A HIGH AVAILABILITY CLUSTER-BASED REPLICA CONTROL PROTOCOL IN DATA GRID

[1]Zulaile Mabni, [3]Rohaya Latip, [2]Hamidah Ibrahim & [1]Azizol Abdullah
[1&2]Faculty of Computer Science and Information Technology and
[3]Institute for Mathematical Research (INSPEM)
Universiti Putra Malaysia, Malaysia

zulaile@tmsk.uitm.edu.my; rohayalt@upm.edu.my;
hamidah.ibrahim@upm.edu.my; azizol@upm.edu.my

ABSTRACT

Data replication is widely used to provide high data availability, and increase the performance of the distributed systems. Many replica control protocols have been proposed in distributed and grid environments that achieved both high performance and availability. However, the previously proposed protocols still require a bigger number of replicas for read and write operations which are not suitable for a large scale system such as data grid. In this paper, a new replica control protocol called Clustering-based Hybrid (CBH) has been proposed for managing the data in grid environments. We analyzed the communication cost and data availability for the operations and compared CBH protocol with recently proposed replica control protocols called Dynamic Hybrid (DH) protocol and Diagonal Replication in 2D Mesh (DR2M) protocol. To evaluate CBH protocol, a simulation model was implemented using Java. Our results show that for the read operations, CBH protocol improves the performance of communication cost and data availability compared to the DH and DR2M protocols.

Key words: Data replication, grid computing, data availability, communication cost.

INTRODUCTION

Grid computing is a distributed network computing system that enables large scale resource sharing between machines distributed across many organizations

Received: 19 March 2016 Acepted: 7 December 2016

and over a wide area network (Foster et al., 2001; Krauter et al., 2002). In grid computing, data grid provides a scalable infrastructure to manage huge amounts of data and support data intensive applications (Chervenak et al., 2000; Abdullah et al., 2004, Yusof et al., 2012). Thus, managing the large network and widely distributed data in the data grid is a challenging problem. To address the challenge, various methods have been proposed in the literature. Data replication is one of the widely used methods to improve data availability and enhance the performance of the distributed database systems (Lamehamedi et al., 2002; Mabni & Latip, 2011). However, some issues arise in managing the replication of data. One of the issues is data availability (Lamehamedi et al., 2003; Latip et al., 2014), because data is geographically distributed over large networks. Another issue is communication cost, where cost can become expensive if the number of read and write operations is high (Choi & Youn, 2012; Latip et al., 2009). The communication cost is calculated based on the number of replicas that need to be accessed. Thus, to obtain low communication cost, the number of replicas need to be as small as possible (Koch, 1993).

In a replicated distributed system, multiple copies or replicas of an object are produced and stored at many sites. The operations that are allowed on the replicated data are read and write operations. A read quorum (RQ) or write quorum (WQ) is defined as a set of copies that is sufficient to execute the read or write operation. In order to maintain a consistent state among the replicas, these "multiple copies or replicas of an object must appear as a single logical object to the transaction which is known as one-copy equivalence" (Bernstein & Goodman, 1984). Thus, to ensure one-copy equivalence, the quorum selected must satisfy the quorum intersection property. The property states that "for any two operations *o[x]* and *o'[x]* on an object *x*, where at least one of them is a write, the quorums must have a non-empty intersection" (Gifford, 1979). Therefore, the basic property for any replica control protocol is to guarantee non-empty intersection between read and write quorums in order to maintain the consistency of the replicated data.

In the literature, many replica control protocols have been proposed in distributed and grid environments which achieved both high performance and availability. However, the previously proposed protocols still require a bigger number of replicas for read and write operations which are not suitable for a large scale system. In this paper, we propose a new replica control protocol called Clustering-based Hybrid (CBH) protocol for the grid environment. The proposed protocol employs a hybrid replication strategy by combining the advantages of two common replica control protocols to improve the performance of earlier protocols. The proposed protocol groups nodes into clusters and organizes these clusters into a tree structure which enables the protocol to minimize the number of replicas for read or write operations. Thus, CBH provides low communication cost as well as high availability.

RELATED WORKS

Data replications have been an active research area in distributed and grid environments. This section describes the previously proposed replica control protocols where the number of replicas involved in executing the read and write operations are different from each other.

Primary Copy Protocol

Primary Copy (PC) algorithm is a simple algorithm that selects one copy of a data object as the primary copy (Stonebraker, 1979; Ahamad et al., 1992; Zhou & Holmes, 1999). In this protocol, each node knows which other nodes it can communicate with by the up-list of nodes. By definition, the node that has the lowest order in the up-list is selected as the primary copy of the data object. The primary copy will maintain the consistency of the object. Any other node is called a non-primary copy. A read operation is executed only at the primary copy while the write operation updates the primary copy and then propagates out to all other nodes that maintain the non-primary copies.

Read-One Write-All

Read-One Write-All (ROWA) protocol is another simple protocol for managing replicated data (Bernstein & Goodman, 1984). In this protocol, a read operation is required to access any single replica. On the other hand, to perform write operation, all *n* replicas need to be accessed. Thus, data consistency is guaranteed in the ROWA protocol.

