ELSEVIER

Information Processing Letters 50 (1994) 69–73

Information Processing Letters

Message complexity of hierarchical quorum consensus algorithm

Her-Kun Chang, Shyan-Ming Yuan *

Department of Computer and Information Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 30050, Taiwan

(Communicated by W.M. Turski; received 20 September 1993; revised 10 January 1994)

Abstract

The hierarchical quorum consensus (HQC) algorithm, which is a generalization of the standard quorum consensus algorithm for managing replicated data, can reduce the quorum size to $N^{0.63}$. This paper analyzes the message complexity of HQC. Moreover, an asymptotic analysis on the ratio of the message complexity to the quorum size is presented. It is shown that the ratio converges to a constant.

Key words: Analysis of algorithms; Replicated data; Hierarchical quorum consensus; Quorum size; Message complexity

1. Introduction

In a distributed system, copies of the same data are kept in different sites to increase the availability of data. To maintain the mutual consistency among replicated copies, a replica control algorithm is required to synchronize read and write operations such that the multiple copies of the same data behave like a single copy. A survey of algorithms for replica control can be found in [2].

Quorum consensus algorithms [1,3,5] are well known for replica control. With quorum consensus, a read operation is allowed to be performed if it can get permissions from a read quorum of r sites. On the other hand, a write operation must get permissions from a write quorum of w sites. To ensure consistency, r and w must satisfy the following constraints:

$$r + w > N, \tag{1}$$

$$2w > N, \tag{2}$$

where N is the number of sites in the system.

Condition (1) ensures that a read operation can access a most recently updated copy (i.e. a copy updated by the last write operation). Version numbers can be used to determine which copy is most recently updated. Condition (2) ensures that the most recently updated copy has the largest version number (see [3] for details).

A major problem with quorum consensus algorithms is that the write quorum size is at least $\lceil (N+1)/2 \rceil$. The hierarchical quorum consensus (HQC) [4] algorithm, which logically organizes the sites in a system into a multilevel hierarchy, can reduce the quorum size to $N^{0.63}$. Quorum size was used to evaluate quorum consensus algorithms based on the assumption that message

* Corresponding author. Email: smyuan@tiger.cis.nctu.edu.tw.

SSDI 0020-0190(94)00012-N

complexity is proportional to quorum size. However, the communication cost of a distributed algorithm is usually estimated by the message complexity, which is defined to be the average number of exchanged messages per operation. In this paper, the message complexity of HQC is analyzed. Also, an asymptotic analysis on the ratio of the message complexity to the quorum size is presented. It is shown that the ratio converges to a constant γ. The presented analysis shows that, for HQC, the message complexity is proportional to the quorum size in most practical cases.

The remainder of this paper is organized as follows. The next section describes the model and notation. Section 3 reviews HQC. In Section 4, the message complexity of HQC is analyzed and an asymptotic analysis on the ratio of the message complexity to the quorum size is presented. Some concluding remarks are given in the final section.

2. Model and notation

A distributed system consists of a set of sites connected by a computer network. There is no shared memory between any pair of sites. The sites can communicate by exchanging messages only. In this paper, we assume that each site stores a replicated copy of the replicated data. The sites can be unreliable. When a site fails, the copy at the site becomes unavailable.

The availability of a site is the probability that the site is operational at any time instant. The *availability* of HQC is the probability that at least one hierarchical quorum can be formed. *Message complexity* is defined to be the expected number of exchanged messages per operation. *Quorum size* is defined to be the average number of sites that form a quorum.

In this paper, we use the following notation:

- p: availability of a single site, $\frac{1}{2} < p < 1$,
- A_i: availability of an i level hierarchy, $i \geqslant 0$,
- C_i: quorum size of an i level hierarchy, $i \geqslant 0$,
- M_i: message complexity of an i level hierarchy, $i \geqslant 0$,
- R_i: M_i/C_i, $i \geqslant 0$.

