leq n. \end{cases} \quad (7)$$

A similar example of the age process is shown in Fig. 3. We represent the area under the age sawtooth as the concatenation of the polygons $A_1, \dots, A_k$ as shown in Figs. 2 and 3. The update j is written to the read quorum $\mathcal{R}$ in the write interval j , and the read quorum waits for $M_k = 3$ write intervals until the next successful write. Note that a strict quorum can be viewed as a special case with deterministic $M_k = 1$. Denote the random variable $\tilde{X}$ as the write delay of a successful update written to at least one node in the read quorum r . Evaluating Fig. 3 gives the area

$$A_k = \frac{1}{2} \left(\sum_{l=j}^{j+M_k-1} Y_l + \tilde{X}_k \right)^2 - \frac{1}{2} \tilde{X}_k^2. \quad (8)$$

Lemma 2. The average area A_k as shown in Fig. 3 is

$$\begin{aligned} \mathbb{E}[A] &= \mathbb{E}[\tilde{X}] \mathbb{E}[M] \mathbb{E}[Y] \\ &\quad + \frac{1}{2} \mathbb{E}[M^2] (\mathbb{E}[Y])^2 + \frac{1}{2} \mathbb{E}[M] \text{Var}[Y]. \end{aligned}$$

Proof. Defining $W = \sum_{l=j}^{j+M_k-1} Y_l$, (8) can be rewritten as

$$A_k = \frac{1}{2} [W^2 + 2\tilde{X}_{k-1}W + \tilde{X}_k^2] - \frac{1}{2} \tilde{X}_k^2.$$

Since M_k and the Y_j are independent, $\mathbb{E}[W] = \mathbb{E}[M] \mathbb{E}[Y]$. It follows that

$$\mathbb{E}[A_k] = \frac{1}{2} \mathbb{E}[W^2] + \mathbb{E}[\tilde{X}] \mathbb{E}[M] \mathbb{E}[Y]. \quad (9)$$

The random sum of random variables W has second moment

$$\begin{aligned} \mathbb{E}[W^2] &= (\mathbb{E}[W])^2 + \text{Var}[W] \\ &= (\mathbb{E}[M])^2 (\mathbb{E}[Y])^2 + \mathbb{E}[M] \text{Var}[Y] + \text{Var}[M] (\mathbb{E}[Y])^2 \\ &= \mathbb{E}[M^2] (\mathbb{E}[Y])^2 + \mathbb{E}[M] \text{Var}[Y]. \end{aligned}$$

Substituting $\mathbb{E}[W^2]$ back into (9) completes the proof. $\square$

It follows from Fig. 3 that the average age is given by

$$\Delta_{(w,r)} = \frac{\mathbb{E}[A]}{\mathbb{E}[M] \mathbb{E}[Y]}. \quad (10)$$

Theorem 1. Consider a Dynamo-style n -node quorum system with write quorum size w and read quorum size r . The source sequentially writes content updates to the system. Assuming

the freshest content is selected from the read quorum and the read operation is instantaneous, the average age of the content observed by the client is

1) for $w + r > n$,

$$\Delta_{(w,r)} = \sum_{i=1}^w \mathbb{E}[X_{i:n}] \frac{\binom{n-i}{r-1}}{\binom{n}{r}} + \frac{1}{2} \frac{\mathbb{E}[X_{w:n}^2]}{\mathbb{E}[X_{w:n}]};$$

2) for $w + r \leq n$,

$$\begin{aligned} \Delta_{(w,r)} &= \sum_{i=1}^w \mathbb{E}[X_{i:n}] \frac{\binom{n-i}{r-1}}{\binom{n}{r} - \binom{n-w}{r}} \\ &\quad + \frac{1}{2} \frac{\binom{n}{r} + \binom{n-w}{r}}{\binom{n}{r} - \binom{n-w}{r}} \mathbb{E}[X_{w:n}] + \frac{1}{2} \frac{\text{Var}[X_{w:n}]}{\mathbb{E}[X_{w:n}]}. \end{aligned}$$

