


CANADIAN THESES ON MICROFICHE

THÈSES CANADIENNES SUR MICROFICHE

 National Library of Canada
Collections Development Branch

Bibliothèque nationale du Canada
Direction du développement des collections

Canadian Theses on
Microfiche Service

Service des thèses canadiennes
sur microfiche

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

**THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED**

**LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE**

Canada

0-315-20327-7



National Library of Canada

Bibliothèque nationale du Canada

CANADIAN THESES ON MICROFICHE

THÈSES CANADIENNES SUR MICROFICHE

68234

NAME OF AUTHOR/NOM DE L'AUTEUR David Wai-Lok Cheung

TITLE OF THESIS/TITRE DE LA THÈSE Site-Optimal Termination Protocols for
Network Partitioning in a Distributed
Database

UNIVERSITY/UNIVERSITÉ Simon Fraser University

DEGREE FOR WHICH THESIS WAS PRESENTED/
 GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE Master of Science

YEAR THIS DEGREE CONFERRED/ANNÉE D'OBTENTION DE CE GRADE 1985

NAME OF SUPERVISOR/NOM DU DIRECTEUR DE THÈSE Tiko Kameda

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

DATED/DATÉ 17 August 1984 SIGNED/SIGNÉ _____

PERMANENT ADDRESS/RÉSIDENCE FIXE _____

SITE-OPTIMAL TERMINATION PROTOCOLS
FOR NETWORK PARTITIONING
IN A DISTRIBUTED DATABASE

by

David Wai-Lok Cheung

B.Sc., Chinese University of Hong Kong, 1971

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in the Department

of

Computing Science

© David Wai-Lok Cheung 1984

SIMON FRASER UNIVERSITY

August 1984

All rights reserved. This thesis may not be
reproduced in whole or in part, by photocopy
or other means, without permission of the author.

APPROVAL

Name: David Wai-Lok Cheung

Degree: Master of Science

Title of Thesis: Site-Optimal Termination Protocols for
Network Partitioning in a Distributed
Database

Examining Committee:

Chairperson: Lou Hafer

Senior Supervisor: Tiko Kameda

Arthur Lee Liestman

Wo-Shun Luk

External Examiner: Francis Chin
Department of Computing Science
University of Alberta

Date Approved: August 7, 1984

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission. J

Title of Thesis/Project/Extended Essay

SITE-OPTIMAL TERMINATION PROTOCOLS FOR NETWORK

PARTITIONING IN A DISTRIBUTED DATABASE

Author: _____

(signature)

David Wai-Lok Cheung

(name)

17 August 1984

(date)

ABSTRACT

In a distributed database system, a transaction submitted at a site may require execution of its subtransactions at a number of sites. In order to guarantee that no partial result of a transaction is reflected in a database, rendering the database inconsistent, all sites involved must unanimously commit or abort the transaction. Thus a *commit protocol* is required.

A distributed database system must guarantee consistency even if there is failure. When failure occurs, it is desirable to have a *termination protocol (TP)* terminate all the affected transactions consistently. However, in the case of network partitioning, it has been shown that there exists no commit protocol that is *nonblocking*, i.e., some participating sites may have to wait for the repair of this type of failure before they can decide to commit or abort a transaction. Hence the goal here is to design a *site optimal termination protocol*, which has the minimum expected number of waiting sites. Such a protocol will maximize the availability of a database in the presence of network partitioning.

We consider the general case in which *realizable component states* of a partition may have different probabilities of occurrence. We study two classes of TP's, namely, *size-based TP's* and *count-based TP's* and show that there exists a *quorum-based TP* that is site optimal in these classes. Results in this thesis indicate that the set of quorum-based TP's plays an essential role in the design of site optimal TP's, both in the decentralized and the centralized cases.

ACKNOWLEDGEMENTS

My greatest thanks are reserved for my Senior Supervisor, Dr. Tiko Kameda. His enthusiasm, patience and willingness to help have been invaluable. Also, his insistence on precision and clarity in writing has been most helpful. It has been a very great pleasure to work with him.

Thanks are due to Dr. Francis Chin, Dr. Arthur Lee Liestman and Dr. Wu-Shun Luk, not only for their reading and commenting on this thesis, but also for their constant concern and encouragement.

I am also indebted to the faculty and staff of the Computing Science Department of Simon Fraser University for their helpful assistance and for a stimulating environment.

Finally, I wish to dedicate this thesis to my dear wife Juana whose invaluable concern and understanding have made this endeavor both possible and worthwhile. Without her support, my dream of going back to a graduate study would not have materialized.

TABLE OF CONTENTS

Approval	ii
Abstract	iii
Acknowledgements	iv
List of Tables	vii
List of Figures	viii
Chapter One. INTRODUCTION	1
1.1. Transaction Atomicity in a Distributed Database System	1
1.2. Commit Protocols	2
1.3. Blocking Property of Two-Phase Commit Protocols	5
1.4. Three-Phase Commit Protocols	6
Chapter Two. TERMINATION PROTOCOLS FOR NETWORK PARTITIONING	9
2.1. Components and Component States of a Partitioned Network	9
2.2. Three-Phase Commit Protocol under Network Partitioning	11
2.3. Termination Protocols	12
2.4. Some Characteristics of DTP's	15
2.5. Site Optimal DTP's for a special case	16
Chapter Three. SITE OPTIMAL SIZE-BASED DECENTRALIZED TERMINATION PRO- TOCOLS	21
3.1. Introduction	21
3.2. Size-Based DTP's	22
3.3. Site Optimal Size-Based DTP's	27

3.4. Count-Based DTP's	31
Chapter Four. SITE OPTIMAL SIZE-BASED CENTRALIZED TERMINATION PROTO- COLS	36
4.1. Introduction	36
4.2. Size-Based Centralized Termination Protocols	40
4.3. Site Optimal Size-Based Centralized Termination Protocols	47
4.4. Count-Based CTP's	55
4.5. Restricted Decentralized Termination Protocols	58
Conclusion	62
References	63
Figures	64

LIST OF TABLES

Table 2.1. The decisions of dp_1 for $n = 4$ sites.	17
Table 2.2. Values of $E(dp_i)$ and $E(dw_i)$ for $n = 9$ sites.	19
Table 3.1. An example of a size-based DTP for $n = 4$ sites.	26
Table 3.2. An example of a count-based DTP for $n = 4$ sites.	32
Table 4.1. Structure of the size-based CTP g in Theorem 4.3.	46
Table 4.2. Structure of a quorum-based CTP cp_i	48
Table 4.3. Structure of the quorum-based CTP cp_0	48
Table 4.4. The decisions of cp_1 when $n = 4$ sites.	48
Table 4.5. Structure of a quorum-based CTP cw_i	50
Table 4.6. An example of a count-based CTP for $n = 4$ sites.	55
Table 4.7. The decisions of cp_0 for $n = 4$ sites.	57

LIST OF FIGURES

Figure 1.1. FSA of the Centralized Two-Phase Commit Protocol.	64
Figure 1.2. FSA of the Decentralized Two-Phase Commit Protocol.	65
Figure 1.3. FSA of the Centralized Three-Phase Commit Protocol.	66
Figure 1.4. FSA of the Decentralized Three-Phase Commit Protocol.	67

CHAPTER 1

INTRODUCTION

1.1. Transaction Atomicity in a Distributed Database System

In a distributed environment, a transaction submitted at a site may require database entities stored at other sites, and thus cooperative execution at a number of sites. These sites are referred to as the *participating sites* of that transaction. A transaction is a logically *atomic* operation which transforms a consistent database state into another consistent state. In order to maintain the consistency of a database, the effect of a transaction should be either fully reflected in the database or not at all.

If a transaction is executed to completion and its effects are permanently incorporated into the database, we say that the transaction is *committed*. If one of the participating sites cannot complete the transaction, then all the other sites have no choice but to *abort* the transaction.

There are many reasons why a transaction cannot be completed: for instance, request for abortion by a subtransaction itself, deadlock and hardware failure.

In our discussion we will use the term *subtransactions* to refer to different parts of a transaction which are executed at the participating sites. In order to guarantee atomicity, all sites involved in a transaction must unanimously commit or abort the transaction. Hence *commit protocols* (see Section 1.2) are required in distributed database systems. In the following section, a well known protocol, called the *two-phase commit protocol* [LAMP-76, GRAY-78], which guarantees atomicity of transaction, will be introduced.

1.2. Commit Protocols

In a distributed database system, any protocol coordinating the subtransactions of a transaction can be modeled as a collection of finite state automata (FSA), one associated with each participating site [SKEE-81]. A finite state automaton in a certain state reads a set of messages from other sites, takes an appropriate action, sends out a set of messages to other sites (FSA) and then changes its state. Initially, the site which issues a transaction sends its subtransactions to the appropriate sites and each of these sites determines individually whether it would commit or abort the transaction. This step can be conceived as a voting step in which every site involved expresses its intention to commit or abort. After all votes are received, a global decision will be made and all the sites will follow this decision to either unanimously abort or unanimously commit the transaction.

Different protocols use different approaches to collecting votes and making global decision. However, they must all follow the same *commit rule*, i.e., a transaction must be aborted if one or more sites have decided to abort it; otherwise, it must be committed. Protocols which follow the commit rule are called *commit protocols*. In the following we introduce two commit protocols, namely, the *centralized two-phase commit protocol* and *decentralized two-phase commit protocol* [GRAY-78].

These protocols both have two phases and four states.

State *q* is the *initial* state before a site has made its voting decision.

State *w* is the *waiting* state in which a site waits for the message containing the global decision after it has sent out its voting message.

State *c* is the *commit* state in which a site has committed its subtransaction.

State *a* is the *abort* state in which the state in which a site has aborted its subtransaction.

The states *c* and *a* are *final* states in which a transaction is *terminated*, whereas state *w* is only a *transient* state.

Centralized two-phase commit protocol

~~One of the sites is designated as the *coordinator*.~~

PHASE ONE

After executing the subtransaction allocated to it, a participating site sends a voting message to the coordinator. If a site votes "no", this reflects its intention to abort the transaction. It aborts its subtransaction after sending out its voting message "no".

If a site votes "yes", the site is ready to commit the transaction if all other sites agree. However, it cannot commit the transaction at this point, it has to wait for the global decision from the coordinator.

PHASE TWO

If all sites have voted "yes", then the coordinator broadcasts a "commit" message. Otherwise, it broadcasts an "abort" message.

All the participating sites then act (i.e., either commit or abort) unanimously according to the message from the coordinator.

This protocol can be represented by a collection of finite state automata (FSA), one associated with each participating site. In what follows we use the term "site" to refer to the FSA at that site.

Initially, all sites are in state *q*. If a site votes "no", it goes into state *a* after it has sent its vote to the coordinator. If it votes "yes", it goes into the waiting state *w*. As for the coordinator, it is also in state *w* before making a global decision. In the second phase, all sites that are in state *w* change their states to either *c* or *a* unanimously according to the global decision received from the coordinator.

The FSA associated with the coordinator and any other site can be represented by two graphs (see Figure 1.1). Note that in these graphs, if a site is in state *q*, all other sites must be in a state which is adjacent to state *q*, i.e., either state *q*, state *w* or state *a*, and no site could be in state *c*. Similarly, if a site is in state *c*, no site could be in state *q* or state *a*.

With this protocol, different sites could be in different states at any given time, but no site could lead another site by more than one state transition during the execution of the protocol. Therefore these FSA are called *synchronous within one state* [SKEE-81a].

The three-phase commit protocol has a decentralized version. In the decentralized case, no coordinator is appointed, but each site will collect all the votes and use them to make the global decision.

Decentralized two-phase commit protocol

PHASE ONE

If a site decides to abort, it broadcasts a "no" message to all other participating sites and aborts the transaction.

If it decides to commit, then it broadcasts a "yes" message and waits for the votes from all other participating sites.

PHASE TWO

After it has received all the voting messages, each site makes a global decision according to the commit rule: i.e., commit if all sites vote to commit, abort otherwise. Since all sites receive the same set of voting messages, they will all take the same final action, either commit or abort.

The FSA associated with this protocol can also be represented by a graph (see Figure 1.2). Since there is no coordinator, all participating sites have the same FSA and these FSA are all synchronous within one state.

1.3. Blocking Property of Two-Phase Commit Protocols

Two-phase commit protocols guarantee atomicity of distributed transactions, but this is only true in case there is no failure. Consider the centralized two-phase commit protocol. Suppose three sites s_1 , s_2 and s_3 , where s_1 is the coordinator. In phase one, site s_2 sends a "yes" to s_1 . After a while, it detects that it is separated from both sites s_1 and s_3 . This could happen because of site failure or network partitioning. In this situation, site s_2 has no information about what has taken place in sites s_1 and s_3 : the transaction could have been aborted or committed in these sites. The only thing that site s_2 could do is to wait until the failure is repaired and then communicate again with s_1 and s_3 in order to reach a global decision.

While site s_2 is blocked, waiting for recovery from the failure, no new transaction can access that part of the database which will be updated by the suspended transaction at s_2 . To see this, suppose the concurrency control scheme used is the "locking scheme". Then a part of the data base which will be updated by the suspended transaction has been locked by the transaction, and hence no new transaction can access it. If another concurrency control scheme, e.g. "time stamping", is used, the problem will still occur. It is this *blocking* property that degrades the performance of the two-phase commit protocol in the presence of failure.

A similar problem occurs for the decentralized two-phase commit protocol. Hence two-phase commit protocols are called *blocking* protocols [SKEE-81b]. If failure occurs, a distributed transaction, executing under a blocking protocol, could have some of its participating sites wait for a long time for recovery from the failure. This is very undesirable and hence the problem of designing *nonblocking protocols* arises. A nonblocking protocol terminates all participating sites to either abort or commit.

1.4. Three-Phase Commit Protocols

As seen above, blocking property degrades the performance of two-phase commit protocols. Is there any protocol that is free from blocking property? Is it possible to design protocols which are nonblocking for certain types of failure? The first nonblocking commit protocol for site failures was proposed by Skeen [SKEE-81b]. He proposed the *three-phase commit protocol* and showed that it is a nonblocking commit protocol for site failures. This type of protocol is essentially a modification of the two-phase commit protocol. The following is a description of his protocol for the centralized model.

Centralized three-phase commit protocol

PHASE ONE

This phase is the same as PHASE ONE of the two-phase commit protocol.

PHASE TWO

If at least one site votes "no", then the coordinator broadcasts an "abort" message and all sites abort the transaction.

If all sites vote "yes", the coordinator broadcasts a "prepare-to-commit" message to every participating site. After each site has received this message, it returns a "confirmation" message to the coordinator.

PHASE THREE

After the coordinator has received "confirmation" messages from all other sites, it broadcasts a "commit" message. A site commits only after it has received this message.

The FSA associated with the coordinator and other participating sites of this protocol can be represented by the two graphs in Figure 1.3, where p is a new state which indicates the state of a site after it has sent out a "confirmation" message but before it has committed, (i.e., entered state

c). Note that the three-phase commit protocol is also synchronous within one state. The significance of the new state p is that the existence of a site in state p indicates that the global decision is a "commit" decision. Since this protocol is synchronous within one state, if a site has committed, no other site could be in state q or a . Note also that if a site has aborted, (i.e., entered state a), no site could be in state p or c .

Once a site failure is detected, all the operational sites can exchange the information they have and use the following protocol to terminate a transaction.

If there exists a site in states p or c , all operational sites commit. Otherwise, all sites abort.

Thus no operational site needs to wait, that is to say, the three-phase commit protocol is non-blocking for site failure. However, this protocol is not nonblocking for a particular type of failure called "network partitioning". In the next chapter, the relation between the three-phase commit protocol and network partitioning will be discussed in detail.

The three-phase commit protocol also has a decentralized version without coordinator.

Decentralized three-phase commit protocol

PHASE ONE

This phase is the same as PHASE ONE of the decentralized two-phase commit protocol.

PHASE TWO

If the set of votes received by a site contains a "no" message, then the site aborts the transaction.

If all the votes received are "yes", then the site broadcasts a "confirmation" message to every other site.

PHASE THREE

After a site has received "confirmation" messages from all other sites, then it commits the transaction.

The FSA of this protocol is represented by the graph in Figure 1.4. The decentralized three-phase commit protocol is also nonblocking for site failures. The same protocol that was used in the centralized case to terminate transactions in the presence of site failure can also be applied to this case. In the rest of this thesis, a "commit protocol" will denote the three phase commit protocol.

CHAPTER 2

TERMINATION PROTOCOLS FOR NETWORK PARTITIONING

2.1. Components and Component States of a Partitioned Network

In a distributed system, sites communicate via a communication network. A message issued at a site may go through some other sites before it reaches its destination. If some sites or communication links fail, it is possible that the sites are divided into subsets such that the sites in a subset can still communicate with each other, whereas sites in different subsets can no longer communicate. Failure of this type is known as *network partitioning* [SKEE-82a]. The sites within a subset can exchange information and try to decide on a concerted action (commit, abort, or wait) to be taken by all the sites within that subset.