For HQC, it can be shown that if $p \leqslant \frac{1}{2}$, then $A_i \leqslant A_{i-1}$, for all $i \geqslant 1$. So we consider only the case that $\frac{1}{2} < p < 1$.

3. Hierarchical quorums

HQC is based on logically organizing sites into a multilevel hierarchy of depth m with root at level m. The physical sites are represented as leaves of the hierarchy at level 0, while the higher level nodes correspond to logical groups. A node at level $i+1$, $0 \leqslant i \leqslant m-1$, is viewed as a logical group which in turn consists of l_i subgroups at level i. The read quorum, r_i, and write quorum, w_i, at level i, $0 \leqslant i \leqslant m-1$, are defined as the number of subgroups that must be collected for read and write operations by a level $i+1$ group, respectively. HQC requires read and write operations to (recursively) assemble quorums of r_{m-1} and w_{m-1} level $m-1$ subgroups, respectively, such that

(a) $r_i + w_i > l_i$, and

(b) $2w_i > l_i$,

for all levels i, $0 \leqslant i \leqslant m-1$.

Fig. 1 shows some examples of hierarchical (read/write) quorums for a 2 level hierarchy, wherein $l_0 = l_1 = 3$ and $r_0 = r_1 = w_0 = w_1 = 2$.

It was shown that the minimum quorum size required for a write operation is $N^{0.63}$ which can be achieved in systems having $N = 3^m * 5^b$ sites where $b = 0$ *or* 1 [4]. In this paper, we discuss only ternary hierarchies, in which every non-leaf (logical) node consists of three nodes in the next lower level and the (read/write) quorums in all levels are 2. For a ternary hierarchy of depth m, the number of sites in the system is $N = 3^m$ and the size of quorums is $2^m + N^{0.63}$.

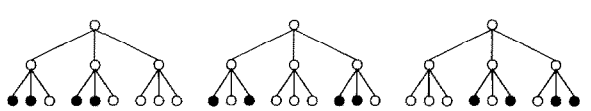

Fig. 1. Examples of hierarchical quorums: dark nodes form a quorum in each case.

For $1 \leqslant i \leqslant m$, A_i can be computed from the following recurrence:

$$A_i = A_{i-1}^3 + 3A_{i-1}^2(1 - A_{i-1})$$
$$= 3A_{i-1}^2 - 2A_{i-1}^3, \quad (3)$$

where $A_0 = p$ and the availability of the hierarchy is A_m.

The quorum size of an i level hierarchy, C_i, $i \geqslant 0$, is

$$C_i = 2^i. \quad (4)$$

4. Analysis of HQC

In this section, the message complexity of HQC is analyzed. An asymptotic analysis on the ratio of the message complexity to the quorum size, say R_i, is presented. It is shown that the sequence $\{R_i\}$ is convergent and a way to estimate the limit of $\{R_i\}$ is given.

4.1. Message complexity of HQC

For an m level ternary hierarchy, with root at level m and leaves (physical sites) at level 0, M_i is the expected number of messages required to form a quorum at level i, $0 \leqslant i \leqslant m$. Consider the construction of a quorum at level i, $1 \leqslant i \leqslant m$,

1. if the first two children (in some predefined order) are available – only $2M_{i-1}$ messages are required;
2. if the first two children are unavailable – since it is impossible to form a quorum, only $2M_{i-1}$ messages are required;
3. otherwise – $3M_{i-1}$ messages are required despite whether the quorum can be formed or not.

Thus,

$$M_i = \left(A_{i-1}^2 + (1 - A_{i-1})^2\right)2M_{i-1}$$
$$+ (2A_{i-1}(1 - A_{i-1}))3M_{i-1}$$
$$= 2\left(1 + A_{i-1} - A_{i-1}^2\right)M_{i-1}, \quad (5)$$

where $M_0 = 1$ and M_m is the message complexity of the m level hierarchy.