Voting Protocol

Voting protocol (VT) was first proposed by Thomas (1979). This protocol was later enhanced by Garcia-Molina & Barbara (1985) where, each replica is assigned a certain number of votes *v*. Every transaction has to collect a read quorum of *r* votes to read a replica, and a write quorum of *w* votes to write the replica. A quorum must satisfy the following two constraints:

i) $r + w > v$

ii) $w > v / 2$

The first constraint ensures that there is a non-empty intersection between every read quorum and every write quorum. The second condition ensures that there is a non-empty intersection between two write quorums. The communication cost of this protocol depends on the quorum size. The bigger the size of the read or write quorum, the higher the communication cost. In this protocol, read operation needs to access several replicas which make the communication cost higher than the ROWA protocol. Meanwhile, for the write operation, this protocol does not need to access all replicas such as in the ROWA protocol, thus increasing its fault-tolerance.

Tree Quorum Protocol

In the Tree Quorum (TQ) protocol (Agrawal & El Abbadi, 1990), replicas are organized in a logical tree structure. Figure 1 shows the diagram of thirteen copies in a tree quorum structure of *height = 2* and degree of node *D = 3*, where every copy represents a replica. In this protocol, a read quorum needs to access only the root replica. If the root is inaccessible, then a read quorum needs to the access majority replicas of its children. Furthermore, for every inaccessible replica, the majority replicas of its children are accessed, and so on and so forth. The examples of valid read quorums of Figure 1 are {1} when the root replica is accessible, and {2,3} when the root replica is inaccessible.

Meanwhile, a write quorum consists of the root, and any majority replicas of the root's children and any majority replicas of their children, and so forth until the leaves are reached. In Figure 1, the examples of valid write quorums are {1,2,3,5,6,8,9}, and {1,3,4,9,10,11,12}.

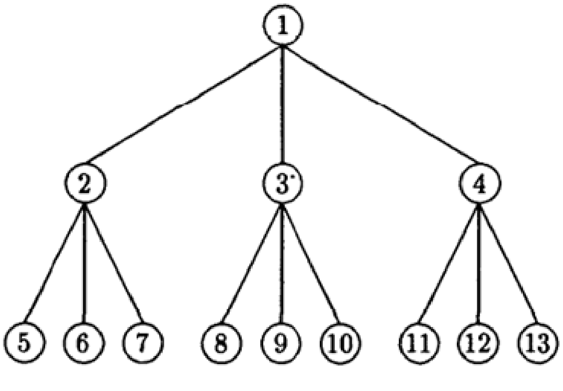

Figure 1. A tree organization of 13 copies of data objects (Agrawal, 1990).

Diagonal Replication on 2D Mesh Protocol

In Diagonal Replication on 2D Mesh structure (DR2M) nodes are organized in a two-dimensional 2D mesh structure (Latip et al., 2008; Latip et al., 2009). Figure 2 illustrates the network of 81 nodes which are divided into four quorums. The nodes are logically grouped by 5 x 5 in each quorum. The nodes in a quorum intersect with the nodes in other quorums to ensure that each quorum can communicate with another quorum. The data is replicated to only the middle node of the diagonal site in each quorum. The replicated data is assigned with vote one. This protocol employs the voting technique where, the write quorum q_w can be formed by any majority of the replicas and the read quorum q_r by a half of the replicas. To ensure that consistency is maintained, $q_w + q_r$ must be greater than the total number of votes assigned to all replicated data (Latip et al., 2008). The read and write quorum sizes of Figure 2 are 2 and 3 respectively.

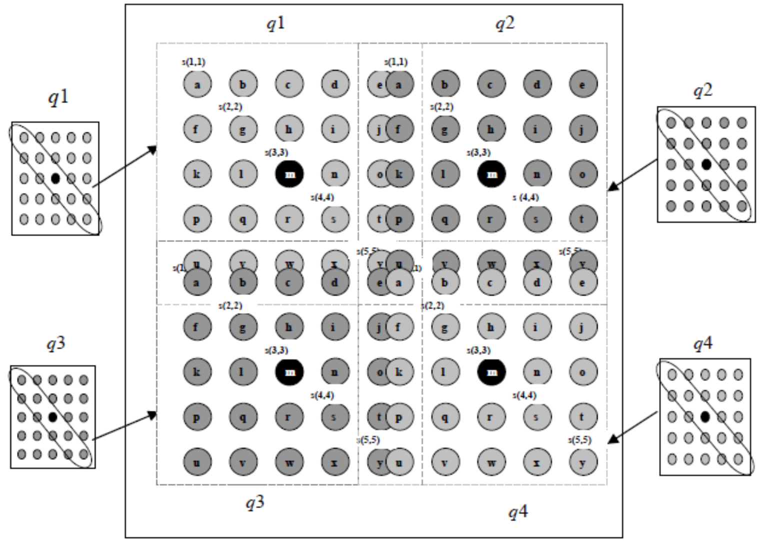

Figure 2. DR2M with 81 nodes; each of the nodes has a data file a, b,…, and y respectively (Latip et al., 2009).