In order to investigate actions to be taken by each site in the event of network partitioning, we define the terms, *component* and *the state of a component* (*component state, for short*),¹ in the context of network partitioning [RAMA-84].

When network partitioning occurs, the participating sites of a transaction are divided into disjoint sets of sites called *components*. Communication between sites in different components is disrupted, whereas communication among the sites within a component is still possible. We thus assume that a pair of sites can communicate with each other or not at all. That is, no failure causes disruption of communication in one direction only. Throughout our discussions, we consider an n -site network and the set of all sites is denoted by I . We use Γ to denote the set of all components and C to denote a typical component in Γ .

Since our main interest is in the design of *termination protocols* (see section 2.3) for network partitioning, we will not concern ourselves with the detection of network partitioning. We

¹ In [RAMA-84] component was called *group* and component state was called *component*. In order to be compatible with the general usage of the term "component", we have adopted new terminology.

assume that site failures as well as network partitioning can be somehow detected, either by operational sites or by the underlying network.

When a transaction is executed under the three-phase commit protocol, the state (q, w, a, etc.) of a site depends on the time when the partitioning occurs. The sites belonging to a component could be in different states, and thus we need a notation to represent the information about the states of the sites in a component.

Let Q be the set of all possible states of the FSA associated with the three-phase commit protocol, i.e., $Q = \{q, w, p, a, c\}$. To represent the fact that site i is in state s , we use an ordered pair (i, s) in $I \times Q$. Let S be a set of ordered pairs from $I \times Q$. S is a *realizable state of component C* (*realizable component state, for short*) [CHIN-83] iff

- (1) $C = \{i \mid (i, s) \in S\}$,
- (2) there do not exist two different ordered pairs in S that have the same first element, and
- (3) the second elements of all the pairs in S are either the same or adjacent states in the FSA associated with the commit protocol.

The first point in the above definition signifies that set S represents the state of component C . The second point ensures that a site can be in exactly one state. The third point follows from one-synchrony of the three-phase commit protocol, i.e., any pair of sites of a component must be in the same or adjacent states. Any set S satisfying these three conditions represents a realizable state of a component in a partition under the three phase commit protocol. See Example 2.1 below for examples of realizable and unrealizable component states.

Throughout our discussion, when we refer to a component state, it is assumed to be realizable unless otherwise stated. For any component state S , we use the notation $\text{comp}(S) = \{i \mid (i, s) \in S\}$ and $\text{state}(S) = \{s \mid (i, s) \in S\}$. With this notation, S is a state of the component $\text{comp}(S)$. Two component states S_1 and S_2 are said to be *concurrent* if $\text{comp}(S_1)$ and $\text{comp}(S_2)$ are disjoint and $\text{state}(S_1) \cup \text{state}(S_2)$ contains one state or only adjacent states. Intui-

tively, this means that the two components $\text{comp}(S_1)$ and $\text{comp}(S_2)$ which are in state S_1 and S_2 , respectively, can occur together in a partition.

Example 2.1.

Suppose there are only three participating sites, i.e., $I = \{1, 2, 3\}$. Then $C_1 = \{1, 2\}$ and $C_2 = \{3\}$ are two disjoint components in a partition. $S_1 = \{(1, p), (2, w)\}$ and $S_2 = \{(3, w)\}$ are two concurrent states of C_1 and C_2 , respectively.

Let $S_3 = \{(3, c)\}$. Although S_3 is a realizable component state, it is not concurrent with S_1 , because states c (the state of site 3 in S_3) and w (the state of site 2 in S_1) are not adjacent.

Let $S_4 = \{(1, q), (2, p)\}$. Then S_4 is an unrealizable component state because state q and state p are not adjacent. \square

2.2. Three-Phase Commit Protocol under Network Partitioning

When network partitioning occurs, can we consistently terminate all the sites without making some of them wait until communication is reestablished? It is reasonable to terminate all the sites in a component by the same action; namely "commit" or "abort", since they can still communicate with each other and can share the information collected within the component.

In the following, when we refer to the termination of a component in a certain state, we mean the termination of the subtransactions by a particular action at all sites in the component. Also, when we say that a component state S is terminated, we mean that the component $\text{comp}(S)$ in state S is terminated to either "commit" or "abort". In the presence of network partitioning, we hope to terminate all concurrent component states consistently. That is, we wish to avoid the situation when one component state is terminated to "commit" and another concurrent component state is terminated to "abort".

Can the three-phase commit protocol terminate all components in all realizable states, i.e., can it terminate all realizable component states? Unfortunately, the answer is negative. It has

been observed that if a protocol can terminate a component C in all realizable states, then the components disjoint from C must wait when they are in certain states, i.e., sites in these components can neither abort nor commit [CHIN-83].

The following example illustrates this observation.

Example 2.2.

Let $I = \{1, 2, 3\}$. Then $S_1 = \{(1, p), (2, p)\}$ and $S_2 = \{(1, w), (2, w)\}$ are states of component $C = \{1, 2\}$. Similarly $S_3 = \{(3, c)\}$ and $S_4 = \{(3, a)\}$ are states of the component $\{3\}$.

Here we assume that a transaction is executed under the decentralized commit protocol. Observe that if f is a protocol that terminates both S_1 and S_2 , it must terminate them to "commit" and "abort", respectively. The reason is that S_1 is concurrent with S_3 and site 3 of S_3 has committed, therefore f must terminate S_1 to "commit" in order to preserve the consistency of the database. Similarly S_2 must be terminated to "abort", since it is concurrent with S_4 .

Let us now consider two component states $S_5 = \{(3, p)\}$ and $S_6 = \{(3, w)\}$. They are both concurrent with S_1 and S_2 . If f terminates one of them to "abort", then it will contradict the decision taken on S_1 . On the other hand, if f terminates one of them to "commit", then it will contradict the decision taken on S_2 . This simple example illustrates the fact that no protocol can consistently terminate all component states. It is now clear that the three-phase commit protocol is blocking for network partitioning. This is true in both the centralized and decentralized cases. \square

2.3. Termination Protocols

It was shown in the last section that no protocol can terminate all realizable component states. We thus wish to have a protocol that minimizes the expected number of waiting sites and hence maximizes the availability of a database when partitioning occurs. Before a detailed discussion of this problem, we first formally define a termination protocol.

A *termination protocol (TP)* can be viewed as a function mapping component states onto decisions to be followed by the sites within the corresponding components. It has to ensure that no two component states that could potentially occur concurrently in a partition are given conflicting decisions.

We use "com", "ab" and "wa" to represent the three decisions "commit", "abort" and "wait", respectively.

Observe that a component which contains a site in state q or state a can always be terminated to *ab*. If a component has a site in state a, then there is no choice but to abort the transaction because at least one site has already aborted the transaction. If a component has a site in state q, then no global decision has been made and no site could have committed the transaction. However, it is possible that sites of some other components in the same partition have aborted the transaction. Therefore, such a component must be terminated to *ab*.

A similar argument applies to the case where a component has a site in state c. Such a component should be terminated to *com*. It follows from the above observation that only those component states with sites in state p and/or w are crucial in defining a termination protocol: a termination protocol is completely defined by mapping these component states to "ab", "com" or wait.

Definition 2.1. [CHIN-83] A *termination protocol (TP, for short) f* is a function from the set of all realizable component states to the set of decisions $\{com, ab, wa\}$ with the following two conditions.

- (1) *f* satisfies the *nonreversal condition*,² i.e., for any component state S , $c \in \text{state}(S)$ implies that $f(S) = com$, and $\{q, a\} \cap \text{state}(S) \neq \emptyset$ implies that $f(S) = ab$.
- (2) *f* satisfies the *consistency condition*, i.e., for any two concurrent component states S_1 and S_2 , $\{f(S_1), f(S_2)\} \neq \{com, ab\}$. \square

² This condition was called *preservation property* in [CHIN-83].

The nonreversal condition of a TP is required because of the observation made before Definition 2.1. The consistency condition simply ensures that, even though two component cannot exchange information, they must be terminated by "non-conflicting" actions.

For convenience we will use Θ to denote the set of all component states with the sites in state p and/or w , i.e., $\Theta = \{S \mid \text{state}(S) \subseteq \{p, w\}\}$. $\Theta_p \subseteq \Theta$ denotes the subset of Θ which contains all the component states that have at least one site in state p . Similarly, $\Theta_w \subseteq \Theta$ denotes the subset which contains all the component states that have at least one site in state w . Throughout the rest of our discussion, when we define a TP f , we will only specify the values of f for the component states in Θ , i.e., $\{f(S) \mid S \in \Theta\}$. The values of f for the component states not belonging to Θ are uniquely determined by the nonreversal condition, and therefore we do not specify them explicitly.

When a TP is used together with the centralized three-phase commit protocol, the TP is called a *centralized termination protocol (CTP)*. Similarly, a TP in the decentralized case is called a *decentralized termination protocol (DTP)*.

As stated above, we wish to design a TP which minimizes the expected number of waiting sites. Such a TP is called a *site optimal termination protocol* [CHIN-83].

Components which result from network partitioning have different probabilities of occurrence. For a component state S , $Pr(S)$ denotes the probability of its occurrence.

Let $E(f)$ denote the expected number of waiting sites under a TP f . Note that if $f(S) = wa$, all sites in $\text{comp}(S)$ wait. Therefore, we have

$$E(f) = \sum_{S \in W} |S| Pr(S),$$

where W is the set of component states that wait under f , i.e., $W = \{S \mid f(S) = wa\}$, and $|S|$ denotes the number of sites in $\text{comp}(S)$.

$E(f)$ gives a measure of performance of a TP f in the presence of network partitioning. If $E(f)$ is small then the availability of the database is high when partitioning occurs. For example,

in the case where locking is used, the locks at a site cannot be released when the site waits. We want to find a site optimal termination protocol, i.e., a TP with the minimum $E(f)$ value.

Definition 2.2 [CHIN-83]. A TP is said to be a *site optimal termination protocol* if it has the minimum expected number of waiting sites. \square

Thus if \mathbf{TP} is the set of all TP's, then $f \in \mathbf{TP}$ is site optimal iff $E(f) = \min\{E(g) \mid g \in \mathbf{TP}\}$.

In the rest of this thesis, we will be mainly concerned with site optimal termination protocols *within* certain subclasses of \mathbf{TP} , i.e., $\min\{E(g) \mid g \in \mathbf{k}\}$, where $\mathbf{k} \subseteq \mathbf{TP}$.

2.4. Some Characteristics of DTP's

The following properties distinguish DTP's from CTP's.

Theorem 2.1 [CHIN-83]. A necessary and sufficient condition for a function f , from the set of realizable component states to the set of decisions $\{\text{com}, \text{ab}, \text{wa}\}$, to be a DTP is

- (1) f satisfies the nonreversal condition, and,
- (2) For any two component states $S_1, S_2 \in \Theta$ such that $\text{comp}(S_1) \cap \text{comp}(S_2) = \emptyset$, $\{f(S_1), f(S_2)\} \neq \{\text{com}, \text{ab}\}$ holds. \square

Comparing this theorem with Definition 2.1, it is seen that the consistency condition in Definition 2.1 is replaced by the second condition in Theorem 2.1.

This reflects the fact that in the distributed case any two disjoint components give rise to pairs of concurrent component states and therefore a DTP must terminate them consistently.

Recall that all the component states in Θ have sites in either state p or state w and these two states are adjacent in the FSA.

For a given component C , the state of C which has all its sites in state p is denoted by p^C . Similarly w^C will denote the component state which has all its sites in state w .

Lemma 2.1 [CHIN-83]. *For a given component C and a DTP f*

- (1) *either $f(p^c) = com$ or $f(p^c) = wa$, and*
- (2) *either $f(w^c) = ab$ or $f(w^c) = wa$. \square*

p^c can occur concurrently with a component state that contains a site in state c , hence p^c cannot be terminated to ab . Similar reasoning applies to (2) above. The conditions on f in this lemma are essential characteristics of a DTP. In Chapter 4 we will see that the first condition does not hold for CTP's.

Lemma 2.2 [CHIN-83]. *For any two disjoint components C_1 and C_2 and any DTP f , at least one of the two values, $f(p^{C_1})$ and $f(w^{C_2})$, must be wa . \square*

If neither value is wa , it follows from Lemma 2.1 that $f(p^{C_1}) = com$ and $f(w^{C_2}) = ab$. However, this violates the consistency condition on TP.

Lemma 2.3 [CHIN-83]. *For any component C_1 , let $S_i, S_j \in \Theta$ be two states of C_1 . If $f(S_i) = com$ and $f(S_j) = ab$ then $f(S_k) = wa$ for every state S_k of C_2 that is disjoint from C_1 . \square*

Note that state S_k in Θ is concurrent with both S_i and S_j . Since S_i and S_j are terminated to conflicting actions, S_k has no other choice but to wait.

In particular, if both p^{C_1} and w^{C_1} are terminated by f , they will be terminated to conflicting actions. Thus all states of C_2 in Θ must wait.

2.5. Site Optimal DTP's for a special case

In this section, we investigate site optimal DTP's in a special case. Although the case is far from general, it gives us some insight into general site optimal termination protocols.

In general, different component states have different probabilities of occurrence. In this section, we assume that the probabilities for different component states are all equal, and therefore the problem of finding a site optimal protocol is reduced to finding a TP which has the minimum

sum of waiting sites over all component states. In this case,

$$E(f) = \sum_{S \in W} |S|$$

where W is the set of waiting component states.

For this case, site optimal DTP's have been found [CHIN-83]. In order to present site optimal DTP's, we first introduce a particular class of DTP's, namely, *quorum-based DTP's* [CHIN-83].

As before let n be the number of participating sites. For a given integer k ($0 \leq k < n/2$), define a DTP dp_k as follows, where $S \in \Theta$.

- (1) If $|S| \leq k$, let $dp_k(S) = wa$.
- (2) If $k < |S| < n-k$ and $p \in \text{state}(S)$, let $dp_k(S) = com$.
- (3) If $k < |S| < n-k$ and $\text{state}(S) = \{w\}$, let $dp_k(S) = wa$.
- (4) If $|S| \geq n-k$ and $p \in \text{state}(S)$, let $dp_k(S) = com$.
- (5) If $|S| \geq n-k$ and $\text{state}(S) = \{w\}$, let $dp_k(S) = ab$.

A DTP defined as above is said to be *quorum-based*. dp_k acts on a component state according to its size as well as whether its sites are all in state w or not. The set of all dp_k 's is denoted by QDp . In Table 2.1, an example of a quorum-based DTP is given. In that example, there are four sites involved (i.e., $n = 4$), and k is equal to 1. The table shows the decision by the quorum-based DTP dp_1 on every realizable component state in Θ . Note that the entries in the first, third and fifth columns represent component states. For example, an entry (p ---) represents the component state $\{(1, p)\}$.

There is another set of quorum-based DTP's denoted by dw_k ($0 \leq k < n/2$). dw_k is defined in the same way as dp_k except that p and w are interchanged and so are *com* and *ab*. The set of all dw_k 's is denoted by QDw . The union of QDp and QDw is denoted by QD . Thus, QD is the set of all quorum-based DTP's.

For a quorum-based DTP dp_k and an integer r ($1 \leq r \leq k$), if S is a component state of size r , then $dp_k(S) = wa$. Since we only consider the case where a component state has all its sites

component state	decision	component state	decision	component state	decision
site 1 2 3 4		site 1 2 3 4		site 1 2 3 4	
p - - -	wa	w - - p	com	w p - w	com
- p - -	wa	- p p -	com	w w - p	com
- - p -	wa	- p w -	com	p - p p	com
- - - p	wa	- w p -	com	p - p w	com
w - - -	wa	- p - p	com	p - w p	com
- w - -	wa	- p - w	com	w - p p	com
- - w -	wa	- w - p	com	p - w w	com
- - - w	wa	- - p p	com	w - p w	com
w w - -	wa	- - p w	com	w - w p	com
w - w -	wa	- - w p	com	- p p p	com
w - - w	wa	p p p -	com	- p p w	com
- w w -	wa	p p w -	com	- p w p	com
- w - w	wa	p w p -	com	- w p p	com
- - w w	wa	w p p -	com	- p w w	com
p p - -	com	p w w -	com	- w p w	com
p w - -	com	w p w -	com	- w w p	com
w p - -	com	w w p -	com	w w w -	ab
p - p -	com	p p - p	com	w w - w	ab
p - w -	com	p p - w	com	w - w w	ab
w - p -	com	p w - p	com	- w w w	ab
p - - p	com	w p - p	com		
p - - w	com	p w - w	com		

Table 2.1. The decisions of dp_1 for $n = 4$ sites.

either in state p or w, the number of component states S such that $|S| = r$ is given by $2^r \binom{n}{r}$. Therefore the total number of waiting sites over all the component states of size r is given by $r 2^r \binom{n}{r}$.

For an integer r ($k < r < n-k$) and for every component C of size r , among all the states of C , w^C is the only waiting component state under dp_k . Therefore the total number of waiting sites over all the component states of size r is given by $r \binom{n}{r}$.

For an integer r ($n-k \leq r < n$), none of the component state of size r waits under dp_k .