4.2. Asymptotic analysis of R_i for HQC

Lemma 1. *For HQC, $\frac{1}{2} < p < 1$, A_i has the following properties:*

(1) $1 - A_{i+m} \leqslant (1 + p - 2p^2)^m(1 - A_i)$,

(2) $(1 + (1 - A_i))(1 + (1 - A_{i+1}))$
$\cdots (1 + (1 - A_{i+m})) < e^{(1-A_i)/p(2p-1)}$.

where $i \geqslant 0$ and $m \geqslant 0$.

Proof. The proof is shown in the Appendix.

Lemma 2. *For HQC, $R_i \geqslant R_{i-1}$, for all $i \geqslant 1$.*

Proof.

$$R_i = M_i/C_i$$
$$= \frac{2\left(1 + A_{i-1} - A_{i-1}^2\right)}{2} \frac{M_{i-1}}{C_{i-1}}$$
$$= \left(1 + A_{i-1}(1 - A_{i-1})\right)R_{i-1}$$
$$> R_{i-1}. \quad \square$$

Lemma 3. *For HQC, $\frac{1}{2} < p < 1$,*

$$R_{i+m} < e^{(1-A_i)/p(2p-1)}R_i,$$

where $i \geqslant 0$ and $m \geqslant 1$.

Proof. For $i \geqslant 0$,

$$R_{i+1} = M_{i+1}/C_{i+1}$$
$$= \frac{2\left(1 + A_i - A_i^2\right)}{2} \frac{M_i}{C_i}$$
$$= \left(1 + A_i(1 - A_i)\right)R_i$$
$$< (1 + (1 - A_i))R_i. \quad (6)$$

By iteration,

$$R_{i+m} < \{(1 + (1 - A_i))(1 + (1 - A_{i+1}))$$
$$\cdots (1 + (1 - A_{i+m-1}))\}R_i$$

According to Lemma 1,

$$R_{i+m} < e^{(1-A_i)/p(2p-1)}R_i. \quad \square$$

Theorem 4. *For HQC, $\frac{1}{2} < p < 1$, the sequence $\{R_i\}$ is convergent.*

Proof. Lemma 2 shows that $\{R_i\}$ is monotonically increasing. By Lemma 3, $R_i < e^{(1-p)/p(2p-1)}$, i.e., $\{R_i\}$ is bounded. Thus, $\{R_i\}$ is convergent to a constant γ, where $\gamma < e^{(1-p)/p(2p-1)}$.

Theorem 4 shows that, for HQC, the message complexity is proportional to the quorum size if the system is sufficiently large.

4.3. Estimation of the limit of $\{R_i\}$ for HQC

It is shown by Lemma 3 that, for HQC, $\frac{1}{2} < p < 1$,

$$R_{i+m} < e^{(1-A_i)/p(2p-1)} R_i$$

where $i \geqslant 0$ and $m \geqslant 1$.

For any $\varepsilon > 0$, if

$$e^{(1-A_i)/p(2p-1)} < (1+\varepsilon)$$

then

$$R_{i+m} < (1+\varepsilon) R_i.$$

Since

$$e^{(1-A_i)/p(2p-1)} < (1+\varepsilon)$$

if

$$A_i > 1 - p(2p-1)\ln(1+\varepsilon). \tag{7}$$

Thus, for any $\varepsilon > 0$, we can find an i satisfying condition (7) such that $R_{i+m} < (1+\varepsilon)R_i$, for any $m \geqslant 1$. Consequently, the limit of $\{R_i\}$, γ, is smaller than $(1+\varepsilon)R_i$. That is, γ is in the interval $(R_i, (1+\varepsilon)R_i)$.