Dynamic Hybrid Protocol

Dynamic Hybrid (DH) protocol is a hybrid replica control protocol that has been proposed recently (Choi & Youn, 2012). In this protocol, the overall topology combined the grid and tree structure where the tree height, number of descendants and grid depth can be adjusted. Figure 3 illustrates the network of DH protocol with 31 replicas in (3,3,2) topology, where the three arguments represent the height *h*, number of descendants *s* and grid depth *g* respectively. In the tree structure of height *h*, the read operation needs to access only the root replica. However, if the root is inaccessible, then the *s* descendants of the root replica have to be accessed. The descendants of the root serve as the new root replica of the sub-tree. The process is repeated until level $h - 1$ is reached. Furthermore, in the grid network of depth *g*, read operation reads *s* replicas or goes to the next level if one of the replicas is inaccessible. Thus, if the root replica is accessible, the read cost is only 1. The examples of valid read quorums of Figure 3 are {R0}, {R1, R2, R3} and {R2, R3, R4, R5, R6}.

Meanwhile, the write operation reads the root replica; one replica of the root's descendants, one replica of these previously selected replicas' descendants and so forth until the leaves are reached. Furthermore, in the grid network of depth *g*, write operation reads only one replica in each level down to the last level. In Figure 3, the examples of valid write quorums are {R0, R1, R4, R13, R22} and {R0, R2, R7, R16, R27}.

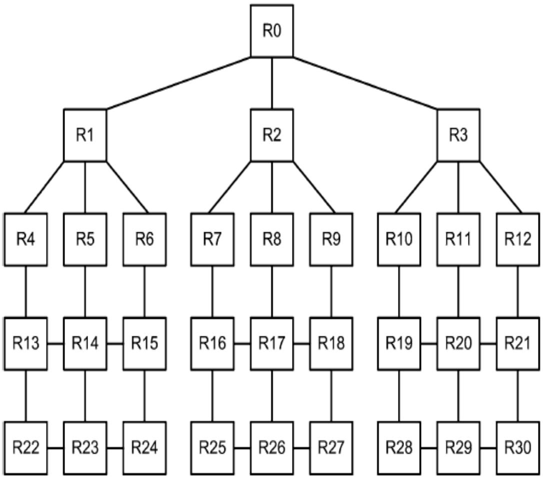

Figure 3. A network for Dynamic Hybrid protocol of 31 replicas in (3,3,2) topology (Choi & Youn, 2012).

Review of Replica Control Protocols

In this section, we have described the previously proposed replica control protocols in distributed and grid environments. The Primary Copy protocol is easy to implement and has a very low read communication cost, however, if the node that maintains a primary copy is not accessible, then a write operation cannot be executed. ROWA produces a very low read communication cost and very high read availability like the Primary Copy protocol since only one replica is required to be accessed. However, ROWA has very high communication cost for the write operation since all replicas must be updated simultaneously. Furthermore, a write operation cannot be performed in case of any node failure. The concept of quorum used in Voting protocol has improved the performance of the protocol compared to ROWA where only the majority of nodes in the network need to be accessed in order to ensure consistency. However, the communication cost for the write operation is still expensive since a write quorum of *w* votes must be larger than the majority votes (Mat Deris et al., 2004). The Tree Quorum protocol allows very low read communication cost since read operation requires only one replica such as in the ROWA protocol. However, as the level of the tree increases, the number of replicas increases rapidly, thus increasing the communication cost. To address the limitation of the Tree Quorum protocol, DR2M protocol has minimized the number of replicas where the primary database is replicated only at the middle of the diagonal site. The number of replicas that need to be accessed for the read and write operations is small; thus it has low read and write communication cost. However, as the network size grows larger, the replicas will be further apart and decrease the performance of the system. A recently proposed protocol named Dynamic Hybrid combined the advantages of the Tree and the Grid protocols to allow low operation cost and high availability. However, as the network size grew, a large number of replicas still needed to be accessed to maintain data consistency and therefore, degraded the performance of the system.

Most of the mentioned replica control protocols perform well in small size systems where the number of replicas is small. However, in a larger system, these replica control protocols require a larger number of replicas to be accessed in order to maintain data consistency and degrade the performance of the system (Abawajy & Mat Deris, 2014). Thus, these protocols are not suitable for a large scale system such as data grid. Therefore, we propose a new quorum-based replica control protocol called Clustering-based Hybrid (CBH) protocol for managing replicated data in a large scale system. CBH protocol minimizes the number of replicas for read or write operations as well as maintains data consistency in a large scale system such as data grid.

CLUSTERING-BASED HYBRID PROTOCOL

In this section, we present the system model and the algorithm for the proposed protocol called Clustering-Based Hybrid (CBH) protocol.