Proof. Substituting Lemma 2 into (10) yields

$$\Delta_{(w,r)} = \mathbb{E}[\tilde{X}] + \frac{\mathbb{E}[M^2]}{2\mathbb{E}[M]} \mathbb{E}[Y] + \frac{1}{2} \frac{\text{Var}[Y]}{\mathbb{E}[Y]}, \quad (11)$$

where M is the number of write intervals between successful writes to the read quorum. With $p = 1 - q$, it follows from (6) and (7) that

$$\frac{\mathbb{E}[M^2]}{2\mathbb{E}[M]} = \frac{1+q}{2(1-q)} = \begin{cases} \frac{1}{2}, & r+w > n \\ \frac{1}{2} \frac{\binom{n}{r} + \binom{n-w}{r}}{\binom{n}{r} - \binom{n-w}{r}}, & r+w \leq n. \end{cases} \quad (12)$$

Thus, for strict quorum $w + r > n$, (11) can be written as

$$\begin{aligned} \Delta_{(w,r)} &= \mathbb{E}[\tilde{X}] + \frac{1}{2} \mathbb{E}[Y] + \frac{1}{2} \frac{\text{Var}[Y]}{\mathbb{E}[Y]} \\ &= \mathbb{E}[\tilde{X}] + \frac{1}{2} \frac{\mathbb{E}[Y^2]}{\mathbb{E}[Y]}. \end{aligned} \quad (13)$$

Denote the node $\hat{i}_{\mathcal{R}}$ as the node with least write delay in the read quorum $\mathcal{R}$, i.e.,

$$\hat{i}_{\mathcal{R}} = \arg \min_{i \in \mathcal{R}} X_i.$$

In addition, we rewrite the write quorum set as

$$\mathcal{W} = \{i_1, i_2, \dots, i_w\}, \quad (14)$$

where $X_{i_k} = X_{k:n}$ is the k -th smallest write delay in the write quorum.

For strict quorum $w + r > n$, the average write delay for a successful update read by the client is given by

$$\mathbb{E}[\tilde{X}] = \mathbb{E}[X_{\hat{i}_{\mathcal{R}}} | \hat{i}_{\mathcal{R}} \in \mathcal{W}] \quad (15a)$$

$$= \sum_{k=1}^w \mathbb{E}[X_{k:n}] \Pr\{\hat{i}_{\mathcal{R}} = i_k | \hat{i}_{\mathcal{R}} \in \mathcal{W}\}, \quad (15b)$$

$$= \sum_{k=1}^w \mathbb{E}[X_{k:n}] \frac{\Pr[\hat{i}_{\mathcal{R}} = i_k]}{1-q}, \quad (15c)$$

$$= \sum_{k=1}^w \mathbb{E}[X_{k:n}] \frac{\binom{n-k}{r-1}}{\binom{n}{r}}. \quad (15d)$$

In (15a), the expectation of $\tilde{X}$ is defined as the expectation of

the minimum of all the write delays X_i in the read quorum $\mathcal{R}$, given that this minimum is also in the write quorum $\mathcal{W}$. (15b) is obtained by averaging over the conditional expectation of all possible order statistics $X_{k:n}$. And the complementary event of the condition $\hat{i}_{\mathcal{R}} \in \mathcal{W}$ is that both subsets do not overlap, $\mathcal{R} \cap \mathcal{W} = \emptyset$, which yields (15c). From (15c) to (15d), it follows from (7) that $q = 0$. And we have $\Pr[\hat{i}_{\mathcal{R}} = i_k] = \binom{n-k}{r-1} / \binom{n}{r}$, since $X_{k:n}$ is smallest in the read quorum $\mathcal{R}$ with size r , and the remaining $r-1$ values are randomly chosen from the subset $\{X_{k+1:n}, \dots, X_{n:n}\}$.

Similarly, for the non-strict quorum with $w + r \leq n$,

$$\mathbb{E}[\tilde{X}] = \mathbb{E}[X_{\hat{i}_{\mathcal{R}}} | \hat{i}_{\mathcal{R}} \in \mathcal{W}] \quad (16a)$$

$$= \sum_{k=1}^{n-r+1} \mathbb{E}[X_{k:n}] \Pr\{\hat{i}_{\mathcal{R}} = i_k | \hat{i}_{\mathcal{R}} \in \mathcal{W}\}, \quad (16b)$$

$$= \sum_{k=1}^{n-r+1} \mathbb{E}[X_{k:n}] \frac{\Pr[\hat{i}_{\mathcal{R}} = i_k]}{1-q}, \quad (16c)$$

$$= \sum_{k=1}^{n-r+1} \mathbb{E}[X_{k:n}] \frac{\binom{n-k}{r-1}}{\binom{n}{r} - \binom{n-w}{r}}. \quad (16d)$$