Hence for any integer k ($0 \leq k < n/2$), if all component states have the same probability, we have

$$E(f) = \sum_{r=1}^k r \binom{n}{r} + \sum_{r=k+1}^{n-k-1} r \binom{n}{r}$$

We can similarly show that $E(dw_k) = E(dp_k)$.

It was shown in [CHIN-83] that a site optimal DTP exists in **QD**. They first showed that for every DTP f , there exists a k such that the number of waiting sites under f is at least as large as under dp_k . They then showed that by comparing all the members in **QD**, a site optimal DTP can be found in **QD**. For $n = 9$, Table 2.2 lists the values $E(dp_k) = E(dw_k)$ for all k . By comparing all values in Table 2.2, we find that dp_2 has the minimum expected number of waiting sites. Hence dp_2 is site optimal for $n = 9$. Since $E(dw_2) = E(dp_2)$, dw_2 is also site optimal if $n = 9$.

Theorem 2.2 [CHIN-83] *Let n be the number of sites involved and let k be the largest integer such that*

$$k 2^k \leq n.$$

Then both dp_k and dw_k are site optimal DTP's. \square

In the following, we will try to abstract some of the characteristics of a quorum-based DTP. Recall that Γ denotes the set of components and let Γ_k denote the set of all the components that are of size k . Recall also that Θ is the set of all the component states with sites in state p and/or w . Let $\Theta(C)$ denote the set of component states S in Θ such that $\text{comp}(S) = C$ and let $\Theta_p(C)$

k	$E(dp_k)/E(dw_k)$
0	2295
1	2232
2	2196
3	3456
4	10368

Table 2.2. Values of $E(dp_k)$ and $E(dw_k)$ for $n = 9$ sites.

denote the set of component states in $\Theta(C)$ which have at least one site in state p . Similarly we define $\Theta_w(C)$ by replacing p with w .

Let $ALL_t(f)$ denote the set of all components C such that all the component states in $\Theta(C)$ are terminated by f , i.e., for each $S \in \Theta(C)$, $f(S) = com$ or ab . Similarly $ALL_{wa}(f)$ denotes the set of all components C such that all component states in $\Theta(C)$ are mapped to wa by f . Let $WONLY_{wa}(f)$ denote the set of all components C with the property that all the component states in $\Theta_p(C)$ are terminated and w^c is the only waiting component state in $\Theta(C)$. $PONLY_{wa}(f)$ is defined similarly by replacing $\Theta_w(C)$ for $\Theta_p(C)$ and p^c for w^c .

With the above notation, we can describe some important properties of a quorum-based DTP dp_k .

- (1) For all $r \leq k$, $\Gamma_r \subseteq ALL_{wa}(dp_k)$.
- (2) For all r ($k < r < n-k$), $\Gamma_r \subseteq WONLY_{wa}(dp_k)$.
- (3) For $r \geq n-k$, $\Gamma_r \subseteq ALL_t(dp_k)$.

The quorum-based DTP dw_k has similar properties. When dp_k is replaced by dw_k and $WONLY_{wa}(dp_k)$ by $PONLY_{wa}(dw_k)$, the above three properties hold for dw_k .

With this background, in the next two chapters, we will generalize the notion of the quorum-based DTP.

CHAPTER 3

SITE OPTIMAL SIZE-BASED DECENTRALIZED TERMINATION PROTOCOLS

3.1. Introduction In Chapter 2, site optimal DTP's were discussed under the assumption that all component states were equally probable. In the general case, different components and component states will have different probabilities of occurrence. Therefore, the expected number of waiting sites involves these probabilities.

In this chapter, site optimal DTP's are investigated in this general context. We introduce a class of DTP's called *size-based DTP's* and discuss site optimality within this class. We will also see that quorum-based DTP's play an important role in the search for optimal size-based DTP's.

Recall that Γ_k denotes the set of all components of size k and $\Theta(C) \subseteq \Theta$ denotes the set of all states of a component C .

Definition 3.1. A DTP f is a *size-based DTP* if it satisfies the following condition: for any positive integer $k < n$, if a component C in Γ_k has a state S such that $f(S) \neq wa$, then every other component C_i in Γ_k has a state S_i such that $f(S_i) \neq wa$. \square

Intuitively, if a size-based DTP f terminates a component of size k in a certain state S , (i.e., map S to com or ab), then f terminates every component of the size k in at least one state.

Recall the notation $ALL_{wa}(f)$ and $ALL_i(f)$ introduced in Section 2.5. A component C belongs to $ALL_{wa}(f)$ (or $ALL_i(f)$, respectively) if the TP f maps all component states in $\Theta(C)$ to wa (or not wa , respectively). The following lemma gives a property of size-based DTP's.

Lemma 3.1. Let f be any size-based DTP. For each positive integer $k < n$, either $\Gamma_k \subseteq ALL_{wa}(f)$ or $\Gamma_k \cap ALL_{wa}(f) = \emptyset$.

Proof. Let $C \in \Gamma_k \cap ALL_{wa}(f)$ and let $C_1 \neq C$ be a component of size k . If C_1 were not a member of $ALL_{wa}(f)$, then it follows from Definition 3.1 that neither would C be a member of

$ALL_{wa}(f)$. Hence C_1 must be a member of $ALL_{wa}(f)$ and this proves that Γ_k is a subset of $ALL_{wa}(f)$. \square

3.2. Size-Based DTP's

In this section, we introduce a partial order among the DTP's, and then we show that some DTP's are good candidates for site optimal size-based DTP's. In addition, we prove important characteristics of a size-based DTP.

For any two TP's, f_1 and f_2 if $f_1(S) = wa$ implies $f_2(S) = wa$ for any component state S , then we denote this relation by $f_1 \ll f_2$. This relation is a partial order on the set of DTP's, since it is transitive and reflexive. Note that we can introduce a similar partial order on the set of CTP's. We will make use of such a partial order later. If $f_1 \ll f_2$ it follows from the definition of the expected value $ES(f)$ that $ES(f_1) \leq ES(f_2)$.

In the following, when we say that a DTP f_1 is *modified* to a DTP f_2 , we mean that some values of f_1 are changed, giving rise to a new DTP f_2 . We specify only those changes explicitly; the other values remain the same. Also a change is always from $f_1(S) = wa$ to $f_2(S) \neq wa$ for some component states S . Therefore $f_2 \ll f_1$ easily follows.

Lemma 3.2. *For a given size-based DTP f and an integer k ($n/2 < k < n$), if a component C in Γ_k has two states S_1 and S_2 such that $f(S_1) = com$ and $f(S_2) = ab$, then there exists a size-based DTP g such that $g \ll f$ and $\Gamma_k \subseteq ALL_i(g)$.*

Proof. Let r be any integer such that $1 \leq r \leq n - k$. Then clearly $r < n/2$. This variable represents the size of a component state concurrent with any state of C . We first show that, for any such r , all component of size r must wait under f regardless of their states, i.e., $\Gamma_r \subseteq ALL_{wa}(f)$.

Recall that I denotes the set of all sites. Let $C_2 \subseteq I$ be a nonempty component of size r disjoint from the component C , i.e., $C_2 \subseteq I - C$. Since $f(S_1) = com$, $f(S_2) = ab$ and both S_1 and S_2 are states of C , it follows from Lemma 2.3 that $C_2 \in ALL_{wa}(f)$. It then follows from Lemma 3.1

that $\Gamma_r \subseteq ALL_{wa}(f)$. Since the above argument is valid for all integers r such that $1 \leq r \leq n - k$, it follows furthermore that f makes a component wait if it is disjoint from any component of size k . Hence, if any component of size k has a state S such that $f(S) = wa$, we can modify f to g in the following way:

Suppose the condition of the lemma holds and let a component of size k has a state S such that $f(S) = wa$.

- (1) If $p \in \text{state}(S)$, let $g(S) = com$.
- (2) If $\text{state}(S) = \{ w \}$, let $g(S) = ab$.

It follows from the definition of g that $g \ll f$, $\Gamma_k \subseteq ALL_k(g)$ and g is a size-based DTP. \square

In the proof of the above lemma, mapping for a component state S having at least one site in state p was changed from $f(S) = wa$ to $g(S) = com$. Only if S had all its sites in state w , it was terminated to ab by g . This scheme of modifying a TP is called the *commit-favouring scheme*. A TP f could also be modified to g in such a way that $g(S) = ab$ if S contains at least one site in state w , and $g(S) = com$ otherwise. This scheme is called the *abort-favouring scheme*.

For any size-based DTP f and any integer ($n/2 < k < n$), if Γ_k satisfies the condition of Lemma 3.2, then there exists a size-based DTP g such that for each component state S , $g(S) = wa$ implies $f(S) = wa$, and no component in Γ_k waits under g regardless of the state it is in. If there is no k such that Γ_k satisfies the condition of Lemma 3.2, does such a g still exist? The lemma below answers this question.

Before we state the lemma, recall that $PONLY_{wa}(f)$ is the set of all components C that are terminated by the TP f , except when C is in state p^C . Similarly $WONLY_{wa}(f)$ is the set of all components C that terminated by f , except when C is in state w^C . Throughout the rest of the thesis, if A is a collection of component states, we use $f(A)$ to denote the set $\{f(S) \mid S \in A\}$.

Lemma 3.3. *For a given size-based DTP f and a positive integer $k < n$, if there is no component $C \in \Gamma_k$ such that $f(\Theta(C))$ contains $\{ab, com\}$ and if there is a component $C_1 \in \Gamma_k$*

such that $f(\Theta(C_1)) \neq \{wa\}$ then there exists a size-based DTP g such that $g \ll f$ and $\Gamma_k \subseteq PONLY_{wa}(g) \cup WONLY_{wa}(g)$.

Proof. If a component C_1 satisfying the condition of the lemma exists, it follows from Definition 3.1 that for all $C \in \Gamma_k$, either $com \in f(\Theta(C))$ or $ab \in f(\Theta(C))$. Let S_j be the state of some $C \in \Gamma_k$ such that $f(S_j) = com$. Let S_i be any state of a component disjoint from C . Thus S_i is concurrent with S_j . Since $f(S_j) = com$, $f(S_i)$ must be either wa or com . Therefore, in general, for every component state S that can occur concurrently with a component state in $\Theta(C)$, $f(S) = wa$ or $f(S) = com$. Thus we can modify f to g on all component states in $\Theta_p(C)$ that wait under f by the commit-favouring scheme. Note that since $\{ab, com\}$ is not a subset of $f(\Theta(C))$, we must have $f(w^C) = wa$. Therefore $C \in WONLY_{wa}(g)$.

If $ab \in f(\Theta(C))$ on the other hand, it follows from a similar argument that f can be modified to g on all component states in $\Theta_w(C)$ that wait under f by the abort-favouring scheme to make C a member of $PONLY_{wa}(g)$. Hence the DTP g thus obtained from f has the property that Γ_k is a subset of $PONLY_{wa}(g) \cup WONLY_{wa}(g)$.

Note that the modification done on f does not affect the property of f being a size-based DTP, and therefore g is also a size-based DTP. Since only component states S with $f(S) = wa$ have been involved in modification, we have $g \ll f$. \square

The following theorem integrates the results of Lemmas 3.1, 3.2 and 3.3.

Theorem 3.1. *For any given size-based DTP f , there exists a size-based DTP g such that $g \ll f$ and for any Γ_k ($1 \leq k \leq n-1$), one of the following three holds:*

- (1) $\Gamma_k \subseteq ALL_{wa}(g)$.
- (2) $\Gamma_k \subseteq PONLY_{wa}(g) \cup WONLY_{wa}(g)$.
- (3) $\Gamma_k \subseteq ALL_l(g)$.

Proof. For any integer k ($1 \leq k \leq n-1$), if Γ_k is not a subset of $ALL_{wa}(f)$ then it satisfies either the condition of Lemma 3.2 or that of Lemma 3.3. In any case, as was shown in Lemma 3.2

and Lemma 3.3, respectively, f can be modified to a size-based DTP g such that $g \ll f$ and $\Gamma_k \subseteq \text{PONLY}_{w_a}(g) \cup \text{WONLY}_{w_a}(g)$ or $\Gamma_k \subseteq \text{ALL}_r(g)$. \square

Observe that if g is a size-based DTP as defined in Theorem 3.1 and if C is a component belonging to $\text{ALL}_r(g)$, then any component that is disjoint from C must belong to $\text{ALL}_{w_a}(g)$. Therefore, for any integer b ($n/2 < b < n$), if Γ_k is a subset of $\text{ALL}_r(g)$ for each k ($b \leq k < n$), then Γ_j is a subset of $\text{ALL}_{w_a}(g)$ for each j ($0 < j \leq n-b$). Observe also that if a component C belongs to $\text{PONLY}_{w_a}(g)$, then any component that is disjoint from C is either a member of $\text{PONLY}_{w_a}(g)$ or a member of $\text{ALL}_{w_a}(g)$. Similarly, if C belongs to $\text{WONLY}_{w_a}(g)$, then any component that is disjoint from C is either a member of $\text{WONLY}_{w_a}(g)$ or a member of $\text{ALL}_{w_a}(g)$. Therefore, in the above theorem, if r is a positive integer less than $n - k$, then Γ_k satisfying condition (3) implies that Γ_r satisfies condition (1). Also if Γ_k satisfies condition (2) then Γ_r satisfies either condition (1) or (2).

From the above observations, we obtain the following result which highlights some important properties of a size-based DTP.

Theorem 3.2. For any size-based DTP f , there exists a size-based DTP h such that $h \ll f$ and there exist two nonnegative integers s and b such that

- (1) $s + b \geq n$ and $b > n/2$,
- (2) for all k ($1 \leq k \leq s$), $\Gamma_k \subseteq \text{ALL}_{w_a}(h)$,
- (3) for each k ($s < k < b$), either $\Gamma_k \subseteq \text{ALL}_{w_a}(h)$ or $\Gamma_k \subseteq \text{PONLY}_{w_a}(h) \cup \text{WONLY}_{w_a}(h)$, and
- (4) for all k ($b \leq k < n$), $\Gamma_k \subseteq \text{ALL}_r(h)$.

Proof. It follows from Theorem 3.1 that there exists a size-based DTP g which has one of the properties mentioned there. We now modify g to h in such a way that h will have the properties (1) through (4).

Let $s = \max\{k \mid \text{for all integer } r \ (1 \leq r \leq k), \Gamma_r \subseteq \text{ALL}_{w_a}(g)\}$.

Let $b = \min\{k \mid \Gamma_k \subseteq \text{ALL}_r(g)\}$. If $b \leq n/2$, let C_1 and C_2 be two disjoint components of size b .

The existence of C_1 and C_2 is guaranteed by the inequality $b \leq n/2$. Then both C_1 and C_2 belong to $ALL_r(g)$, which contradicts Lemma 2.3. Hence $b > n/2$. If $r \leq n - b$, then Γ_r must be a subset of $ALL_{wa}(g)$, since each component in Γ_r is disjoint from some component in Γ_b . Hence $s \geq n - b$, i.e., $s + b \geq n$.

It follows from the minimality of b that for any integer k ($s < k < b$), Γ_k satisfies condition (3). Hence all the conditions mentioned above, except possibly condition (4), are satisfied by g .

For any $k > b$, if Γ_k is not a subset of $ALL_r(g)$, we can modify g to h on all component states of size k that wait under f by the commit-favouring scheme. This modification is feasible because if $1 \leq r \leq n - k$ ($\leq n - b$), all the Γ_r 's are subsets of $ALL_{wa}(g)$. It follows from the way h is defined that h is also a size-based DTP and it satisfies condition (4). Since h is modified from g , it also satisfies all the other conditions. \square

Definition 3.2. A size-based DTP h is a *standardized size-based DTP* if there exist two nonnegative integers s and b such that

- (1) $s + b \geq n$ and $b > n/2$,
- (2) for all k ($1 \leq k \leq s$), $\Gamma_k \subseteq ALL_{wa}(h)$,
- (3) for each k ($s < k < b$), either $\Gamma_k \subseteq ALL_{wa}(h)$ or $\Gamma_k \subseteq PONLY_{wa}(h) \cup WONLY_{wa}(h)$,
- (4) for all k ($b \leq k < n$), $\Gamma_k \subseteq ALL_r(h)$. \square

Definition 3.2 is based on Theorem 3.2. With this definition, Theorem 3.2 can be restated as follows: given any size-based DTP f , there exists a standardized size-based DTP $h \ll f$. An example of a standardized size-based DTP is given below in Table 3.1. In this example, the number of sites is four and the values of s and b are 1 and 3, respectively. Note that all component states of size 1 are made to wait and all component states of size 3 are terminated. For the component states of size 2, the decisions depend on the sites they contain and on the states of these sites.

component state	decision	component state	decision	component state	decision
site 1 2 3 4		site 1 2 3 4		site 1 2 3 4	
p---	wa	w--p	com	w p - w	com
-p--	wa	-pp-	com	w w - p	com
--p-	wa	-pw-	com	p - p p	com
---p	wa	-wp-	com	p - p w	com
w---	wa	-p-p	com	p - w p	com
-w--	wa	-p-w	com	w - p p	com
--w-	wa	-w-p	com	p - w w	com
---w	wa	--pp	wa	w - p w	com
w w --	ab	--pw	ab	w - w p	com
w - w -	wa	--wp	ab	- p p p	com
w -- w	wa	ppp-	com	- p p w	com
- w w -	wa	ppw-	com	- p w p	com
- w - w	wa	pw p -	com	- w p p	com
-- w w	ab	w p p -	com	- p w w	com
pp--	wa	p w w -	com	- w p w	com
p w --	ab	w p w -	com	- w w p	com
w p --	ab	w w p -	com	w w w -	ab
p - p -	com	pp - p	com	w w - w	ab
p - w -	com	pp - w	com	w - w w	ab
w - p -	com	pw - p	com	- w w w	ab
p -- p	com	w p - p	com		
p -- w	com	p w - w	com		

Table 3.1. An example of a size-based DTP for $n = 4$ sites.