System Model

The system consists of N sites that communicate with each other by exchanging messages through a communication link. We assumed that sites fail independently and communication links do not fail to deliver messages. It is assumed that access requests to the physical replicas are performed by executing transactions which are partially ordered read and write operations. In the CBH protocol, the N sites in the network are logically grouped into several nonintersecting groups. The N sites are divided into $\sqrt{N}$ disjoint groups with each group having approximately $\sqrt{N}$ sites (Madhuram & Kumar, 1994; Mabni et al., 2014; Latip et al., 2014). Each group is called a cluster. These clusters are logically organized as a tree of height, h and descendants, s. We defined the nodes in the tree to be a sequence of clusters C_0, C_1,… C_i, C_{i+1}, … C_n. We assumed that the nodes in each cluster are logically organized into two dimensional grid structures. For example, if the CBH protocol consists of 81 nodes, it will be divided into 9 clusters with 9 nodes in each cluster. The nodes in each cluster will be logically organized in the form of 3 x 3 grid. In Figure 4, an example of a ternary tree of *height = 2* with 81 nodes is presented. Each cluster designates the middle node of the cluster as the cluster head which is colored in black in Figure 4 and has the replica or primary copy of the data object. The center of the cluster is selected because it is the shortest path to get a copy of the data from most of the directions in the cluster.

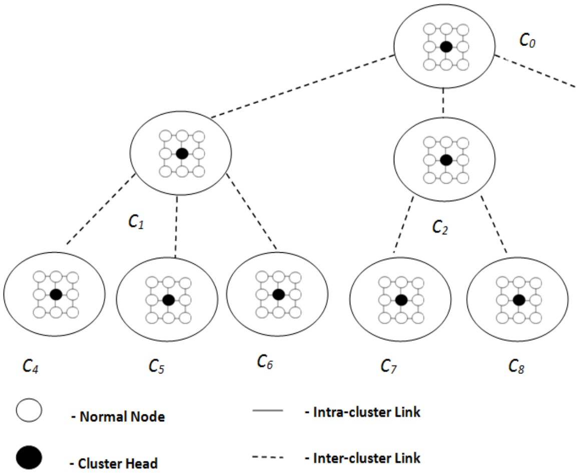

Figure 4. System model of CBH in a ternary tree of height = 2 with 81 nodes.

Proposed Algorithm

Here, we describe the hybrid algorithm of the CBH protocol, where it combines the advantages of two common replica control protocols, namely the Tree Quorum (TQ) protocol and the Primary Copy (PC) protocol. The hybrid algorithm logically groups the nodes into a tree structure. Figure 4 shows a system with 81 nodes for which we use the TQ protocol on top of the PC protocol as the replication strategy. The system consists of 81 nodes where nine clusters called C_0, $C_{1,}$ …, C_8 are "logical replicas" as shown in Figure 4. The logical replica C_0 serves as the root cluster, whereas the logical replicas C_1,…, C_8 are its descendants. Each logical replica contains a cluster of physical nodes with one middle node called "physical replica" which has the replica or the primary copy of the data object. The physical replica in the root cluster C_0 is called root replica. Thus, for a system with N nodes, there will be $\sqrt{N}$ clusters, and $\sqrt{N}$ replicas. We assume that every physical replica is assigned exactly one vote. To illustrate the algorithm, the replication strategy involves two strategies: "Local Replication", where the PC protocol is used for the replication strategy for managing the physical replica within a cluster and "Global Replication", where the TQ protocol is used as the replication strategy for managing the logical replicas between clusters.

Read Operation

In the CBH protocol, for global replication, a read operation is based on the TQ protocol, where the root replica C_0 is accessed if it is accessible. However, if the root replica C_0 is inaccessible then a majority of physical replicas of its children are added as members of this quorum. Furthermore, for every inaccessible physical replica, a majority of physical replicas of its children are added as members, and so forth. On the other hand, for local replication, a read operation is based on the PC protocol, where a logical replica can be read if the physical replica that it contains can be accessed. This means that for reading a logical replica, the precondition is a read quorum of $RQ = 1$ if its contained physical replica is accessible for read operation. Thus, in Figure 4, by employing the TQ Protocol, the minimal read cost is 1 if the root replica is accessible. However, if the root replica is inaccessible, then the read cost is 2 since the majority of the physical replicas of its children have to be accessed. The examples of valid read quorums of Figure 4 are $\{C_0\}$ if the root replica is available and $\{C_1, C_2\}$ if the root replica is not available.

Write Operation

In the CBH protocol, for global replication, a write operation is based on the TQ protocol, where the root replica C_0 and any majority of the physical replicas of the root's children, and any majority of the physical replicas of their children, and so forth are accessed until the leaves are reached. As for local replication, a write operation is based on the PC protocol, where a logical replica is accessible if a write operation can be performed on its physical replica. This means that for writing a logical replica, the precondition is a write quorum of $WQ = 1$ if its contained physical replica is accessible for write operation. Therefore, in Figure 4, by employing the TQ protocol, we obtain a write cost of 7. An example of valid write quorum of Figure 4 is $\{C_0, C_1, C_2, C_4, C_5, C_7, C_8\}$.

Correctness of CBH Algorithm

A replica control protocol is said to achieve one-copy equivalence if any read quorum has a non-empty intersection with any write quorum. Here, we demonstrate that the CBH protocol guarantees a non-empty intersection.

> ***Theorem***: The CBH protocol guarantees the intersection of read and write quorums.

Local replication

Proof: In any cluster, there is only one physical replica or primary copy that maintains the consistency of the object. Thus, the quorum intersection property within a cluster is guaranteed.