From (16c) to (16d), we apply $q = \binom{n-w}{r} / \binom{n}{r}$ in (7) for the case $r + w \leq n$. Note that the length of a write interval is $Y = X_{w:n}$. To get Theorem 1, we substitute (15d) back to (13) for $w + r > n$, and substitute (12) and (16d) back to (11) for $w + r \leq n$. $\square$

Corollary 1. Let $\beta = 1 - \alpha = 1 - w/n$. For shifted exponential (λ, c) write delay X and a given read quorum size r , the average age at the client can be approximated for large n as:

1) for $w + r > n$,

$$\Delta_{(w,r)} \approx \frac{1-2\beta^r}{2\lambda} \log \frac{1}{\beta} + (1-\beta^r)(c + \frac{1}{\lambda r}) + \frac{c}{2}.$$

2) for $w + r \leq n$,

$$\Delta_{(w,r)} \approx \frac{1}{\lambda r} + \frac{1}{2\lambda} \log \frac{1}{\beta} + c + \frac{c(1+\beta^r)}{2(1-\beta^r)}.$$

Corollary 2. Denote $\omega = \beta^r$, the optimal ω^* that minimizes Corollary 1 for positive λ and c is

$$\omega^* = (\lambda cr + 1) - \sqrt{(\lambda cr + 1)^2 - 1}. \quad (17)$$

Proofs for both corollaries are provided in the appendix.

IV. EVALUATION

Figure 4 compares the simulation results of the average age $\Delta_{(w,r)}$ as a function of the write quorum size w given different pre-determined read quorum r . In this experiment, the total number of nodes $n = 100$, and every write to a node has shifted exponential delay with $c = 1$ and different λ . The approximation in Corollary 1 is marked with $\times$, and the near-optimal write quorum size w^* in Corollary 2 is marked with $\circ$. By looking at the three curves in a single figure, We observe that for a given c and n , the optimal write quorum size

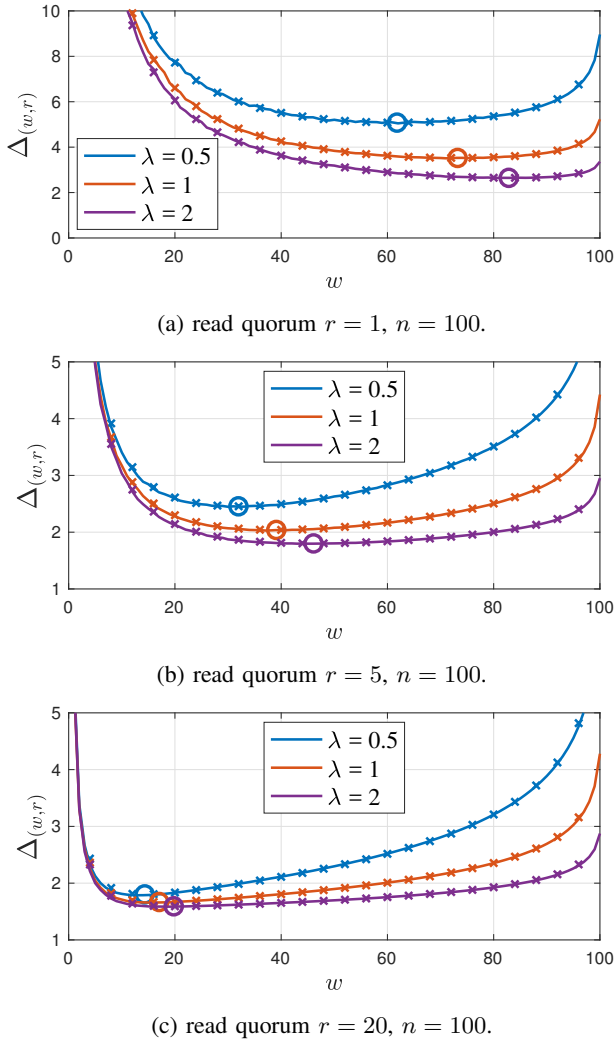


Fig. 4: Average age as a function of write quorum w for shifted exponential write delay. $\times$ marks the approximate average age, and $\circ$ marks the minimized approximate age $\hat{\Delta}(w^*)$.