3.3. Site Optimal Size-Based DTP's

In Theorem 3.2, we have shown that for every size-based DTP f there exists a standardized size-based DTP $h \ll f$. Recall the class of quorum-based DTP's defined in Section 2.5. In this section, we show that we can always find a quorum-based DTP dp_i or dw_i which satisfies $E(dp_i) \leq E(h)$ or $E(dw_i) \leq E(h)$.

Assume that partitioning has occurred and consider a component C of size k . Let $Pr(C)$ denote the probability of occurrence of the component C and let $P(r, s, k)$, ($0 < k < n$), be the sum of the probabilities of all states of C with exactly r sites in state p and s sites in state w . For example, the component state p^C has all its sites in state p , hence r equals k , s equals 0 and the

probability of its occurrence is the product of $Pr(C)$ and $P(k, 0, k)$, i.e.,

$$Pr(p^C) = Pr(C) P(k, 0, k).$$

Similarly

$$Pr(w^C) = Pr(C) P(0, k, k).$$

Recall that under a quorum-based DTP dp_k ($k < n/2$), all components of size less than or equal to k are wait regardless of their states; each component C of size between k and $n - k$, exclusive, waits in the state w^C and no other component waits. For convenience, let PC_i denote the sum of the probabilities of all components of size i , i.e.,

$$PC_i = \sum_{C \in \Gamma_i} Pr(C),$$

and let

$$P_k = \sum_{r+s=k} P(r, s, k).$$

Theorem 3.3. For an integer k ($0 \leq k < n/2$), the expected number of waiting sites under the quorum-based DTP dp_k is given by the following formulae:

$$E(dp_0) = \sum_{i=1}^{n-1} i PC_i P(0, i, i), \text{ and}$$

$$E(dp_k) = \sum_{i=1}^k i PC_i P_i + \sum_{i=k+1}^{n-k-1} i PC_i P(0, i, i) \text{ for } k > 0.$$

Proof. Suppose $k = 0$. For any integer i ($1 \leq i \leq n-1$) and for every component C of size i , the state w^C is the only waiting state of C under dp_0 . The sum of the probabilities of these component states is given by the product of PC_i and $P(0, i, i)$. Hence

$$E(dp_0) = \sum_{i=1}^{n-1} i PC_i P(0, i, i).$$

Suppose $k \geq 1$. For any integer i ($1 \leq i \leq k$) and for any component C of size i , all states of C wait under dp_k . The sum of the probabilities of occurrence of all these component states is given by PC_i .

For any integer i ($k+1 \leq i < n-k$) and for every component C of size i , the state w^C is the only waiting state of C under dp_k . The sum of the probabilities of these component states is given by the product of PC_i and $P(0, i, i)$. Also for any integer i ($n-k \leq i < n$), no component of size i waits under dp_k . Hence

$$E(dp_k) = \sum_{i=1}^k i PC_i P_i + \sum_{i=k+1}^{n-k-1} i PC_i P(0, i, i) \text{ for } k > 0.$$

The argument used in the above proof also applies to the quorum-based DTP's dw_k , proving the following theorem.

Theorem 3.4. *For an integer k ($0 \leq k < n/2$), the expected number of waiting sites under the quorum-based DTP dw_k is given by the following formulae.*

$$E(dw_0) = \sum_{i=1}^{n-1} i PC_i P(i, 0, i), \text{ and}$$

$$E(dw_k) = \sum_{i=1}^k i PC_i P_i + \sum_{i=k+1}^{n-k-1} i PC_i P(i, 0, i) \text{ for } k > 0. \quad \square$$

Under some conditions, for every size-based DTP f , there exists a quorum-based DTP dp_k or dw_k such that $E(dp_k) \leq E(f)$ or $E(dw_k) \leq E(f)$, as stated in the following theorem.

Theorem 3.5. *If $P(0, k, k) \leq P(k, 0, k)$ for all integers k ($1 \leq k \leq n-1$), then for every size-based DTP f , there exists a quorum-based DTP dp_i ($1 \leq i < n/2$) such that $E(dp_i) \leq E(f)$.*

Proof. It follows from Theorem 3.2 that there exists a size-based DTP h such that $h \ll f$ and there exist two nonnegative integers s and b such that

- (1) $s + b \geq n$ and $b > n/2$,
- (2) for all k ($1 \leq k \leq s$), $\Gamma_k \subseteq ALL_{w_0}(h)$,
- (3) for all k ($s < k < b$), either $\Gamma_k \subseteq ALL_{w_0}(h)$ or $\Gamma_k \subseteq PONLY_{w_0}(h) \cup WONLY_{w_0}(h)$, and
- (4) for all k ($b \leq k < n$), $\Gamma_k \subseteq ALL_t(h)$.

We want to compare the expected number of waiting sites under h and that under the quorum-based DTP $dp_{\bar{b}}$ where $\bar{b} = n - b$. Since $s + b \geq n$, we have $\bar{b} \leq s < n/2$.

For any integer k ($1 \leq k \leq \bar{b}$), it follows from (2) and the definition of quorum-based DTP that all components of size k wait under both h and $dp_{\bar{b}}$ regardless of their states. Hence these two size-based DTPs have the same expected number of waiting sites for components of size k in the range $1 \leq k \leq \bar{b}$.

For any integer k ($\bar{b} < k < b$), it follows from (2) and (3) that, under h , a component C of size k either always waits or at least when C is in one of w^C and p^C . Under $dp_{\bar{b}}$, the component C waits only when it is in state w^C . Since $P(0, k, k) \leq P(k, 0, k)$, we have $Pr(w^C) \leq Pr(p^C)$. Hence the expected number of waiting sites from C under $dp_{\bar{b}}$ is at most as large as that under h .

For any integer k ($b \leq k < n$), and for any component C of size k , it follows from (4) that C never waits under h . This is also true for $dp_{\bar{b}}$. Hence the expected number of waiting sites under h and $dp_{\bar{b}}$ are both equal to zero.

In each case, the expected number of waiting sites under $dp_{\bar{b}}$ is not larger than that under h . Hence $E(dp_{\bar{b}}) \leq E(h)$. Since $h \ll f$, this proves that $E(dp_{\bar{b}}) \leq E(f)$. \square

By replacing the quorum-based DTP dp_i by the quorum-based DTP dw_i , we get a similar result.

Theorem 3.6. *If $P(k, 0, k) \leq P(0, k, k)$ for all integers k ($1 \leq k \leq n-1$), then for every size-based DTP f , there exists a quorum-based DTP dw_i ($1 \leq i < n/2$) such that $E(dw_i) \leq E(f)$. \square*

From these theorems, we see that the set **QD** of quorum-based DTPs plays an important role in the search for site optimal size-based DTPs. For every size-based DTP f , there exists a size-based DTP q in **QD** such that $q \ll f$, therefore by comparing all the quorum-based DTPs, we can find the site optimal size-based DTPs.

Theorem 3.7. If $P(0, k, k) \leq P(k, 0, k)$ for all integers k ($1 \leq k \leq n-1$), let m be an index such that

$$E(dp_m) = \min\{E(dp_k) : 1 \leq k < n/2\}$$

then dp_m is site optimal in the set of size-based DTP's.

Proof. The optimality of dp_m follows from Theorem 3.5. \square

Theorem 3.8. If $P(k, 0, k) \leq P(0, k, k)$ for all integers k ($1 \leq k \leq n-1$), let m be an index such that

$$E(dw_m) = \min\{E(dw_k) : 1 \leq k < n/2\}$$

then dw_m is site optimal in the set of size-based DTP's.

Proof. The theorem follows from Theorem 3.6. \square

This concludes our search for site optimal DTP's among all size-based DTP's. In the next section, we will introduce an interesting subclass of size-based DTP's, called *count-based DTP's*, which is a generalization of quorum-based DTP's.

3.4. Count-Based DTP's

It is natural to assume that when a DTP decides to terminate a component, it bases its decision only on the states of the sites in the component, and not on what sites are in the component. In other words, two component states which have equal number of sites in each state, will be mapped to the same decision.

Given a component state S , let $n_p(S)$ denote the number of sites in state p and let $n_w(S)$ denote the number of sites in state w . Two component states S_1 and S_2 are *state equivalent* if $|S_1| = |S_2|$, $n_p(S_1) = n_p(S_2)$ (or equivalently, $n_w(S_1) = n_w(S_2)$).

Definition 3.3. A DTP f is a *count-based DTP* if for any two state equivalent component states S_1 and S_2 , $f(S_1) = f(S_2)$.

An example of a count-based dependent DTP with four sites is illustrated in Table 3.2.

Theorem 3.9. *Any count-based DTP is a size-based DTP.*

Proof. Let f be a count-based DTP and let k be an integer such that $1 \leq k \leq n-1$. Suppose $C \in \Gamma_k$ and has a state S such that $f(S) \neq wa$. Then for every component C_i in Γ_k has a state S_i such that S and S_i are state equivalent, and it follows from Definition 3.2 that $f(S) = f(S_i)$. In particular $f(S_i) \neq wa$. Hence f is also a size-based DTP. \square

Consider the size-based DTP represented in Table 3.1. Let $S_1 = \{(1, w), (4, p)\}$ and $S_2 = \{(3, p), (4, w)\}$. These two component states have the same numbers of p 's and w 's but are

component state	decision	component state	decision	component state	decision
site		site		site	
1 2 3 4		1 2 3 4		1 2 3 4	
p ---	wa	w - - p	ab	w p - w	ab
- p --	wa	- p p -	wa	w w - p	ab
-- p -	wa	- p w -	ab	p - p p	com
- - p	wa	- w p -	ab	p - p w	wa
w ---	wa	- p - p	wa	p - w p	wa
- w --	wa	p - w	ab	w - p p	wa
-- w -	wa	- w - p	ab	p - w w	ab
--- w	wa	-- p p	wa	w - p w	ab
w w --	ab	-- p w	ab	w - w p	ab
w - w -	ab	-- w p	ab	- p p p	com
w -- w	ab	p p p -	com	- p p w	wa
- w w -	ab	p p w -	wa	- p w p	wa
- w - w	ab	p w p -	wa	- w p p	wa
-- w w	ab	w p p -	wa	- p w w	ab
p p --	wa	p w w -	ab	- w p w	ab
p w --	ab	w p w -	ab	- w w p	ab
w p --	ab	w w p -	ab	w w w -	ab
p - p -	wa	p p - p	com	w w - w	ab
p - w -	ab	p p - w	wa	w - w w	ab
w - p -	ab	p w - p	wa	- w w w	ab
p - - p	wa	w p - p	wa		
p - - w	ab	p w - w	ab		

Table 3.2. An example of a count-based dependent DTP for $n = 4$ sites.

terminated to com and ab, respectively. Therefore this size-based DTP is not a count-based DTP. Hence the set of count-based DTP's is a proper subset of the set of size-based DTP's.

Recall the definition of a standardized size-based DTP. Here we define a similar DTP, namely, *standardized count-based DTP*.

Definition 3.4. A count-based DTP h is said to be *standardized* if there exist two nonnegative integers s and b such that

- (1) $s + b \geq n$ and $b > n/2$,
- (2) for all k ($1 \leq k \leq s$), $\Gamma_k \subseteq ALL_{w_a}(h)$,
- (3) for each k ($s < k < b$), either $\Gamma_k \subseteq ALL_{w_a}(h)$ or $\Gamma_k \subseteq PONLY_{w_a}(h)$, or $\Gamma_k \subseteq WONLY_{w_a}(h)$,
- (4) for all k ($b \leq k < n$), $\Gamma_k \subseteq ALL_l(h)$. \square

It was proved in Theorem 3.2 that any size-based DTP f can be modified to a standardized size-based DTP $h \ll f$. It turns out that if f is a count-based DTP, then f can be modified to a standardized count-based DTP $h \ll f$ as shown in the next theorem.

Theorem 3.10. *For any count-based DTP f , there exists a standardized count-based DTP h such that $h \ll f$.*

Proof. The existence of h follows from Theorem 3.2. and h inherits the properties of a count-based DTP from f .

In the condition (3) of Theorem 3.2, for all k ($s < k < b$), either $\Gamma_k \subseteq ALL_{w_a}(h)$ or $\Gamma_k \subseteq PONLY_{w_a}(h) \cup WONLY_{w_a}(h)$. Since h is a count-based DTP, $\Gamma_k \cap PONLY_{w_a}(h) \neq \emptyset$ implies that $\Gamma_k \subseteq PONLY_{w_a}(h)$. Therefore, for each k ($s < k < b$), either $\Gamma_k \subseteq ALL_{w_a}(h)$ or $\Gamma_k \subseteq PONLY_{w_a}(h)$ or $\Gamma_k \subseteq WONLY_{w_a}(h)$. \square

Theorem 3.11. *For any count-based DTP f , there exists a quorum-based DTP q such that $q \ll f$.*

Proof. For any given count-based DTP f , it follows from Theorem 3.10 that there exists a standardized count-based DTP h such that $h \ll f$.

Let $P = \{ k \mid s < k < b \text{ and } \Gamma_k \subseteq \text{PONLY}_{wa}(h) \}$ and $W = \{ k \mid s < k < b \text{ and } \Gamma_k \subseteq \text{WONLY}_{wa}(h) \}$. If $P \neq \emptyset$, then let $m_p = \min\{ k \mid k \in P \}$. If $W \neq \emptyset$, then let $m_w = \min\{ k \mid k \in W \}$.

We now consider four cases.

Case A: suppose both P and W are empty sets, then for all k ($s < k < b$), $\Gamma_k \subseteq \text{ALL}_{wa}(h)$. Let $\bar{b} = n - b$. Since $s + b \geq n$, therefore $\bar{b} \leq s$. By comparing the waiting component states under $dp_{\bar{b}}$ and h , it is clear that $dp_{\bar{b}} \ll h$.

Case B: suppose $W = \emptyset$ and $P \neq \emptyset$, then for all k ($s < k < b$), either $\Gamma_k \subseteq \text{ALL}_{wa}(h)$ or $\Gamma_k \subseteq \text{PONLY}_{wa}(h)$. Let $\bar{b} = n - b$. Since $s + b \geq n$, therefore $\bar{b} \leq s$. By comparing the waiting component states under $dp_{\bar{b}}$ and h , it is clear that $dp_{\bar{b}} \ll h$.

Case C: suppose $P = \emptyset$ and $W \neq \emptyset$, then for all k ($s < k < b$), either $\Gamma_k \subseteq \text{ALL}_{wa}(h)$ or $\Gamma_k \subseteq \text{WONLY}_{wa}(h)$. Let $\bar{b} = n - b$. Since $s + b \geq n$, therefore $\bar{b} \leq s$. By comparing the waiting component states under $dw_{\bar{b}}$ and h , it is clear that $dw_{\bar{b}} \ll h$.

Case D: suppose both P and Q are nonempty sets, then either $m_1 < m_2$ or $m_2 < m_1$.

If $m_1 < m_2$. Let $m_2 \leq r < b$ and C be a component of size r . Consider a subcomponent C_1 of C with size equal to m_1 , since $\Gamma_{m_1} \subseteq \text{PONLY}_{wa}(h)$, therefore $h(w^{C_1}) = ab$. It is clear that, if $h(w^C) = wa$, then the value of w^C can be modified to ab because it contains the set w^{C_1} . Similarly, if $h(p^C) = wa$, then p^C can be modified to com . It then follows from Lemma 3.2 that h can be modified so that for all $m_2 \leq r < b$, $\Gamma_r \subseteq \text{ALL}_{wa}(h)$ and it is also true that $s + m_2 \geq n$. Let $\bar{m} = n - m_2$. By comparing the waiting component states under $dp_{\bar{m}}$ and h , it is clear that $dp_{\bar{m}} \ll h$.

If $m_2 < m_1$, let $\bar{m} = n - m_1$. The same proof applies and $dp_{\bar{m}} \ll h$.

Since $h \ll f$ and there always exists a quorum-based DTP q such that $q \ll h$, it follows that $q \ll f$. \square

Because of Theorem 3.11, we can compare all the quorum-based DTP's to find the site-optimal count-based DTP. Recall that QD denotes the set of all quorum-based DTP's.