Global replication

In (Agrawal & El Abbadi, 1990), the Tree Quorum protocol was proven to satisfy the intersection property. Since the Tree Quorum protocol was used in the global replication, the proof is as follows:

> *Proof*: The proof is by induction on the height of the trees.

Basis: The theorem holds for a tree of height zero, since there is only one physical replica in the tree.

> *Induction Hypothesis*: Assume that the theorem holds for a tree of height h.

> *Induction Step*: Consider a tree of height $h + 1$. The read quorum (RQ) and write quorum (WQ) for the CBH protocol are as follows:

RQ = {Root Replica} or {Majority of physical replicas of sub trees of height h}.

WQ = {Root Replica} and {Majority of physical replicas of sub trees of height h}.

In the CBH protocol, the intersection property is guaranteed since any write quorum must access the root replica whenever the root replica is accessible. On the other hand, if the root replica is inaccessible, the read quorum must access a majority of physical replicas for sub trees of height *h.* Therefore, it is guaranteed to have at least one sub tree in common with any write quorum. Since the sub trees are of height *h*, the induction hypothesis guarantees that read and write quorums will have a non-empty intersection.

Thus, by induction, the CBH protocol guarantees a non-empty intersection between the read and write quorums.

PERFORMANCE ANALYSIS AND COMPARISON

A simulation model was developed using Java to validate the CBH protocol. In this section, we evaluate and compare the performances which are communication costs and data availability of DR2M, DH and CBH protocols.

Communication Cost Analysis

The communication cost of an operation is directly proportional to the size of read and write quorum required to execute the operation. Thus, the bigger the number of replicas involved in the read or write operation, the higher the communication cost. Therefore, for the cost analysis, we represent the communication cost in terms of the number of replicas involved in the read or write operation.

Diagonal Replication on 2D Mesh Protocol

In DR2M (Latip et al., 2008; Latip et al., 2009), voting approach is used to assign a certain number of votes to every copy of the replicated data objects. The selected node in the diagonal sites is assigned vote one or zero. The communication cost for read and write operation is directly proportional to the size of the quorum. The DR2M communication cost for the read operation $C_{DR2M, R}$ is:

$$C_{DR2M,R} = \lfloor r/2 \rfloor \quad (1)$$

whereas, the communication cost for the write operation $C_{DR2M, W}$ is:

$$C_{DR2M,W} = \lfloor (r+1)/2 \rfloor \quad (2)$$

where r is the number of replicas in the whole network for executing read or write operations.

Dynamic Hybrid Protocol

The read operation of the DH protocol (Choi & Youn, 2012) needs to access only the root replica if the root replica is available. The minimum read cost $C_{DH, R}$ is:

$$C_{DH,R} = 1 \quad (3)$$

and the write cost $C_{DH, W}$ that depends on the value of h and g is:

$$C_{DH,W} = h + 1 + g \quad (4)$$

Clustering-Based Hybrid (CBH) Protocol

In the CBH protocol, the communication cost is estimated based on the TQ protocol as given in Chung (1994) , where h denotes the height of the tree, D is the degree of the logical replicas in the tree, and M is the majority of D where:

$$M = \left\lceil \frac{D+1}{2} \right\rceil \quad (5)$$

Therefore, for a tree of height h, the maximum quorum size is M^h and the communication cost for the read operation $C_{CBH, R}$ is in the range of $1 \leq C_{CBH, R} \leq M^h$. Meanwhile, the communication cost for the write operation $C_{CBH, W}$ is:

$$C_{CBH,W} = \sum \mathrm{M}^{\mathrm{i}} \quad (6)$$

where $i = 0,\ldots,h$.

Comparison of Communication Costs

For the read communication costs, both the CBH and the DH protocols have the same minimum read cost of 1. This is due to the fact that CBH and DH need to consult only the root replica if the root replica is accessible. On the other hand, for an example of 81 nodes, the DR2M protocol requires a higher read communication cost, which is 2.

Table 1 illustrates the write communication costs of the CBH, DH and DR2M protocols for an example system with a different total number of nodes, n = 81, 121, 225, and 289. Here, for a fair comparison, we assume that for the DH and the CBH protocols, the number of descendants is 3. For the write costs as illustrated in Table 1, it is apparent that DR2M has the lowest overall write communication cost. This is because in DR2M, the write quorum size is smaller than those of the other two protocols. The average write cost for CBH is 8.0 and for DH is 9.5. Thus, CBH has reduced the average write cost by up to 15.8% compared to DH.

Considering the read communication cost, in the best case, CBH and DH are better than DR2M. As for the write communication cost, CBH provides lower write costs than DH since the number of replicas required for the write operation is smaller.

Table 1

Comparison for the Write Communication Costs of the Protocols

Protocols	Number of nodes in the system			
	N = 81	N = 121	N = 225	N = 289
DR2M	3	3	3	3
DH	6	7	11	14
CBH	7	7	9	9

Data Availability Analysis

In this section, we analyze the read and write availability of the protocols. The availability of the protocol is defined as the probability of successfully forming a read and write quorum in that protocol. The read and write availability is determined by the probabilistic failure model where every replica is available independently with a probability *p*. Every replica is assumed to have the same availability *p* in estimating the availability of an operation.