Table 1 illustrates the smallest value of i that satisfies condition (7) for several combinations of p and ε. Several notable observations from Table 1 are:

Table 1
The smallest value of i that satisfies condition (5)

p	ε			
	10^{-2}	10^{-3}	10^{-4}	10^{-5}
0.6	7	7	8	8
0.7	5	5	6	6
0.8	3	4	4	5
0.9	2	3	3	4

- In most practical applications, $p \geqslant 0.9$ (most current workstations can achieve this availability) and $i \geqslant 2$, the ratio of the message complexity to the quorum size is almost a constant.
- Even the site availability is low, e.g. $p = 0.6$, the ratio converges very fast ($i \geqslant 7$ for $p = 0.6$).

5. Conclusions

This paper analyzes the message complexity of the hierarchical quorum consensus algorithm. Also, an asymptotic analysis on the ratio of the message complexity to the quorum size is presented. It is shown that the ratio converges to a constant γ, where $\gamma < e^{(1-p)/p(2p-1)}$. Finally, a way to estimate the value of γ is given. The result shows that, in most practical cases ($p \geqslant 0.9$ and $i \geqslant 2$), the ratio is almost a constant. Even for some impractical cases, the ratio converges very fast. Quorum size was used to appraise the performance of quorum consensus algorithms based on the assumption that the message complexity is proportional to the quorum size. However, the assumption was not proved in previous works. The presented analysis shows that the assumption is valid in most practical cases.

6. Appendix

Lemma 5. *If* $0 < x_i < 1$, *for all* i, *then*

$$(1+x_1)(1+x_2)\cdots(1+x_n) < e^{\sum x_i}.$$

Proof. It is known that

$$\ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots.$$

Hence, for any $0 < z < 1$, $\ln(1+z) < z$. Then, if $0 < x_i < 1$, for all i, we have

$$\ln((1+x_1)(1+x_2)\cdots(1+x_n)) < \sum x_i$$

That is,

$$(1+x_1)(1+x_2)\cdots(1+x_n) < e^{\sum x_i}. \qquad \square$$

Proof of Lemma 1. For $i \geqslant 0$, by (3),

$$\begin{aligned}1 - A_{i+1} &= 1 - 3A_i^2 + 2A_i^3 \\ &= (1 - A_i^2) - 2A_i^2(1 - A_i) \\ &= (1 + (1 - 2A_i)A_i)(1 - A_i) \\ &\leqslant (1 + (1 - 2p)A_i)(1 - A_i) \\ &\leqslant (1 + (1 - 2p)p)(1 - A_i) \\ &= (1 + p - 2p^2)(1 - A_i).\end{aligned}$$

By iteration, for $i \geqslant 0$ and $m \geqslant 0$,

$$1 - A_{i+m} \leqslant (1 + p - 2p^2)^m (1 - A_i). \tag{8}$$

Next, according to (8),

$$\begin{aligned}&(1 - A_i) + (1 - A_{i+1}) + \cdots + (1 - A_{i+m}) \\ &\quad \leqslant (1 - A_i)\{1 + (1 + p - 2p^2) \\ &\qquad + \cdots + (1 + p - 2p^2)^m\} \\ &\quad < (1 - A_i)\frac{1}{1 - (1 + p - 2p^2)} \\ &\quad = \frac{1 - A_i}{p(2p - 1)}.\end{aligned}$$

Thus, by Lemma 5,

$$\begin{aligned}&(1 + (1 - A_i))(1 + (1 - A_{i+1})) \\ &\quad \cdots (1 + (1 - A_{i+m})) < e^{(1 - A_i)/p(2p-1)}.\end{aligned}$$

□

7. References

[1] M. Ahamad and M.H. Ammar, Performance characterization of quorum-consensus algorithms for replicated data, *IEEE Trans. Software Engrg.* **15** (4) (1989) 492–496.

[2] S.B. Davidson, H. Garcia-Molina and D. Skeen, Consistency in partitioned networks, *ACM Comput. Surveys* **17** (3) (1985) 341–370.

[3] D.K. Gifford, Weighted voting for replicated data, in: *Proc. 7th ACM Symp. on Operating System Principles* (1979) 150–162.

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

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