Corollary 3.1. *If q is a quorum-based DTP such that*

$$E(q) = \min\{ES(f) \mid f \in QD\},$$

then q is the site optimal count-based DTP.

Proof. The proof follows directly from Theorems 3.11. \square

Because the set of count-based DTP's is a proper subset of the set of size-based DTP's, in the search for a site-optimal count-based DTP, we have a stronger result in Corollary 3.1, i.e., the condition $P(0, k, k) \leq P(k, 0, k)$ in Theorem 3.7 (or $P(k, 0, k) \leq P(0, k, k)$ in Theorem 3.8) has been removed.

CHAPTER 4

SITE OPTIMAL SIZE-BASED CENTRALIZED TERMINATION PROTOCOLS

4.1. Introduction

In this chapter we continue our discussion on site optimal termination protocols, this time, in the centralized case. In the previous chapter, we discussed extensively the problem of finding a DTP site optimal within the class of size-based DTP's and it was found that each site optimal size-based DTP is a quorum-based DTP. We prove an analogous result in what follows.

We first investigate the properties of a CTP that distinguish it from a DTP. These properties help us in the search for site optimal CTP's. We then define the *size-based CTP* and try to find a CTP site optimal within the class of size-based CTP's. We also introduce the *quorum-based CTP*, analogous to the quorum-based DTP.

Recall the centralized three-phase commit protocol described in Section 1.4, in which coordinator sites collect the votes and broadcast decisions. In order to simplify our discussion, we assume that there is only one coordinator and, without loss of generality, we consider site 1 as the coordinator. Whenever a decision is made, the coordinator is the first site to act on the decision. For example, in the second phase, after the coordinator has broadcast "prepare-to-commit" messages, it is the first site to go into state p . Also, in the third phase, after it has broadcast "commit" messages, it is the first site to commit the transaction.

Recall that Γ denotes the set of all components. Because the coordinator has some special properties, we separate Γ into two sets: Γ^+ denotes the set of all components that contain the coordinator and Γ'' denotes the set of all components that do not contain the coordinator. Note that if a component is a member of Γ^+ , then any component that is disjoint from it is in Γ'' .

Recall the fundamental property of DTP stated in Lemma 2.3: if f is a DTP and C_1, C_2 are two disjoint components, then the fact that one of them belongs to $ALL_r(f)$ implies that the other belongs to $ALL_w(f)$. A CTP does not possess this property, unless both C_1 and C_2 belong to Γ'' (see Lemma 4.4).

Lemma 4.1. *Let f be any CTP and consider any two disjoint components $C_1 \in \Gamma'$ and $C_2 \in \Gamma''$. Among the component states in $\Theta(C_2)$, only w^{C_2} can be concurrent with w^{C_1} .*

Proof. Since the coordinator is the first site to go into state p , if C_1 has all its sites, including the coordinator, in state w , then all the sites in C_2 must also be in state w . \square

This lemma has two important implications:

Property One: For any component C_1 in Γ' , if its current state is w^{C_1} then no other site can be in state p .

Property Two: For any component C_2 in Γ'' , if its current state contains a site in state p , then the coordinator must be in state p .

Due to the above two properties of a CTP, Lemma 2.3 does not apply to CTP. The corresponding lemmas for CTP's are proved below as Lemmas 4.2 and 4.3.

Lemma 4.2. *For any CTP f and two disjoint components $C_1 \in \Gamma'$ and $C_2 \in \Gamma''$, if C_1 has two component states S_1, S_2 that $f(S_1) = com$, $f(S_2) = ab$, then $f(w^{C_2}) = wa$.*

Proof. No matter how many sites of S_1 and S_2 are in states p or w , they are concurrent with the component state w^{C_2} . Therefore the consistency condition requires w^{C_2} to wait under f . \square

Note that, if f was a DTP then it follows from Lemma 2.3 that C_2 would have to wait under f not only in w^{C_2} but also in all other states as well. The following example shows that in general this is not the case for a CTP.

Example 4.1.

Let $I = \{1, 2, 3\}$ be the set of sites and define a CTP f as follows.

- (1) For every component C that contains the coordinator site 1 and any component state S of C , if $p \in \text{state}(S)$, let $f(S) = \text{com}$; otherwise let $f(S) = \text{ab}$.
- (2) For every component C that does not contain the coordinator and any component state S of C , if $p \in \text{state}(S)$, let $f(S) = \text{com}$; otherwise let $f(S) = \text{wa}$.

(2) above does not cause inconsistency because of Property Two. Consider two components $C_1 = \{1\}$ and $C_2 = \{2, 3\}$. It is clear that w^{C_2} is the only state of C_2 that waits under f . \square

Lemma 4.3. *For any CTP f and two disjoint components $C_1 \in \Gamma$ and $C_2 \in \Gamma^*$, if C_2 has two component states S_1, S_2 such that $f(S_1) = \text{com}$, $f(S_2) = \text{ab}$, then $f(S) = \text{wa}$ for every state $S \in \Theta_p(C_1)$.*

Proof. Suppose $S \in \Theta_p(C_1) = \Theta(C_1) - \{w^{C_1}\}$. Since it contains the coordinator and the coordinator must be in state p , therefore S can occur concurrently with any component state in $\Theta(C_2)$, in particular, S_1 and S_2 . Since these two component states are mapped to conflicting decisions by f , S must wait under f . \square

If the component C_2 in Lemma 4.3 belongs to $ALL_r(f)$, then $f(S) = \text{wa}$ for every component state $S \in \Theta_p(C_1)$. The following example shows that there exists a CTP f such that $f(w^{C_1}) \neq \text{wa}$.

Example 4.2. Let $I = \{1, 2\}$ be the set of sites, where site 1 is the coordinator. Define a CTP f as follows:

- (1) $f(\{(1, p)\}) = \text{wa}$ and $f(\{(1, w)\}) = \text{ab}$.
- (2) $f(\{(2, p)\}) = \text{com}$ and $f(\{(2, w)\}) = \text{ab}$.

Consider two components $C_1 = \{1\}$ and $C_2 = \{2\}$. According to (1), f maps w^{C_1} to ab and the other states of C_1 to wa . Note that the component states $\{(1, w)\}$ and $\{(2, p)\}$ are not concurrent and this makes it possible to map them to ab and com , respectively. \square

The above two lemmas show a crucial difference between a CTP and a DTP. The following lemma shows that there is also some similarity between them.

Lemma 4.4. *For any CTP f and two disjoint components $C_1, C_2 \in \Gamma$, if C_1 has two states S_1, S_2 such that $f(S_1) = com$, $f(S_2) = ab$, then $f(S) = wa$ for every state of C_2 .*

Proof. Since the two components C_1 and C_2 do not contain the coordinator, the proof of Lemma 2.3 carries over. \square

Lemma 4.5. *For any TP f and any component C , if $f(w^C) = ab$ and $f(p^C) = com$, then there exists a TP g such that $g \ll f$ and $C \in ALL_f(g)$.*

Proof. For every component state S_i that is concurrent with w^C and p^C , the consistency condition requires $f(S_i) = wa$. Therefore f can be modified to g in the following way. If $S \in \Theta(C)$ waits under f , let $g(S) = com$. This is consistent with the value of $f(S_i)$. Therefore $g \ll f$ and $C \in ALL_f(g)$. \square

Theorem 4.1. *Any site optimal CTP f has the property that $f(w^C) = ab$ for all $C \in \Gamma$.*

Proof. If f does not have the property, i.e., if $f(w^C) = wa$ for some C , then it can be modified to $g \ll f$ by defining $g(w^C) = ab$. This modification will not introduce any inconsistency, because w^C contains the coordinator and is concurrent with only those component states that contain all sites in state w and which therefore cannot be terminated to com . A contradiction, since $E(g) < E(f)$. \square

Theorem 4.1 implies that, in the search for site optimal CTP's, without loss of generality, we may assume that CTP's have the property mentioned in the theorem, i.e., $f(w^C) = ab$ for all $C \in \Gamma$.

4.2. Size-Based Centralized Termination Protocols

In this section, we introduce the *size-based CTP* and investigate site optimal CTP's in this class (see Section 4.3).

In the decentralized case, no component can be aborted by any TP if all its sites are in state p (see Lemma 2.1). In the centralized case, however, if such a component state contains the coordinator, it can be aborted by a TP as shown in the next lemma.

Lemma 4.6. *Each CTP f must satisfy the following two properties.*

- (1) *For every $C \in \Gamma$, $f(w^C) \neq com$ holds, but $f(p^C)$ can take any of the three values com , wa , and ab , and*
- (2) *for every $C \in \Gamma''$, $f(w^C) \neq com$ and $f(p^C) \neq ab$.*

Proof. The coordinator is always the first site to go into a new state. When it is in state p , no other site could be in state c , and therefore a component containing site 1 could be aborted if it is in state p , i.e., it is possible for a CTP f to have $f(p^C) = ab$ for some component C in Γ . The rest of the lemma follows from Lemma 2.1. \square

Recall that Γ_k is the set of components of size k . Let Γ_k^+ denote the set of components of size k containing the coordinator, i.e., the intersection of Γ_k and Γ^+ . Similarly, let Γ_k'' denote the intersection of Γ_k and Γ'' . Recall also that $\Theta_p(C)$ is the set of component states in $\Theta(C)$ which have at least one site in state p and $\Theta_w(C)$ is the set of component states in $\Theta(C)$ which have at least one site in state w .

Definition 4.1. A CTP f is said to be a *size-based CTP* if it satisfies the following two conditions. Let k be any positive integer less than n .

- (1) If a component $C \in \Gamma_k^+$ has a state $S \in \Theta_p(C)$ such that $f(S) \neq wa$, then for any $C_i \in \Gamma_k^+$ has a state $S_i \in \Theta_p(C_i)$ such that $f(S_i) \neq wa$.
- (2) If a component $C \in \Gamma_k''$ has a state $S \in \Theta(C)$ such that $f(S) \neq wa$, then for any $C_i \in \Gamma_k''$ has a state $S_i \in \Theta(C_i)$ such that $f(S_i) \neq wa$. \square

The definition of size-based CTP is similar to that of size-based DTP (see Section 3.2) except that we consider Γ^+ and Γ^- separately. In (1) of Definition 4.1, we only consider component states in $\Theta_p(C_i)$ instead of $\Theta(C_i)$. Since we assume that the component C_i in state w^{C_i} is always aborted. (See Theorem 4.1).

In the following $WA(f)$ denotes the set of all component states in Θ that a CTP f maps to wa , i.e., $WA(f) = \{S \mid f(S) = wa\}$.

Lemma 4.7. *For any size-based CTP f and any positive integer $k < n$, if some component $C \in \Gamma_k^+$ has a state $S \in \Theta_p(C)$ such that $f(S) \neq wa$, then there exists a size-based CTP $g \ll f$ such that $\Gamma_k^+ \subseteq ALL_i(g)$.*

Proof. Because of Lemma 4.5, without loss of generality, we may assume that if C does not belong to $ALL_i(f)$, then either $f(w^C) = wa$ or $f(p^C) = wa$. Also, because of Theorem 4.1, without loss of generality, we may assume that $f(w^C) = ab$ for all $C \in \Gamma^+$.

If some component $C \in \Gamma_k^+$ has a state $S \in \Theta_p(C)$ such that $f(S) \neq wa$, then it follows from the definition of size-based CTP that for any C_i in Γ_k^+ , either $com \in f(\Theta_p(C_i))$ or $ab \in f(\Theta_p(C_i))$ or both. Since all the component states of $\Theta_p(C_i)$ have the same set of concurrent component states, i.e., for any two component states $S_1, S_2 \in \Theta_p(C_i)$, S_1 is concurrent with a component state S_3 iff S_2 is concurrent with S_3 , we can modify f to g as follows.

For all component states $S \in \Theta_p(C_i) \cap WA(f)$, if $com \in f(\Theta_p(C_i))$, then let $g(S) = com$; otherwise, i.e., if $ab \in f(\Theta_p(C_i))$, let $g(S) = ab$. Then g terminates C_i if it is in any states in $\Theta_p(C_i)$, and since $f(w^{C_i}) = ab$, we have $\Gamma_k^+ \subseteq ALL_i(g)$. Since $f(S) \neq wa$ implies $g(S) \neq wa$ for any component state S , it follows that $g \ll f$. \square

Recall the set $PONLY_{\neq wa}(f)$ defined in Section 2.5, which consists of components C such that the TP f terminates C unless it is in state p^C . Similarly, let $PONLY_i(f)$ be the set of components C such that the TP f makes C wait unless it is in state \bar{p}^C , i.e., $PONLY_i(f) = \{C \in \Gamma \mid f(p^C) \neq wa \text{ and for all } S \in \Theta_-(C), f(S) = wa\}$. $WONLY_i(f)$ is defined

analogously by replacing $\Theta_r(C)$ and p^C by $\Theta_p(C)$ and w^C , respectively.

Lemma 4.8. *Given a size-based CTP f , there exists a size-based CTP $g \ll f$ such that for every positive integer $k < n$ either $\Gamma_k \subseteq ALL_r(g)$ or $\Gamma_k \subseteq WONLY_r(g)$.*

Proof. Because of Lemma 4.5, without loss of generality, we may assume that if C does not belong to $ALL_r(f)$, then either $f(w^C) = wa$ or $f(p^C) = wa$. Also, by Theorem 4.1, without loss of generality, we may assume that $f(w^C) = ab$ for all $C \in \Gamma$.

For every positive integer $k < n$, if Γ_k is not a subset of $WONLY_r(f)$, then there exists a component $C \in \Gamma_k$ that has a state $S \in \Theta_p(C)$ with $f(S) \neq wa$. It follows from Lemma 4.7 that f can be modified to a size-based CTP g such that $\Gamma_k \subseteq ALL_r(g)$. Hence the size-based CTP g has the required property. \square

Recall the definition of a component state as a set of (site, state) pairs. If S is a subset of a component state S_1 , we say that S is a *component substate* of S_1 .

Theorem 4.2. *For any given size-based CTP f , there exists a size-based CTP $g \ll f$ which has the following property.*

There exists a nonnegative integer s ($0 \leq s < n$) such that

- (1) *For each integer k ($1 \leq k \leq s$), $\Gamma_k \subseteq WONLY_r(g)$, and*
- (2) *For each integer k ($s < k \leq n-1$), $\Gamma_k \subseteq ALL_r(g)$.*

Proof. It follows from Lemma 4.8 that there exists a size-based CTP h such that $h \ll f$ and for each positive integer $k \leq n-1$, either $\Gamma_k \subseteq WONLY_r(h)$ or $\Gamma_k \subseteq ALL_r(h)$.

If $\{r \mid \Gamma_r \subseteq ALL_r(h)\}$ is nonempty, let $s = \min\{r \geq 1 \mid \Gamma_r \subseteq ALL_r(h)\} - 1$; otherwise, let $s = n-1$. It is clear that $\Gamma_{s+1} \subseteq ALL_r(h)$ and if $s \geq 1$ then by Lemma 4.8, for each integer k ($1 \leq k \leq s$), $\Gamma_k \subseteq WONLY_r(h)$. Suppose k is an integer, if any, such that $s+1 < k \leq n-1$ and $\Gamma_k \subseteq WONLY_r(h)$. If there is no such k then the proof is complete. If there is such a k then by Lemma 4.8, $h(S) = wa$ for every $S \in \Theta_r(C)$ such that $C \in \Gamma_k$. Let S be any state of a component which contains $s+1$ sites including the coordinator and let S_1 be a substate of S . By the definition

of the constant s we have $h(S_1) \neq wa$. Note that any component state S_2 that is concurrent with S_1 is also concurrent with S_1 , hence we can modify h to g by defining $g(S) = h(S_1)$. If $g(S)$ is defined this way for all component states $S \in \Theta_r(C)$ such that $C \in \Gamma_k$, then $\Gamma_k \subseteq ALL_r(g)$. \square

Having investigated the components which contain the coordinator, we now turn our attention to those which do not contain the coordinator.

Lemma 4.9. *For any size-based CTP f , and each positive integer $k < n$, if there exists $C \in \Gamma_k$ such that $\{com, ab\} \subseteq f(\Theta(C))$, then there exists a size-based CTP $g \ll f$ such that $\Gamma_k \subseteq ALL_r(g)$.*

Proof. Because of Lemma 4.5, without loss of generality, we may assume that if C does not belong to $ALL_r(f)$, then either $f(w^C) = wa$ or $f(p^C) = wa$. Also because of Theorem 4.1, without loss of generality, we may assume that $f(w^C) = ab$ for all $C \in \Gamma$.

Let $C \in \Gamma_k$ be a component such that $\{com, ab\} \subseteq f(\Theta(C))$. Let $C_1 \subseteq I - C$ contain the coordinator and let $C_2 \subseteq I - C - \{1\}$. Since $\{com, ab\} \subseteq f(\Theta(C))$, it follows from Lemma 4.3 and the assumption in the previous paragraph that $C_1 \in WONLY_r(f)$. Since $1 \leq |C_1| \leq n-k$, $\Gamma_r \subseteq WONLY_r(f)$ for all r ($1 \leq r \leq n-k$).