w^* increases as the exponential rate λ increases. Comparing across figures with different read quorum r , we observe that the average age decreases as r increases, and the optimal write quorum size w also decreases significantly. For example, the client reads the data from only a single node in Fig. 4a. In this case, in order to minimize the staleness of the content update, the source should consider a write operation to be complete once the update is written to 60 out of 100 nodes when $\lambda = 0.5$. However, if the client is connected to 5 nodes as shown in Fig. 4b, it is best for the source to only write to around 30 nodes. This observation also implies that choosing a partial/non-strict quorum is usually a winning strategy to minimize the content staleness in replicated storage system. Although strict quorum guarantees that any read after the write commit can return the most recent write, the content written to the system is more likely to become stale due to larger write delay.

V. CONCLUSION

We have examined a real-time data replicated storage system in which content updates are replicated and stored in a dynamo-style quorum systems. Either a write or read request goes to all the nodes in the system, and it is considered complete once there are at least w or r responses. Assuming that the write delay dominates the latency, the freshness of the replicated storage system is measured by the average age of the content returned by a read at any time t . As the write quorum size w increases, an instantaneous read from a read quorum with predetermined size r is more likely to get the latest version generated by the source. However, the age of the content also increases as the write quorum size w increases. We have derived the optimal w given a read quorum size r such that the average age of the content returned by a read is minimized, and showed by experiment that the optimal w satisfies the non-strict quorum $w + r \leq n$.

The analysis presented in this work is based on the assumption that the read delay is negligible compared to the write delay, and the feedback channels from the nodes are instantaneous. We are also aware of more general cases where the read delay is also significant and random such that the content may become stale during the read process. Under this scenario, a different write quorum w should be chosen to deal with the possibility of stale data due to the read delay.

ACKNOWLEDGMENT

Part of this research is based upon work supported by the National Science Foundation under grant CNS-1422988.

REFERENCES

- [1] H. Attiya, A. Bar-Noy, and D. Dolev, "Sharing Memory Robustly in Message-Passing Systems," *J. ACM*, pp. 1–12, 1995.
- [2] D. Peleg and A. Wool, "The availability of quorum systems," *Information and Computation*, vol. 123, no. 2, pp. 210–223, 1995.
- [3] D. Malkhi, M. Reiter, and R. Wright, "Probabilistic quorum systems," in *Proceedings of ACM Symposium on Principles of Distributed Computing*, 1997, pp. 267–273.
- [4] D. Malkhi and M. Reiter, "Byzantine quorum systems," *Distributed Computing*, vol. 11, no. 4, pp. 203–213, 1998.
- [5] H. Scalability, "Latency is everywhere and it casts you sales - How to crush it," 2009, <http://highscalability.com/latency-everywhere-and-it-costs-you-sales-how-crush-it>.
- [6] P. Bailis, S. Venkataraman, M. J. Franklin, J. M. Hellerstein, and I. Stoica, "Probabilistically Bounded Staleness for Practical Partial Quorums," *Proceeding of VLDB*, 2012.
- [7] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Voshall, and W. Vogels, "Dynamo: amazon's highly available key-value store," *ACM SIGOPS operating systems review*, vol. 41, no. 6, pp. 205–220, 2007.
- [8] A. Lakshman and P. Malik, "Cassandra: A decentralized structured storage system," *SIGOPS Oper. Syst. Rev.*, vol. 44, no. 2, Apr. 2010.
- [9] W. Golab, X. Li, and M. A. Shah, "Analyzing Consistency Properties for Fun and Profit," in *ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC)*, New York, NY, USA, 2011, pp. 197–206.
- [10] M. R. Rahman, L. Tseng, S. Nguyen, I. Gupta, and N. Vaidya, "Characterizing and Adapting the Consistency-Latency Tradeoff in Distributed Key-Value Stores," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 11, no. 4, pp. 1–36, 2017.
- [11] S. Kaul, R. D. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. INFOCOM*, Apr. 2012, pp. 2731–2735.
- [12] M. Costa, M. Codreanu, and A. Ephremides, "Age of information with packet management," in *Proc. IEEE Int. Symp. Inform. Theory*, 2014, pp. 1583–1587.