Diagonal Replication on 2D Mesh Protocol

The DR2M (Latip et al., 2008; Latip et al., 2009) read availability $A_{DR2M,R}$ is represented as:

$$A_{DR2M,R} = \sum_{i=q_r}^{n} \binom{n}{i} \left(p^i(1-p)^{n-i}\right) \tag{7}$$

and write availability $A_{DR2M,W}$ is represented as:

$$A_{DR2M,W} = \sum_{i=q_w}^{n} \binom{n}{i} \left(p^i(1-p)^{n-i}\right) \tag{8}$$

In Equation 7 and Equation 8, n is the number of the column or row of the grid. For example, in Figure 2, the value of n is 5. p is the probability that a copy is available with a value between 0 and 0.9. The q_R and q_W are the number of quorums for the read and write operations, respectively.

Dynamic Hybrid protocol

Meanwhile, in the DH protocol (Choi & Youn, 2012), the overall availability is obtained using the availability of the tree and grid structure. The availability of the read operation of the grid structure is:

$$\wp_{\boldsymbol{read}}^{\boldsymbol{G(g)}} = \boldsymbol{p^s} + (\boldsymbol{1} - \boldsymbol{p^s}).\wp_{\boldsymbol{read}}^{\boldsymbol{G(g-1)}} \tag{9}$$

with $\wp_{\boldsymbol{read}}^{\boldsymbol{G(0)}} = \boldsymbol{p^s}$

Thus, the overall read availability of the DH protocol for a tree of height h is:

$$\wp_{\boldsymbol{read}}^{\boldsymbol{(l)}} = \boldsymbol{p} + (\boldsymbol{1} - \boldsymbol{p}).(\wp_{\boldsymbol{read}}^{\boldsymbol{(l-1)}})^{\boldsymbol{s}} \tag{10}$$

with

$$\wp_{\boldsymbol{read}}^{\boldsymbol{(0)}} = \wp_{\boldsymbol{read}}^{\boldsymbol{G(g)}} \ \boldsymbol{where} \ l = h - 1$$

On the other hand, the availability of the write operation of the grid structure is:

$$\wp_{\boldsymbol{write}}^{\boldsymbol{G(l)}} = p\left(\sum_{k=1}^{s} \binom{k}{s} p^k(1-\boldsymbol{p})^{s-k}\right).\wp_{\boldsymbol{write}}^{\boldsymbol{G(l-1)}} \tag{11}$$

where l = g -1

$$\wp_{\boldsymbol{write}}^{\boldsymbol{G(0)}} = \sum_{k=1}^{s} \binom{k}{s} \boldsymbol{p}^k(1-\boldsymbol{p})^{s-k}$$

Thus, the overall write availability of the DH protocol for tree of height *h* is:

$$\wp_{\boldsymbol{write}}^{(\boldsymbol{l})} = p\left(\sum_{k=1}^{s}\binom{k}{s}\,(\wp_{\boldsymbol{write}}^{(\boldsymbol{l-1})})^{k}\left(1-\wp_{\boldsymbol{write}}^{(\boldsymbol{l-1})}\right)^{s-k}\right) \tag{12}$$

with

$$\wp_{\boldsymbol{write}}^{(\boldsymbol{0})} = p\left(\sum_{k=1}^{s}\binom{k}{s}\,\boldsymbol{p}^{k}(1-\boldsymbol{p})^{s-k}\right).\ \wp_{\boldsymbol{write}}^{\boldsymbol{G(l)}}$$

where l = h -2

Clustering-Based Hybrid (CBH) protocol

The overall availability of the CBH protocol is obtained using the combination of the availability for the PC protocol and the TQ protocol. The availability for the read operation of the PC protocol is as given in Equation 13, where *p* is the probability of data file accessing between 0.1 and 0.9 and *i* is the increment of *n*.

$$A_{PC,\,R} = \sum_{i=1}^{n=1}\binom{n}{i} p^{i}\,(1 - p)^{n-i} \tag{13}$$

$$= 1 - (1 - p)^{n}$$

Thus, the overall availability for read operation of CBH protocol for a tree of height *h + 1* is:

$$A_{CBH,R_{h+1}} = \boldsymbol{A_{PC,R}} + \left(\boldsymbol{1} - \boldsymbol{A_{PC,R}}\right) \sum_{i=M}^{D}\binom{D}{i} A_{CBH,R_h}{}^{i}\,(1 - A_{CBH,R_h})^{\mathrm{D-i}} \tag{14}$$

with

$$A_{CBH,R_0} = \boldsymbol{A_{PC,R}}$$

The availability for the write operation of PC protocol is:

$$A_{PC,\,W} = \sum_{i=1}^{n=1}\binom{n}{i} p^{i}\,(1 - p)^{n-i} \tag{15}$$

$$= 1 - (1 - p)^{n}$$

Therefore, the overall availability of the write operation for the CBH protocol for a tree of height *h + 1* is:

$$A_{CBH,W_{h+1}} = A_{PC,\,w} \sum_{i=M}^{D} \binom{D}{i} A_{CBH,W_h}{}^{i} \, (1 - A_{CBH,W_h})^{\mathrm{D-i}}. \tag{16}$$

$$A_{CBH,W_0} = \boldsymbol{A_{PC,W}}$$

Comparison of Data Availabilities

Figure 5 shows the comparison of the read availability of the CBH, DH and DR2M protocols for 121 nodes. Here, we assume that the DH and the CBH protocols have descendants *s = 3*. The result in Figure 5 indicates that CBH has the highest read availability compared to the DR2M and the DH protocols. This is due to the fact that CBH needs only to access a small number of replicas for the read operations. The result shows that for read availability, the CBH protocol has an average of 10.9% higher compared to the DR2M protocol and 16.8% higher compared to the DH protocol for all probabilities of data accessing.

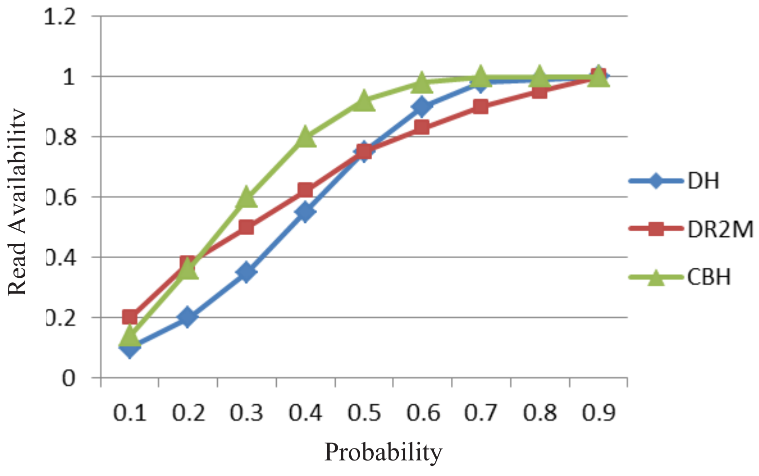

Figure 5. Comparison of read availability for 121 nodes.

The comparison for the write availability of the CBH, DH and DR2M protocols for 121 nodes is depicted in Figure 6. It shows that DR2M has the highest write availability than the DH and the CBH protocols. This is because of the fact that in DR2M, the write quorum size is small and the read and write operations execute at the same quorum. On the other hand, CBH has the lowest write availability as compared to the DH and the DR2M protocols.

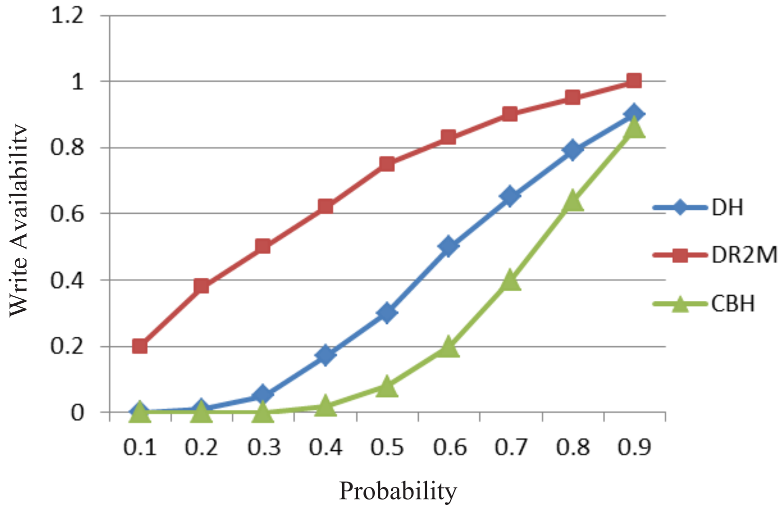

Figure 6. Comparison of write availability for 121 nodes.

CONCLUSION

A new replica control protocol named Clustering-Based Hybrid (CBH) protocol has been proposed in this paper for the management of replicated data in a large scale distributed system such as data grid. In the proposed protocol, the *N* sites in the network are logically grouped into several nonintersecting groups called clusters and organized in a tree structure. The CBH protocol employs a hybrid replication strategy by combining the advantages of the Primary Copy (PC) protocol and the Tree Quorum (TQ) protocol to improve the performance and availability of the protocol. In CBH, grouping the nodes into clusters and having only one replica in each cluster has resulted in a small number of replicas involved in performing the read and write operations. In comparison with the DR2M protocol and the DH protocol, the CBH protocol provides lower read communication cost while providing higher read availability which is suitable for large scale systems in grid environments.

ACKNOWLEDGMENT

We wish to acknowledge the support of the Malaysian Ministry of Higher Education (MOHE) under the Fundamental Research Grant Scheme (FRGS) No: 08-01-13-1220FR.

REFERENCES

Abawajy, J. H., & Mat Deris, M. (2014). Data replication approach with consistency guarantee for data grid. *IEEE Transaction on Computers, 63(12)*, 2975-2987.

Abdullah, A., Othman, M., Sulaiman, M. N., Ibrahim, H., & Othman, A. T. (2004). A simulation study of data discovery mechanism for scientific data grid environment. *Journal of Information and Communication Technology (JICT)*, 3 (1). 19-32.