Similarly it follows from Lemma 4.4 that $C_2 \in ALL_{n-1}(f)$. Since $1 \leq |C_2| \leq n-k-1$, by Definition 4.1, we have $\Gamma_r \subseteq ALL_{n-1}(f)$ for all r ($1 \leq r \leq n-k-1$).

We now modify f to $g \ll f$ as follows. For any component state $S \in \Theta(C) \cap WA(f)$, if $p \in \text{state}(S)$, let $g(S) = com$; otherwise, i.e., if $\text{state}(S) = \{w\}$, let $g(S) = ab$.

If the above modification is repeated for all components C in Γ_k , then we have $g \ll f$ and $\Gamma_k \subseteq ALL_r(g)$. \square

If the condition of Lemma 4.9 does not hold, namely, if there is no $C \in \Gamma_k$ such that $\{com, ab\}$ is a subset of $f(\Theta(C))$, can f be modified to a "better" size-based CTP? The following lemma shows that it is still possible, although the result is not as good as that of Lemma 4.9: Γ_k can be contained in $PONLY_{n-1}(g) \cup WONLY_{n-1}(g)$ but not in $ALL_r(g)$.

Lemma 4.10. For any size-based CTP f and each integer k ($1 \leq k < n$), if there is no component $C \in \Gamma_k^{\text{wa}}$ such that $\{com, ab\}$ is a subset of $f(\Theta(C))$, but if there is a component $C \in \Gamma_k^{\text{wa}}$ has a state S such that $f(S) \neq wa$, then there exists a size-based CTP g such that $g \ll f$ and $\Gamma_k^{\text{wa}} \subseteq \text{PONLY}_{wa}(g) \cup \text{WONLY}_{wa}(g)$.

Proof. Let S be as given in the lemma. It follows from Definition 4.1 that for any $C_i \in \Gamma_k^{\text{wa}}$, either ab or com belongs to $f(\Theta(C_i))$. Suppose $com \in f(\Theta(C_i))$ and let $S_i \in \Theta_p(C_i)$ be such that $f(S_i) = com$. For any $S_j \in \Theta_p(C_i) \cap \text{WA}(f)$, a component state S_k is concurrent with S_j iff it is concurrent with S_i . Since $f(S_i) = com$ implies that $f(S_k) \in \{com, wa\}$, it is possible to modify f to g on S_j by defining $g(S_j) = com$. Therefore we obtain $g(\Theta_p(C_i)) = \{com\}$. Since $\{com, ab\}$ is not a subset of $f(\Theta(C_i))$, we have $f(w^{C_i}) = wa$ and g inherits this value and so $C_i \in \text{WONLY}_{wa}(g)$.

Similarly, if $ab \in f(\Theta(C_i))$, f can be modified to g so that $C_i \in \text{PONLY}_{wa}(g)$. Hence $\Gamma_k^{\text{wa}} \subseteq \text{PONLY}_{wa}(g) \cup \text{WONLY}_{wa}(g)$. \square

Lemma 4.11. For any size-based CTP f and each integer k ($1 \leq k < n$), if $\Gamma_k^{\text{wa}} \cap \text{ALL}_{wa}(f) \neq \emptyset$, then $\Gamma_k^{\text{wa}} \subseteq \text{ALL}_{wa}(f)$.

Proof. Assume there exists $C \in \Gamma_k^{\text{wa}} - \text{ALL}_{wa}(f)$ and let $S \in \Theta(C)$ be such that $f(S) \neq wa$. It follows from Definition 4.1 that $\Gamma_k^{\text{wa}} \cap \text{ALL}_{wa}(f) = \emptyset$, a contradiction. Hence $\Gamma_k^{\text{wa}} \subseteq \text{ALL}_{wa}(f)$. \square

Theorem 4.3. For any given size-based CTP f , there exists a size-based CTP $g \ll f$ such that for each integer k ($1 \leq k < n$), one of the following three relations holds.

- (1) $\Gamma_k^{\text{wa}} \subseteq \text{ALL}_{wa}(g)$.
- (2) $\Gamma_k^{\text{wa}} \subseteq \text{PONLY}_{wa}(g) \cup \text{WONLY}_{wa}(g)$.
- (3) $\Gamma_k^{\text{wa}} \subseteq \text{ALL}(f)$.

Proof. It follows from Lemmas 4.9, 4.10 and 4.11. \square

The following theorem summarizes the main results of this section. The first part of this theorem comes from Theorem 4.2.

Theorem 4.4. For any size-based CTP f , there exists a size-based CTP $g \ll f$ which has the following property.

(1) There exists an integer s ($0 \leq s < n$) such that

Case 1a. for each integer k ($1 \leq k \leq s$), $\Gamma_k^+ \subseteq \text{WONLY}_s(g)$, and

Case 1b. for each integer k ($s < k < n$) $\Gamma_k^+ \subseteq \text{ALL}_s(g)$.

(2) There exists an integer b ($\max\{n-s-1, (n-1)/2\} < b < n$) such that

Case 2a. for each integer k ($1 \leq k < n-b$), $\Gamma_k^+ \subseteq \text{ALL}_{s+b}(g)$,

Case 2b. for each integer k ($n-b \leq k < b$), $\Gamma_k^+ \subseteq \text{ALL}_{s+b}(g)$ or $\Gamma_k^+ \subseteq \text{PONLY}_{s+b}(g) \cup \text{WONLY}_{s+b}(g)$, and

Case 2c. for each integer k ($b \leq k < n$), $\Gamma_k^+ \subseteq \text{ALL}_s(g)$.

Proof. Part (1) follows from Theorem 4.2 and the constant s is as defined in Theorem 4.2. It remains to show the existence of b . Let $b = \min\{k : \Gamma_k^+ \subseteq \text{ALL}_s(g)\}$. Note that $b \geq n-s$. For, otherwise $(s+1) + b \leq n$ and there exist two concurrent component states S_1 and S_2 such that $|S_1| = s+1$, $|S_2| = b$ and S_1 contains the coordinator. Then $\Gamma_b^+ \subseteq \text{ALL}_s(g)$, and $\Gamma_{s+1}^+ \subseteq \text{ALL}_s(g)$, contradicting Lemma 4.3.

To prove that $b > (n-1)/2$, assume otherwise, i.e., $b \leq (n-1)/2$. Then there are two disjoint components, C_1 and C_2 , in Γ_b^+ . Since $\Gamma_b^+ \subseteq \text{ALL}_s(g)$, we have $C_1, C_2 \in \text{ALL}_s(g)$, contradicting Lemma 4.4.

To prove 2c, suppose that $k \geq b$ and Γ_k^+ is not a subset of $\text{ALL}_s(g)$. Let $S \in \Theta(C) \cap \text{WA}(g)$ for some component $C \in \Gamma_k^+$, and S_1 be a proper substate of S with $|S_1| = b$. Then $g(S_1) \neq wa$ by the definition of b . Note that any component state S_2 that is concurrent with S is also concurrent with S_1 , and therefore $g(S)$ can be changed to $g(S_1)$. This applies to every component state in $\Theta(C) \cap \text{WA}(g)$ for all $C \in \Gamma_k^+$. Hence, after g is modified, $\Gamma_k^+ \subseteq \text{ALL}_s(g)$.

To prove 2a, suppose $1 \leq k < n-b$ and $C \in \Gamma_k^+$. Let $C_1 = I - C - \{1\}$. Since $|C_1| = n-k-1 > b$, we have $C_1 \in \text{ALL}_s(g)$ from 2c. We thus have $g(\Theta(C)) = wa$ from Lemma 4.4. It

then follows from Theorem 4.3 that $\Gamma_k'' \subseteq ALL_{wa}(g)$.

Finally to prove 2b., suppose $n-b \leq k < b$ and $C \in \Gamma_k''$. Because of the minimality of b , Γ_k'' is not a subset of $ALL_r(g)$. It then follows from Theorem 4.3 that $\Gamma_k'' \subseteq ALL_{wa}(g)$ or $\Gamma_k'' \subseteq PONLY_{wa}(g) \cup WONLY_{wa}(g)$. \square

The structure of the size-based CTP g described in Theorem 4.4 is illustrated in Table 4.1.

Definition 4.2 A CTP f is a *standardized CTP* if it has the following property.

(1) There exists an integer s ($0 \leq s < n$) such that

Case 1a. for each integer k ($1 \leq k \leq s$), $\Gamma_k' \subseteq WONLY_r(g)$, and

Case 1b. for each integer ($s < k < n$) $\Gamma_k' \subseteq ALL_r(g)$.

(2) There exists an integer b ($\max\{n-s-1, (n-1)/2\} < b < n$) such that

Case 2a. for each integer k ($1 \leq k < n-b$), $\Gamma_k'' \subseteq ALL_{wa}(g)$,

k	Γ_k'	Γ_k''	k
1	$WONLY_r(g)$	$ALL_{wa}(g)$	1
.			.
.			.
.			.
s			.
$s+1$	$ALL_r(g)$	$ALL_{wa}(g)$ or $PONLY_{wa}(g) \cup WONLY_{wa}(g)$.
.			.
.			.
.			.
.			.
.			.
.	.	$b-1$	
.	.	b	
$n-1$.	$ALL_r(g)$	$n-1$

Table 4.1. Structure of the size-based CTP g in Theorem 4.3.

Case 2b. for each integer k ($n-b \leq k < b$), $\Gamma_k^* \subseteq ALL_{wa}(g)$ or
 $\Gamma_k^* \subseteq PONLY_{wa}(g) \cup WONLY_{wa}(g)$, and

Case 2c. for each integer k ($b \leq k < n$), $\Gamma_k^* \subseteq ALL_l(g)$.

Definition 4.2 is based on Theorem 4.4. With this definition, Theorem 4.4 can be restated as follows: for any size-based CTP f , there exists a standardized size-based CTP $g \ll f$.

4.3. Site Optimal Size-Based Centralized Termination Protocols

In Theorem 4.4, it was shown that a size-based CTP can be modified to a "better" size-based CTP which has the properties mentioned in the theorem, unless it already possesses those properties. As shown in Chapter 3, in the decentralized case, a site optimal size-based DTP can be found in the set of quorum-based DTP's. In this section, we define and investigate *quorum-based CTP's* and show that a site optimal size-based CTP exists among them.

For a given integer k ($0 \leq k < n/2$), define a CTP cp_k as follows. (Recall the definition of quorum-based DTP's, denoted by dp_k and dw_k .) Again, a component is treated differently depending on whether it contains the coordinator or not.

(1) Let S be the state of a component which contains the coordinator.

Case 1a. $1 \leq |S| \leq k$: If $p \in \text{state}(S)$, let $cp_k(S) = wa$; otherwise, i.e., if $\text{state}(S) = \{w\}$, let $cp_k(S) = ab$.

Case 1b. $k < |S| < n$: If $p \in \text{state}(S)$, let $cp_k(S) = com$; otherwise, i.e., if $\text{state}(S) = \{w\}$, let $cp_k(S) = ab$.

(2) Let S be the state of a component which does not contain the coordinator.

Case 2a. $1 \leq |S| \leq k-1$: Let $cp_k(S) = wa$.

Case 2b. $k \leq |S| < n-k$: If $p \in \text{state}(S)$, let $cp_k(S) = com$; otherwise, i.e., if $\text{state}(S) = \{w\}$, let $cp_k(S) = wa$.

Case 2c. $n-k \leq |S| \leq n-1$: If $p \in \text{state}(S)$, let $cp_k(S) = \text{com}$; otherwise, i.e., if $\text{state}(S) = \{w\}$, let $cp_k(S) = ab$. \square

The set of all cp_k 's is denoted by QCp . The following facts follow directly from the definition of cp_k . The first two are concerned with the components in Γ' and the last three with the components in Γ'' .

- (1) For all r ($1 \leq r \leq k$), $\Gamma_r' \subseteq \text{WONLY}_r(cp_k)$.
- (2) For all r ($k < r \leq n-1$), $\Gamma_r' \subseteq \text{ALL}_r(cp_k)$.
- (3) For all r ($1 \leq r < k$), $\Gamma_r'' \subseteq \text{ALL}_{w_0}(cp_k)$.
- (4) For all r ($k \leq r < n-k$), $\Gamma_r'' \subseteq \text{WONLY}_{w_0}(cp_k)$.
- (5) For all r ($n-k \leq r \leq n-1$), $\Gamma_r'' \subseteq \text{ALL}_r(cp_k)$.

The structure of cp_k ($k \geq 1$) is illustrated in Table 4.2, and the structure of cp_0 is illustrated in Table 4.3.

As an example, in Table 4.4, we list the values of cp_1 for $n = 4$. (Site 1 is the coordinator.)

There is another set of quorum-based CTP's denoted by cw_k ($0 \leq k < n/2$) defined as follows.

- (1) Let S be the state of a component which contains the coordinator.

Case 1a. $1 \leq |S| \leq k$: If $p \in \text{state}(S)$, let $cw_k(S) = wa$; otherwise, i.e., if $\text{state}(S) = \{w\}$, let $cw_k(S) = ab$.

Case 1b. $k < |S| < n$: Let $cw_k(S) = ab$.

- (2) Let S be the state of a component which does not contain the coordinator.

Case 2a. $1 \leq |S| \leq k-1$: Let $cw_k(S) = wa$.

Case 2b. $k \leq |S| < n-k$: If $w \in \text{state}(S)$, let $cw_k(S) = ab$; otherwise, i.e., if $\text{state}(S) = \{p\}$, let $cw_k(S) = wa$.

r	Γ_r	Γ_r^{**}	r
1	$WONLY_r(cp_k)$	$ALL_{wa}(cp_k)$	1
.			.
.			.
k-1			k-1
k	$ALL_r(cp_k)$	$WONLY_{wa}(cp_k)$	k
k+1			k+1
.			.
.			.
n-k	$ALL_r(cp_k)$	$ALL_r(cp_k)$	n-k
.			.
.			.
n-1			n-1

Table 4.2. Structure of a quorum-based CTP cp_k .

r	Γ_r	Γ_r^{**}	r
1	$ALL_r(cp_0)$	$WONLY_{wa}(cp_0)$	1
.			.
.			.
.			.
n-1	$ALL_r(cp_0)$	$WONLY_{wa}(cp_0)$	n-1
.			.
.			.
.			.

Table 4.3. Structure of the quorum-based CTP cp_0 .

component state	decision	component state	decision	component state	decision
site 1 2 3 4		site 1 2 3 4		site 1 2 3 4	
p - - -	wa	p - - p	com	p w - w	com
- p - -	com	p - - w	com	p - p p	com
- - p -	com	- p p -	com	p - p w	com
- - - p	com	- p w -	com	p - w p	com
w - - -	ab	- w p -	com	p - w w	com
- w - -	wa	- p - p	com	- p p p	com
- - w -	wa	- p - w	com	- p p w	com
- - - w	wa	- w - p	com	- p w p	com
w w - -	ab	- - p p	com	- w p p	com
w - w -	ab	- - p w	com	- p w w	com
w - - w	ab	- - w p	com	- w p w	com
- w w -	wa	p p p -	com	- w w p	com
- w - w	wa	p p w -	com	w w w -	ab
- - w w	wa	p w p -	com	w w - w	ab
p p - -	com	p w w -	com	w - w w	ab
p w - -	com	p p - p	com	- w w w	ab
p - p -	com	p p - w	com		
p - w -	com	p w - p	com		

Table 4.4. The decisions of cp_1 when $n = 4$ sites.

Case 2c. Suppose $n-k \leq S \leq n-1$. If $w \in \text{state}(S)$, let $cw_k(S) = ab$; otherwise, i.e., if $\text{state}(S) = \{p\}$, let $cw_k(S) = com$. \square

The set of all cw_k 's is denoted by QCw . Note that after the coordinator has broadcast "commit" messages, it is the first site to go into state c. If the coordinator is still in state p, no site could be in state c, and therefore a component which contains the coordinator can be aborted even if it has all its sites in state p. This makes 1b. of the above definition possible. In Table 4.5, we illustrate the structure of cw_k .

The union of QCp and QCw is denoted by QC . Thus QC is the set of all quorum-based CTP's.

Recall that PC_i is the sum of the probabilities of all components of size i . (See Section 3.3).

PC_i^c denote the sum of the probabilities of all components of size i that contain the

$$E(cp_k) = \sum_{i=1}^k i PC_i^*(P_i - P(0, i, i)) + \sum_{i=1}^{k-1} i PC_i^{**} P_i + \sum_{i=k}^{n-k-1} i PC_i^{**} P(0, i, i).$$

Proof. Suppose $k = 0$.

For any integer i ($1 \leq i < n-1$), and for any component $C \in \Gamma_i^*$, C waits under cp_0 only when it is in w^C . (See Table 4.3). The sum of the probabilities of these component states $\{w^C\}$ is given by the product of PC_i^* and $P(0, i, i)$. Hence

$$E(cp_0) = \sum_{i=1}^{n-1} PC_i^* P(0, i, i).$$

Suppose $k > 0$. First we consider components in Γ^* . For any integer i ($1 \leq i \leq k$) and for any component $C \in \Gamma_i^*$, C waits under cp_k unless it is in state w^C . The sum of the probabilities of the states in which components wait is given by the product of PC_i^* and $P_i - P(0, i, i)$.