- [13] L. Huang and E. Modiano, "Optimizing age-of-information in a multi-class queueing system," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun. 2015, pp. 1681–1685.
- [14] Y. Sun, E. Uysal-Biyikoglu, R. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," in *Proc. INFOCOM*, 2016.
- [15] E. Najm, R. D. Yates, and E. Soljanin, "Status updates through m/g/1/1 queues with harq," in *Proc. IEEE Int. Symp. Inform. Theory*, 2017.
- [16] Basho Riak, <http://www.basho.com/products/#riak>.
- [17] R. D. Yates, E. Najm, E. Soljanin, and J. Zhong, "Timely updates over an erasure channel," in *Proc. IEEE Int. Symp. Inform. Theory*, 2017.
- [18] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Minimizing the age of the information through queues," *arXiv preprint arXiv:1709.04956*, 2017.
- [19] Y. Sang, B. Li, and B. Ji, "The power of waiting for more than one response in minimizing the age-of-information," *arXiv preprint arXiv:1704.04848*, 2017.
- [20] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A first course in order statistics*. SIAM, 2008.

APPENDIX

Proof of Corollary 1: For shifted exponential r.v. X ,

$$\begin{aligned} \frac{E[X_{w:n}^2]}{E[X_{w:n}]} &= E[X_{w:n}] + \frac{\text{Var}[X_{w:n}]}{E[X_{w:n}]} \\ &= E[X_{w:n}] + \frac{H_{n^2} - H_{(n-w)^2}}{2\lambda^2 c + 2\lambda(H_n - H_{n-w})}. \end{aligned} \quad (18)$$

Note that the sequence H_{n^2} is monotonically increasing and $\lim_{n \rightarrow \infty} H_{n^2} = \pi^2/6$, thus $H_{n^2} - H_{(n-w)^2}$ is negligible and

$$\lim_{n \rightarrow \infty} \frac{\text{Var}[X_{w:n}]}{E[X_{w:n}]} = 0, \quad (19)$$

It follows from (19) that

$$\lim_{n \rightarrow \infty} \frac{E[X_{w:n}^2]}{E[X_{w:n}]} = E[X_{w:n}]. \quad (20)$$

With large n , we also approximate the harmonic number by $H_i \approx \log i + \gamma$, thus

$$\begin{aligned} E[X_{w:n}] &\approx c + \frac{1}{\lambda} (\log n - \log(n-w)) \\ &= c + \frac{1}{\lambda} \left(\log \frac{n}{n-w} \right). \end{aligned} \quad (21)$$

Note that this approximation only holds when $w < n$.

Let's substitute (19) and (21) into Theorem 1, and approximate the binomial coefficient by $\binom{n}{k} \approx \frac{n^k}{k!}$. For $w+r > n$, Theorem 1 is then rewritten as

$$\Delta \approx \sum_{i=1}^w E[X_{i:n}] \frac{(n-i)^{r-1} r}{n^r} + \frac{1}{2} E[X_{w:n}] \quad (22a)$$

$$\begin{aligned} &\approx \sum_{i=1}^w \left(\frac{1}{\lambda} \log \left(\frac{n}{n-i} \right) + c \right) \frac{(n-i)^{r-1} r}{n^r} \\ &\quad + \frac{1}{2\lambda} \log \left(\frac{n}{n-w} \right) + \frac{c}{2} \end{aligned} \quad (22b)$$

$$\begin{aligned} &\approx r \int_{x=0}^{\alpha=\frac{w}{n}} \left(\frac{1}{\lambda} \log \left(\frac{1}{1-x} \right) + c \right) (1-x)^{r-1} dx \\ &\quad + \frac{1}{2\lambda} \log \left(\frac{1}{1-\alpha} \right) + \frac{c}{2} \end{aligned} \quad (22c)$$

$$= \frac{(1 - (1-\alpha)^r (1 - r \log(1-\alpha)))}{\lambda r} + c(1 - (1-\alpha)^r)$$

$$+ \frac{1}{2\lambda} \log \left(\frac{1}{1-\alpha} \right) + \frac{c}{2} \quad (22d)$$

$$\begin{aligned} &= \frac{1 - 2(1-\alpha)^r}{2\lambda} \log \frac{1}{1-\alpha} \\ &\quad + (1 - (1-\alpha)^r) \left(c + \frac{1}{\lambda r} \right) + \frac{c}{2}. \end{aligned} \quad (22e)$$