Agrawal, D., & El Abbadi, A. (1990). The Tree Quorum protocol:An efficient approach for managing replicated data. *Proceedings of the 16th International Conference on Very Large Databases*, 243-254.

Ahamad, M., Ammar, M.H., & Cheung, S.Y. (1992). Replicated data management in distributed systems. *Readings in Distributed Computing Systems*, 572-591. doi:10.1.1.45.5283

Bernstein, P. A., & Goodman, N. (1984). An algorithm for concurrency control and recovery in replicated distributed database. *ACM Transaction Database Systems, 9*(4), 596-615. doi:10.1145/1994.2207

Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., & Tuecke, S. (2000). The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications, 23*(3), 187-200. doi: 10.1006/jnca.2000.0110

Choi, S. C., & Youn, H. Y. (2012). Dynamic hybrid replication effectively combining tree and grid topology. *The Journal of Supercomputing, 59*(3), 1289-1311. doi: 10.1007/s11227-010-0536-6

Chung, S. M. (1994). Enhanced tree quorum algorithm for replica control in distributed database systems. *Data and Knowledge Engineering, Elsevier, 12*(1), 63-81. doi:10.1016/0169-023X(94)90022-1

Foster, I., Kesselman, C., & Tuecke, S. (2001). The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of High Performance Computing Applications, 15*(3), 200-222. doi:10.1177/109434200101500302

Garcia-Molina, H., & Barbara, D. (1985). How to assign votes in a distributed system. *Journal of the ACM (JACM)*, 32(4), 841-860.

Gifford, D. K. (1979). Weighted voting for replicated data. *Proceedings of the 7th Symposium on Operating System Principles,* 150-162.

Koch, H. (1993). An efficient replication protocol exploiting logical tree structures. *The 23rd Annual International Symposium on Fault-Tolerant Computing*, 382-391.

Krauter, K., Buyya, R., & Maheswaran, M. (2002). A taxanomy and survey of grid resource management systems for distributed computing. *International Journal of Software Practice and Experience,, 32*(2), 135-164. doi:10.1002/spe.432

Lamehamedi, H., Shentu, Z., & Syzmanski, B. (2003). Simulation of dynamic data replication strategies in data grids. *Proceedings of the 17th International Symposium on Parallel and Distributed Processing, 1-10*.

Lamehamedi, H., Syzmanski, B., Shentu, Z., & Deelman, E. (2002). Data replication in grid environment. *Proceedings of the Fifth International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP'02)*, 378-383.

Latip, R., Ibrahim, H. , Othman, M., Sulaiman, M. N., & Abdullah, A. (2008). High availability with diagonal replication in 2D Mesh (DR2M) protocol for grid environment. *Journal of Computer and Information Science, 1*(2), 95-1005. doi: 10.5539/cis.v1n2p95

Latip, R., Ibrahim, H., Othman, M., Abdullah, A., & Sulaiman, M.N. (2009). Quorum-based data replication in grid environment. *International Journal of Computational Intelligence Systems (IJCIS), 2*(4), 386-397 doi:10.2991/ijcis.2009.2.4.7

Latip, R., Mabni, Z., Ibrahim, H., Abdullah, A., & Hussin, M. (2014). A clustering-based hybrid replica control protocol for high availability in grid environment. *Journal of Computer Science*, 10(12), 2442-2449.

Mabni, Z., & Latip, R. (2011). A comparative study on quorum-based replica control protocols for grid environment. In A. Abd Manaf et al. (Ed.), *Informatics Engineering and Information Science* (Vol. 253, pp. 364-377): Springer Berlin Heidelberg.

Mabni, Z., Latip, R., Ibrahim, H., & Abdullah, A. (2014). Cluster-based replica control protocol for improving data availability in data grid. *Proceedings of the Malaysian National Conference on Databases 2014 (MANCoD'14)*, 75-80.

Madhuram, S., & Kumar, A. (1994). A hybrid approach for mutual exclusion in distributed computing systems. *Sixth IEEE Symposium on Parallel and Distributed Processing*, 18-25.

Mat Deris, M., Abawajy, J. H., & Suzuri H. M. (2004). An efficient replicated data access approach for large-scale distributed systems. *IEEE International Symposium on Cluster Computing and the Grid*, 588-594.

Stonebraker, M. (1979). Concurrency control and consistency of multiple copies of data in distributed ingres. *IEEE Transaction on Software Engineering, 5*(3), 188-194. doi:10.1109/TSE.1979.234180

Thomas, R., H. (1979). A majority consensus approach to concurrency control for multiple copy databases. *ACM Transaction Database Systems, 4*(2), 80-229.

Yusof, Y., Madi, M., & Hassan, S. (2012). Dynamic replication strategy based on exponential model and dependency relationships in data grid. *Journal of Information and Communication Technology (JICT)*, 11, 193-206.

Zhou, W., & Holmes, R. (1999). The design and simulation of a hybrid replication control protocol. *Fourth International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN '99)*, 210-215. doi: 10.1109/ISPAN.1999.778941