We now consider the components in Γ^{**} . For any integer i ($1 \leq i \leq k-1$) and for any component $C \in \Gamma_i^{**}$, C always waits under cp_k . The sum of the probabilities of the states in which components wait is given by the product of P_i and PC_i^{**} . For any integer i ($k \leq i \leq n-k-1$) and for any component $C \in \Gamma_i^{**}$, C waits under cp_k only when it is in state w^C . The sum of the probabilities of the states in which components wait is given by the product of PC_i^{**} and $P(0, i, i)$.

Hence

$$E(cp_k) = \sum_{i=1}^k i PC_i^*(P_i - P(0, i, i)) + \sum_{i=1}^{k-1} i PC_i^{**} P_i + \sum_{i=k}^{n-k-1} i PC_i^{**} P(0, i, i).$$

The following similar result applies to a quorum-based CTP cw_k .

Theorem 4.6. For any integer k ($0 \leq k < n/2$), the expected number of waiting sites under the quorum-based CTP cw_k is given by the following formulae:

$$E(cw_0) = \sum_{i=1}^{n-1} i PC_i^{**} P(i, 0, i),$$

and for $k > 0$.

$$E(cw_k) = \sum_{i=1}^k i PC_i^*(P_i - P(0, i, i)) + \sum_{i=1}^{k-1} i PC_i^* P_i + \sum_{i=k}^{n-k-1} i PC_i^{**} P(i, 0, i). \square$$

Theorem 4.7. *If $P(0, k, k) \leq P(k, 0, k)$ for all integers k ($1 \leq k \leq n-1$), then for any size-based CTP f , there always exists a quorum-based CTP cp_i ($1 \leq i < n/2$) such that $E(cp_i) \leq E(f)$.*

Proof. It follows from Theorem 4.4 that there exists a size-based CTP h such that $h \ll f$ having the following properties.

(1) There exists an integer s ($0 \leq s < n$) such that

Case 1a. for any integer k ($1 \leq k \leq s$), $\Gamma_k^* \subseteq \text{WONLY}_i(h)$, and

Case 1b. for any integer ($s < k < n$) $\Gamma_k^* \subseteq \text{ALL}_i(h)$.

(2) There exists an integer b ($\max\{n-s-1, (n-1)/2\} < b < n$) such that

Case 2a. for any integer k ($1 \leq k < n-b$), $\Gamma_k^{**} \subseteq \text{ALL}_{w_d}(h)$,

Case 2b. for any integer k ($n-b \leq k < b$), $\Gamma_k^{**} \subseteq \text{ALL}_{w_d}(h)$ or $\Gamma_k^{**} \subseteq \text{PONLY}_{w_d}^*(h) \cup \text{WONLY}_{w_d}(h)$, and

Case 2c. for any integer k ($b \leq k < n$), $\Gamma_k^{**} \subseteq \text{ALL}_i(h)$.

We want to compare the expected numbers of waiting sites between h and the quorum-based CTP $cp_{\bar{b}}$, where $\bar{b} = n - b$. Since $s + b \geq n$, we have $\bar{b} \leq s$.

We first consider the components in Γ^* .

Case A. $1 \leq k \leq \bar{b}$: Since $\bar{b} \leq s$, it follows from 1a. that if $S \in \Theta_p(C)$ for some $C \in \Gamma_k^*$ then $S \in \text{WA}(h)$ and $h(w^C) \neq wa$. It follows from the way $cp_{\bar{b}}$ is defined that $S \in \text{WA}(cp_{\bar{b}})$ and $cp_{\bar{b}}(w^C) \neq wa$. Therefore, h and $cp_{\bar{b}}$ have the same number of expected waiting sites.

Case B. $\bar{b} < k \leq n-1$: For any $C \in \Gamma_k^*$, we have $\Theta(C) \cap \text{WA}(cp_{\bar{b}}) = \emptyset$, i.e., C never waits under $cp_{\bar{b}}$. Therefore, the expected number of waiting sites under $cp_{\bar{b}}$ is not larger than that under h .

Now we consider the components in Γ'' .

Case C. $1 \leq k < \bar{b}$: It follows from 2a. that $\Gamma_k'' \subseteq ALL_{w_0}(h)$. Since $\Gamma_k'' \subseteq ALL_{w_0}(cp_{\bar{b}})$, h and $cp_{\bar{b}}$ have the same number of waiting sites.

Case D. $\bar{b} \leq k < b$: It follows from 2b. that for any $C \in \Gamma_k''$, either w^C or $p^C \in WA(h)$. For $cp_{\bar{b}}$, w^C is the only waiting component state in $\Theta(C)$. Since $P(0, k, k) \leq P(k, 0, k)$, $Pr(w^C) \leq Pr(p^C)$. Hence the expected number of waiting sites under $cp_{\bar{b}}$ is not larger than that under h in this case.

Case E. $b \leq k \leq n-1$: It follows from 2c that $\Gamma_k'' \subseteq ALL_r(h)$. Also $\Gamma_k'' \subseteq ALL_r(cp_{\bar{b}})$ holds (see definition of $cp_{\bar{b}}$). Hence, in this case, h and $cp_{\bar{b}}$ have the same number of waiting sites.

In each of the above five cases, the expected number of waiting sites under $cp_{\bar{b}}$ is not larger than that under h , and therefore $E(cp_{\bar{b}}) \leq E(h)$. Since $h \ll f$, this implies that $E(cp_{\bar{b}}) \leq E(f)$. \square

By replacing $cp_{\bar{b}}$ by $cw_{\bar{b}}$, we have a parallel result.

Theorem 4.8. *If $P(k, 0, k) \leq P(0, k, k)$ for all integers k ($1 \leq k \leq n-1$), then for any size-based CTP f , there exists a quorum-based CTP $cw_{\bar{b}}$ ($1 \leq \bar{b} < n/2$) such that $E(cw_{\bar{b}}) \leq E(f)$. \square*

With the results of Theorems 4.7 and 4.8, we can compare all the quorum-based CTP's to find a site optimal size-based CTP.

Theorem 4.9. *If $P(0, k, k) \leq P(k, 0, k)$ for all integers k ($1 \leq k < n$), let m be an index such that*

$$E(cp_m) = \min\{E(cp_k) \mid 1 \leq k < n/2\}.$$

Then cp_m is a site optimal CTP in the set of size-based CTP's.

Proof. The theorem follows from Theorems 4.7 and 4.5. \square

Theorem 4.10. *If $P(k, 0, k) \leq P(0, k, k)$ for all integers k ($1 \leq k < n$), let m be an index such that*

$$E(cw_m) = \min\{E(cw_k) \mid 1 \leq k < n/2\}.$$

Then cw_m is a site optimal CTP in the set of size-based CTP's.

Proof. The theorem follows from Theorem 4.8 and 4.6. \square

This concludes our search for site optimal CTP's in the class of size-based CTP's.

4.4. Count-Based CTP's

Recall the definition of a count-based DTP given in Section 3.4. A count-based DTP maps any two state equivalent component states to the same decision. In this section we introduce an analogous concept in the centralized case.

Definition 4.3. A CTP f is a *count-based CTP*, if for any two state equivalent component states S_1 and S_2 , the following two conditions are satisfied.

- (1) If both S_1 and S_2 contain the coordinator, then $f(S_1) = f(S_2)$.
- (2) If neither S_1 nor S_2 contains the coordinator, then $f(S_1) = f(S_2)$.

An example of a count-based CTP is illustrated below in the Table 4.6.

Theorem 4.11. *Each count-based CTP is a size-based CTP.*

Proof. Let f be a count-based CTP and let k be an integer such that $1 \leq k \leq n-1$. Suppose that a component $C \in \Gamma_k^+$ has a state S in $\Theta_f(C)$ such that $f(S) \neq wa$. For an arbitrary component C_i in Γ_k^+ , let $S_i \in \Theta(C_i)$ be state equivalent to S . It follows from Definition 4.3 that $f(S_i) = f(S) \neq wa$. Therefore, any $C_i \in \Gamma_k^+$ has a state $S_i \in \Theta_f(C_i)$ such that $f(S_i) \neq wa$. Similarly, Definition 4.3 (2) implies Definition 4.1 (2). Hence f is also size-based. \square

By definition, a quorum-based CTP is also a count-based CTP. Hence we have the following two results.

component state	decision	component state	decision	component state	decision
site		site		site	
1 2 3 4		1 2 3 4		1 2 3 4	
p - - -	ab	p - - p	wa	p w - w	ab
- p - -	wa	p - - w	ab	p - p p	com
- - p -	wa	- p p -	wa	p - p w	wa
- - - p	wa	- p w -	ab	p - w p	wa
w - - -	ab	- w p -	ab	p - w w	ab
- w - -	wa	- p - p	wa	- p p p	ab
- - w -	wa	- p - w	ab	- p p w	wa
- - - w	wa	- w - p	ab	- p w p	wa
w w - -	ab	- - p p	wa	- w p p	wa
w - w -	ab	- - p w	ab	- p w w	ab
w - - w	ab	- - w p	ab	- w p w	ab
- w w -	ab	p p p -	com	- w w p	ab
- w - w	ab	p p w -	wa	w w w -	ab
- - w w	ab	p w p -	wa	w w - w	ab
p p - -	wa	p w w -	ab	w - w w	ab
p w - -	ab	p p - p	com	- w w w	wa
p - p -	wa	p p - w	wa		
p - w -	ab	p w - p	wa		

Table 4.6. An example of a count-based CTP for $n = 4$ sites.

Theorem 4.12. *Given any count-based CTP f , there exists a count-based CTP $g \ll f$ with the following property.*

(1) *There exists an integer s ($0 \leq s < n$) such that*

Case 1a. *for each integer k ($1 \leq k \leq s$), $\Gamma_k \subseteq \text{WONLY}_f(g)$, and*

Case 1b. *for each integer ($s < k < n$) $\Gamma_k \subseteq \text{ALL}_f(g)$.*

(2) *There exists an integer b ($\max\{n-s-1, (n-1)/2\} < b < n$) such that*

Case 2a. *for each integer k ($1 \leq k < n-b$), $\Gamma_k \subseteq \text{ALL}_{w_0}(g)$.*

Case 2b. *for each integer k ($n-b \leq k < b$), either $\Gamma_k \subseteq \text{ALL}_{w_0}(g)$ or $\Gamma_k \subseteq \text{PONLY}_{w_0}(g)$ or*

$\Gamma_k \subseteq \text{WONLY}_{w_0}(g)$, and

Case 2c. for each integer k ($b \leq k < n$), $\Gamma_k \subseteq ALL(g)$.

Proof. The proof follows from Theorem 4.4 and the proof of Theorem 3.10. \square

Theorem 4.13. For any count-based CTP f , there exists a quorum-based CTP $q \ll f$.

Proof. The proof follows from Theorem 4.12 and the proof of Theorem 3.11. \square

Hence, a site optimal count-based CTP can be found by comparing all the quorum-based CTP's. Recall that QC denotes the set of all quorum-based CTP's.

Theorem 4.14. Let q be a quorum based CTP such that

$$E(q) = \min\{E(q) \mid q \in QC\}.$$

Then q is a site optimal count-based CTP.

Proof. The proof follows from Theorem 4.13. \square

In the following, we show an example of a site optimal count-based CTP in a special case, where all component states have the same probability. In this particular case, as was explained in Section 2.5, during the calculation of the expected number of waiting sites under a CTP, we have only to calculate the number of waiting sites and then multiply it with the probability of a component state.

We list the values of cp_0 in Table 4.7. The values of cp_1 were listed in Table 4.4. From these two tables, it is seen that the number of waiting sites under cp_0 is 12, whereas the number of waiting sites under cp_1 is 10. In this particular case, $E(cp_1) = E(cp_0)$. Therefore, cp_1 is a site optimal count-based CTP.

Ramarao has proposed a "highly optimal" CTP h which he claimed was a site optimal CTP in the general case [RAMA-84]. It turns out that the CTP h is actually cp_0 . As was pointed out above, cp_0 is not site optimal. Thus the above cp_1 provides a counter example to his claim.

component state	decision	component state	decision	component state	decision
site		site		site	
1 2 3 4		1 2 3 4		1 2 3 4	
p - - -	com	w - - p	com	w p - w	com
- p - -	com	- p p -	com	w w - p	com
- - p -	com	- p w -	com	p - p p	com
- - - p	com	- w p -	com	p - p w	com
w - - -	ab	- p - p	com	p - w p	com
- w - -	wa	- p - w	com	w - p p	com
- - w -	wa	- w - p	com	p - w w	com
- - - w	wa	- - p p	com	w - p w	com
w w - -	ab	- - p w	com	w - w p	com
w - w -	ab	- - w p	com	- p p p	com
w - - w	ab	p p p -	com	- p p w	com
- w w -	wa	p p w -	com	- p w p	com
- w - w	wa	p w p -	com	- w p p	com
- - w w	wa	w p p -	com	- p w w	com
p p - -	com	p w w -	com	- w p w	com
p w - -	com	w p w -	com	- w w p	com
w p - -	com	w w p -	com	w w w -	ab
p - p -	com	p p - p	com	w w - w	ab
p - w -	com	p p - w	com	w - w w	ab
w - p -	com	p w - p	com	- w w w	wa
p - - p	com	w p - p	com		
p - - w	com	p w - w	com		

Table 4.7. The decisions of cp_0 for $n = 4$ sites.

4.5. Restricted Decentralized Termination Protocols

If S is a realizable component state in the centralized case, then it is also realizable in the decentralized case. The converse is not true in general. Consider a component state which contains the coordinator. If the coordinator is in state w and some other sites are in state p , then this component state is realizable in the decentralized case, but not in the centralized case. Hence, the set of realizable component states in the centralized case is a proper subset of the set of realizable component states in the decentralized case.

To see another difference between the two cases, consider two disjoint components C_1 and C_2 . In the decentralized case, any component state $S_1 \in \Theta(C_1)$ is concurrent with any component state $S_2 \in \Theta(C_2)$. This is no longer true in the centralized case. For example, if C_1 contains the coordinator, by Property One (see Section 4.1), $S_1 = w^{C_1}$ is not concurrent with $S_2 = p^{C_2}$. However, if S_1 and S_2 are two concurrent component states in the centralized case, then they are also concurrent in the decentralized case.

The above observations make it possible to apply a DTP to the centralized case by restricting its domain. Let f be a DTP and let R be the set of all realizable component states in the centralized case. Then R is a proper subset of the domain of f . If we restrict f to R , it satisfies both the nonreversal and consistency conditions. (See Definition 2.1). Therefore, we can consider f to be a CTP. CTPs obtained in this way are called *restricted decentralized terminated protocols (RDTP)*. Some members of this class have been regarded as possible candidates for a site optimal CTP [RAMA-84]. However, we show here that this is not true and, in fact, there always exists a CTP which is strictly better than any RDTP.

Lemma 4.12. *Given any CTP f , if there exist two disjoint components C_1 and C_2 such that $C_1 \in ALL_r(f)$ and $C_2 \notin ALL_{wa}(f)$, then f is not a RDTP.*

Proof. Suppose f is a RDTP. It follows from Lemma 2.3 that $C_2 \in ALL_{wa}(f)$, a contradiction. Hence f cannot be a RDTP. \square

The following lemma shows that for every RDTP, there exists a "better" CTP.

Theorem 4.15. *If $P(0, k, k) \leq P(k, 0, k)$ for all k ($1 \leq k \leq n-1$), then for any RDTP f , there exists a CTP g which is not a RDTP such that $E(g) < E(f)$.*

Proof. Because of Lemma 4.5, without loss of generality, we may assume that if C does not belong to $ALL_r(f)$, then either $f(w^C) = wa$ or $f(p^C) = wa$.

The proof is divided into three cases.

Case A: suppose $ALL_i(f) \cap \Gamma'' \neq \emptyset$. Let $C_1 \in ALL_i(f) \cap \Gamma''$ and let C_2 be the complementary component of C_1 , i.e., $C_2 = I - C_1$. Since $C_1 \in ALL_i(f)$ and f is a RDTP, it follows from Lemma 2.3 that $f(w^{C_2}) = wa$. Because C_2 contains the coordinator, only the component state w^{C_1} from among those in $\Theta(C_1)$ is concurrent with w^{C_2} . Since $f(w^{C_1}) = ab$, f can be modified to g by defining $g(w^{C_2}) = ab$. Therefore $E(g) < E(f)$, regardless of the relative values of $P(0, k, k)$ and $P(k, 0, k)$.

Case B: $ALL_i(f) \cap \Gamma'' = \emptyset$ but $ALL_i(f) \neq \emptyset$. Recall that $\Theta_p(C)$ denotes the set of states of C which contain at least one site in state p . Define a CTP h as follows.

- (1) For any $C \in \Gamma$ let $h(S) = com$ for all $S \in \Theta_p(C)$.
- (2) For all $C \in \Gamma'$, let $h(w^C) = ab$.
- (3) For all $C \in \Gamma''$, let $h(w^C) = wa$.