In (22b), we use the limit in (20) as an approximate. In (22c), we denote $\alpha = w/n$ and approximate the sum $\sum_{i=1}^w f(i)$ by the integral $\int_{i=0}^w f(i) di$.

For $w+r \leq n$,

$$\begin{aligned} \Delta &\approx \sum_{i=1}^w E[X_{i:n}] \frac{(n-i)^{r-1} r}{n^r - (n-w)^r} \\ &\quad + \frac{n^r + (n-w)^r}{2(n^r - (n-w)^r)} E[X_{w:n}] \end{aligned} \quad (23a)$$

$$\begin{aligned} &\approx \sum_{i=1}^w \left(\frac{1}{\lambda} \log \left(\frac{n}{n-i} \right) + c \right) \frac{(n-i)^{r-1} r}{n^r - (n-w)^r} \\ &\quad + \frac{n^r + (n-w)^r}{2(n^r - (n-w)^r)} \left(\frac{1}{\lambda} \log \left(\frac{n}{n-w} \right) + c \right) \end{aligned} \quad (23b)$$

$$\begin{aligned} &\approx \int_{x=0}^{\alpha} \left(\frac{1}{\lambda} \log \left(\frac{1}{1-x} \right) + c \right) \frac{(1-x)^{r-1} r}{1 - (1-\alpha)^r} dx \\ &\quad + \frac{1 + (1-\alpha)^r}{2(1 - (1-\alpha)^r)} \left(\frac{1}{\lambda} \log \left(\frac{1}{1-\alpha} \right) + c \right) \end{aligned} \quad (23c)$$

$$\begin{aligned} &= \frac{1}{\lambda c} - \frac{(1-\alpha)^r}{\lambda(1 - (1-\alpha)^r)} \log \left(\frac{1}{1-\alpha} \right) + c \\ &\quad + \frac{(1 + (1-\alpha)^r)}{2\lambda(1 - (1-\alpha)^r)} \log \left(\frac{1}{1-\alpha} \right) + \frac{c(1 + (1-\alpha)^r)}{2(1 - (1-\alpha)^r)} \end{aligned} \quad (23d)$$

$$= \frac{1}{\lambda r} + \frac{1}{2\lambda} \log \left(\frac{1}{1-\alpha} \right) + c + \frac{c(1 + (1-\alpha)^r)}{2(1 - (1-\alpha)^r)}. \quad (23e)$$

To obtain (23a), we use the limit in (19) as an approximate and substitute it back to Theorem 1. To simplify the expression we further denote $\beta = 1 - \alpha$ to complete the proof.

Proof of Corollary 2: We first prove by contradiction that the optimal β doesn't fall into strict quorum region. For $w+r > n$, taking the derivative of the approximation 1) in corollary 1 gives

$$\frac{d\Delta}{d\beta} = \frac{\beta^{r-1}(\lambda cr - r \log \beta)}{\lambda}. \quad (24)$$

Thus we have the optimal $\beta^* = e^{\lambda c}$ by setting (24) to zero. Since λ and c are positive, $e^{\lambda c} > 1$ contradicts $\beta \in (0, 1)$. For non-strict quorum $w+r \leq n$, we let the derivative of the approximation 2) to be zero, i.e.,

$$\frac{d\Delta}{d\beta} = \frac{\beta^{2r} - 2(\lambda cr + 1)\beta^r + 1}{\lambda\beta(\beta^r - 1)} = 0 \quad (25)$$

Since $\beta \in (0, 1)$, it is equivalent that

$$\beta^{2r} - 2(\lambda cr + 1)\beta^r + 1 = 0. \quad (26)$$

We define $\omega = \beta^r$, and the solution to (26) is given by

$$\omega^* = (\lambda cr + 1) - \sqrt{(\lambda cr + 1)^2 - 1}. \quad (27)$$