CTP h is defined above in such a way that it maps to wa only those states of components which do not contain the coordinator and have all their sites in state w .

No $C_i \in \Gamma''$ belongs to $ALL_i(f)$ and this implies that either $f(w^{C_i}) = wa$ or $f(p^{C_i}) = wa$. C_i waits under h only when it is in state w^{C_i} . Since $P(0, k, k) \leq P(k, 0, k)$ for every $k \leq n-1$, $\Pr(w^{C_i}) \leq \Pr(p^{C_i})$. Therefore the expected number of waiting sites in C_i under h is not larger than that under f . Hence $E(h) \leq E(f)$.

Because $ALL_i(f) \cap \Gamma' \neq \emptyset$, there exists a component C in Γ' such that $C \in ALL_{wa}(f)$. Since C waits under h only when it is w^C , more sites in C wait under f . Hence $E(h) < E(f)$.

Case C: $ALL_i(f) = \emptyset$. Since no $C \in \Gamma$ belongs to $ALL_i(f)$, therefore either $f(w^C) = wa$ or $f(p^C) = wa$. Again we compare f with the CTP h defined in Case B above. Since $\Pr(w^C) \leq \Pr(p^C)$ and a component in Γ' never waits under h , we have $E(h) < E(f)$. \square

—The following is a parallel result to Theorem 4.15 when $P(k, 0, k) \leq P(0, k, k)$.

Theorem 4.16. *If $P(k, 0, k) \leq P(0, k, k)$ for all k ($1 \leq k \leq n-1$), then for any RDTP f , there exists a CTP g which is not a RDTP such that $E(g) < E(f)$. \square*

Theorems 4.15 and 4.16 confirm the fact that no site optimal CTP can be found among the RDTP's.

CONCLUSION

The handling of network partitioning is in general a difficult problem. Most of the known systems treat it as a catastrophic failure and handle it manually. In this thesis, our main concern is to design protocols which maximizes the availability of a database in the presence of network partitioning. Transactions are normally executed under the three-phase commit protocol and a termination protocol (TP) is invoked only when a failure occurs.

We have extensively investigated two classes of TP's: count-based TP's and size-based TP's. It was shown that, in these classes, "best" TP's with the minimum expected number of waiting sites can be found among the quorum-based TP's.

The methodology used in the search for these "best" TP's was to introduce a partial order among all size-based TP's and to identify a subset which contained all candidates for the "best" TP's. The subset thus identified is the set of quorum-based TP's. We have also succeeded in demonstrating that this approach applies equally well to the decentralized and the centralized cases.

Along with the development of this methodology, characteristics of TP's were examined extensively. In particular, some of the essential characteristics of CTP's have been found which give us a better insight into the properties of CTP's.

REFERENCES

- [CHIN-83] Chin, F. and Ramarao, K.V.S. Optimal Termination Protocols for Network Partitioning. *Proc. 2nd ACM SIGACT-SIGMOD Symp. on Principles of Database Systems*, Mar, 1983.
- [GRAY-78] Gray, J. Notes on Database Operating Systems. *Operating Systems: An Advanced Course*, Lecture Notes in Computer Science 60, Springer-Verlag, N.Y. 1978, pp.393-481.
- [LAMP-76] Lampson, B. and Sturgis, H. Crash Recovery in a Distributed Storage System. *Tech. Report*. CS Lab, Xerox Parc, Palo Alto, California, 1976.
- [RAMA-84] Ramarao, K.V.S. Resilient Distributed Database System. *Ph.D. Thesis*, CS Dept., University of Alberta, 1984.
- [SKEE-81a] Skeen, D. and M. Stonebraker. A Formal Model of Crash Recovery in a Distributed System. *Proc. 5th Berkeley Workshop on Distributed Data Management and Computer Networks*, 1981, pp.129-142.
- [SKEE-81b] Skeen, D. Nonblocking Commit Protocol. *Proc. ACM SIGMOD Conf. on Management of Data*, 1981, pp. 133-142.
- [SKEE-81c] Skeen, D. A Decentralized Termination Protocol. *Memo. No. UCB/ERL MA 81/50*, EECS Dept., Univ. of Calif., Berkeley, 1981.
- [SKEE-82a] Skeen, D. Crash Recovery in Distributed Database Management System. *Ph.D. Thesis*. EECS Dept., Univ. of Calif., Berkeley, 1982.
- [SKEE-82b] Skeen, D. A Quorum-based Commit Protocol. *Computer Science T.R. 82-483*, Cornell Univ., 1982.

FIGURES

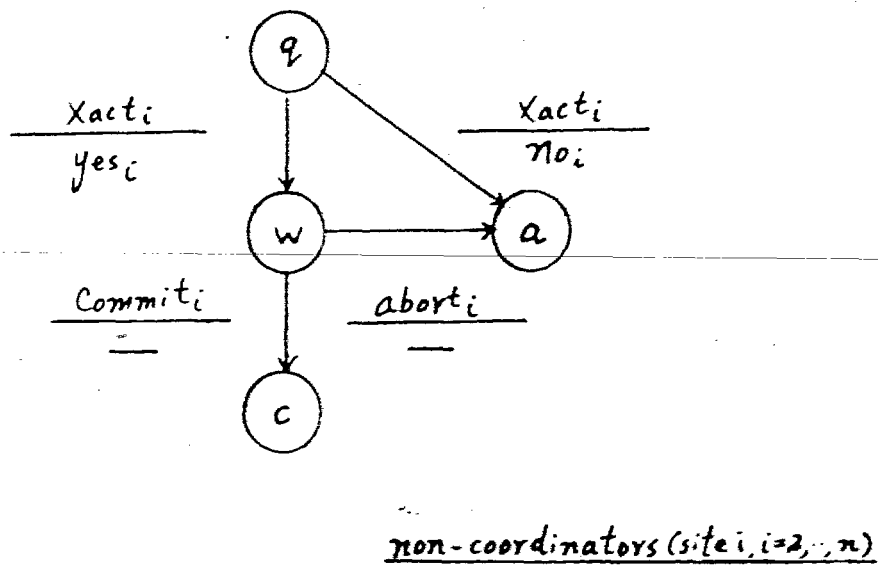
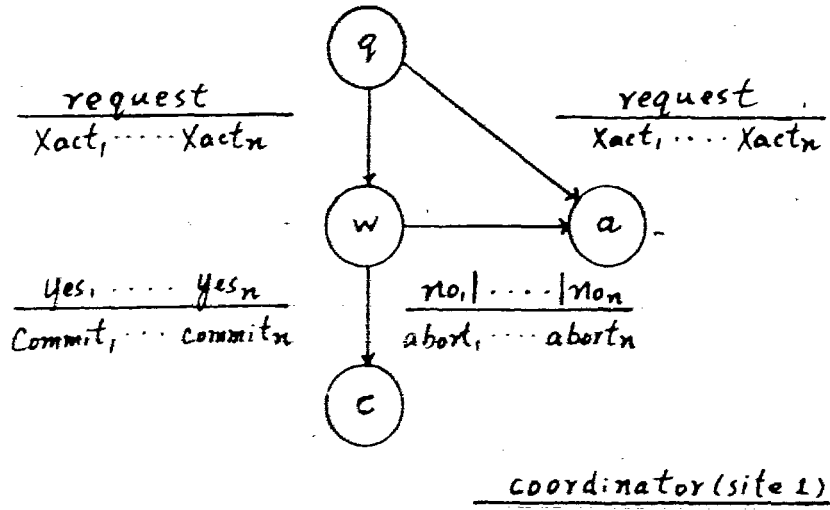


Figure 1.1. FSA of the Centralized Two-Phase Commit Protocol.

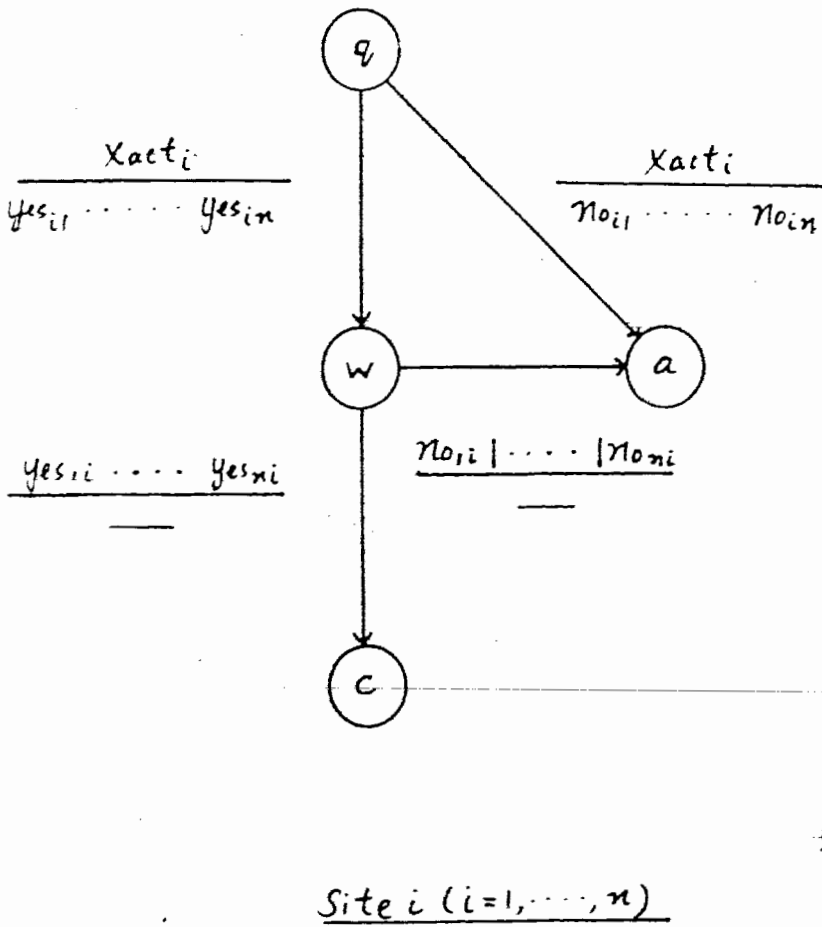


Figure 1.2. FSA of the Decentralized Two-Phase Commit Protocol.

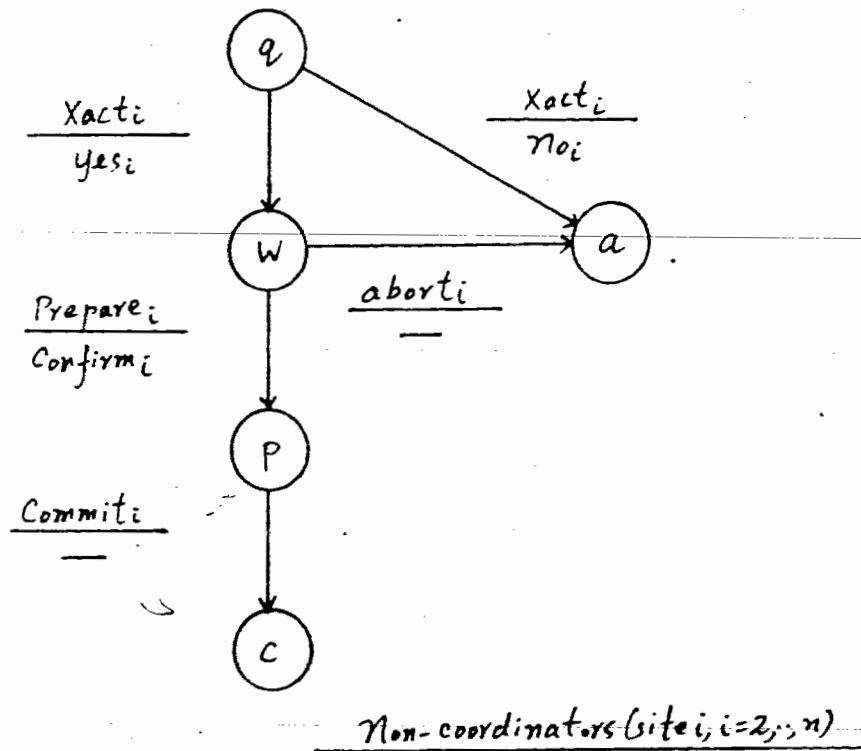
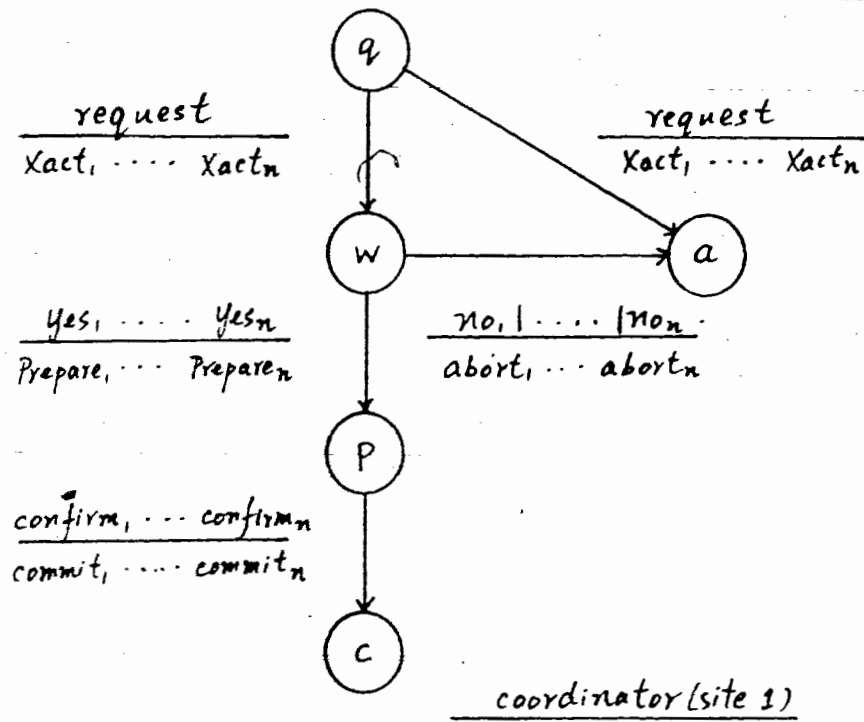
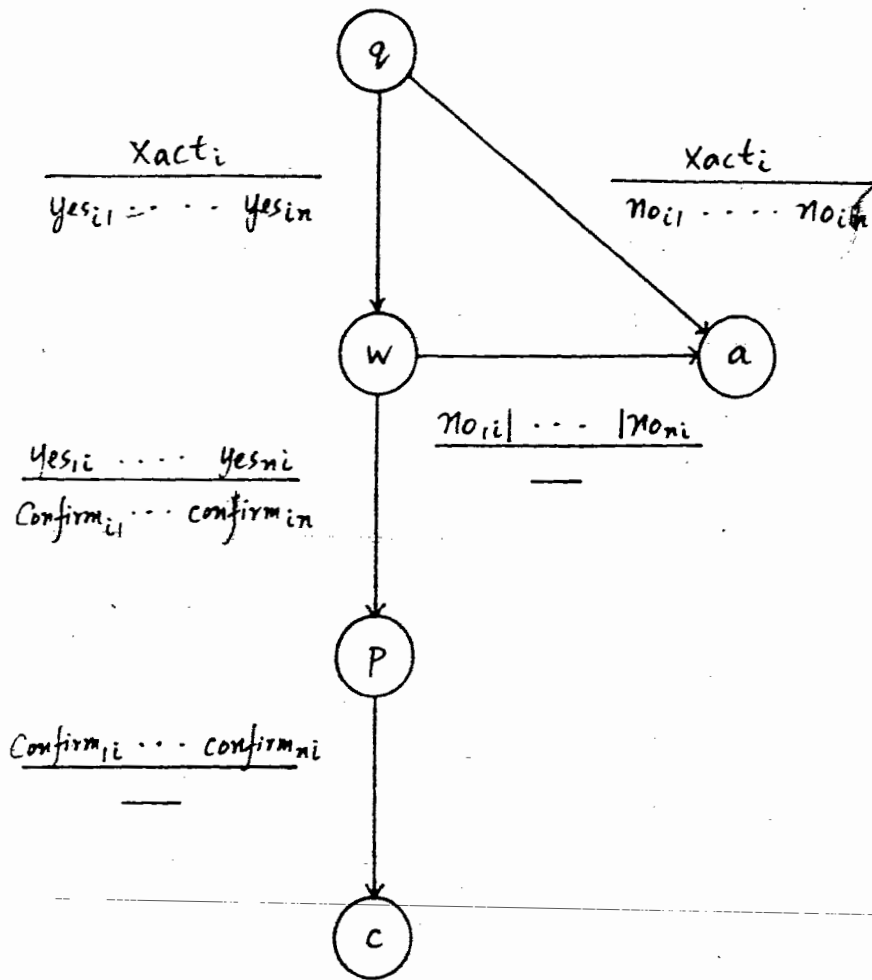


Figure 1.3. FSA of the Centralized Three-Phase Commit Protocol.



Site i ($i=1, \dots, n$)

Figure 1.4. FSA of the Decentralized Three-Phase Commit Protocol.