システムソフトウェアと オペレーティング・システム 74－9
マルチメディア通信と分散処理 81－9
（1997. 2. 27）

多重化データベースにおける site quorumを用いたデータの一貫性制御

千葉 佳史 多田 知正 樋口 昌宏 藤井 護

大阪大学基礎工学部

e-mail: {chiba, tada, higuchi, fujii}@ics.es.osaka-u.ac.jp

多重化データベースにおけるデータ一貫性制御アルゴリズムの一つとして仮想分割が提案されている．仮想分割では，障害によりデータベースシステムが互いに通信できない２つ以上の断片に分かれたとき，各データ項目が高々１つの断片で読み書き可能となるよう制御を行なう．各データ項目がどの断片において読み書きが可能となるかを決定する基準として data quorum と site quorum が考えられる．本研究では，仮想分割アルゴリズムを data quorum を基準として用いた場合と site quorum を基準として用いた場合について，ネットワーク分割が生じても実行可能となるトランザクションの比率に着目して比較を行なった．複製が各サイトにランダムに配置されている多重化データベースにおいては site quorum の方が実行可能なトランザクションの比率は大きくなる．さらに，２つの基準の差を定量的に評価するため，いくつかの多重化データベースモデルを設定し，その上でネットワーク分割をシミュレートし，実行可能なトランザクションの比率を求めた．その結果，各データ項目の複製数が少ない場合，及び多くのデータ項目に読み書きするトランザクションが多い場合に比率の差がより大きくなることがわかった．

Data consistency control using Site Quorum in Replicated Database

Yoshifumi Chiba, Harumasa Tada, Masahiro Higuchi and Mamoru Fujii

Faculty of Engineering Science,
Osaka University

Virtual Partition (VP) has been proposed as an algorithm to preserve database consistency. Under VP, when a database system is divided into two or more components by some failures and sites in each component cannot communicate with sites in other components, each data item can be read or written (is valid) in only one component. There are two criteria to determine the validity of each data item in each component — the data quorum and the site quorum. In this paper, we compare the data quorum with the site quorum as the criterion in VP by the ratio of executable transactions and unexecutable ones. In replicated databases in which replicas are distributed randomly, the ratio is larger under site quorum than under data quorum. Moreover, to evaluate the difference of two criteria quantitatively, we introduced some typical models of replicated database system, simulated the network partition on the models and measured the ratio of executable transactions. ¿From simulation results, we learned that the difference of two cirteria is larger in tha case that the number of replicas is small and the case that many transactions accesses many data items.

1 まえがき

多重化データベースとは各データ項目の複製が複数のサイトに配置された分散データベースである．複数のサイトに複製を配置することで障害に対するデータの可用性が高くなるという特徴がある．しかし，同じデータ項目の複製が複数存在するため各データ項目に対して複製の一貫性を維持する必要がある．複製の一貫性を維持する最も簡単な方法は，同じデータ項目に対する異なる内容を持つ２つ以上の複製が共に有効にならないようにすることである [1]．データ一貫性制御アルゴリズムの一つとして仮想分割が提案されている [1]．仮想分割においては複製の一貫性を維持するためにこの方法を用いる．また，この方法によりネットワーク分割時にデータの一貫性を維持するためには，各データ項目は高々１つの断片でのみ読み書き可能とする必要がある．

ネットワーク分割時に各データ項目がどの断片で読み書き可能となるか決定する方法として data quorum が提案されている [1]. data quorum において，トランザクションは読み書きを行なうデータ項目 i について，アクセス可能な i の複製の数を調べる．その数が i の全複製数の過半数 (quorum) を越えている時のみ，i への読み書きを行なう．別の方法として，過半数のサイトにアクセス可能なサイトではトランザクションはアクセス可能なすべてのデータ項目への読み書きを可能とする方法が提案されている．この方法は site quorum と呼ばれる [1]．

仮想分割においては，障害検出時に data quorum と site quorum のどちらかを用いて各データ項目について読み書き可能かどうか調べる．[1]．そこで本研究では，data quorum を用いた場合と site quorum を用いた場合についてネットワーク分割が生じても実行不可能とならないトランザクションの比率について比較を行なった．各データ項目の複製（全データ項目同数）が各サイトに等確率で配置され，各トランザクションが各データ項目に等確率で読み書きし各サイトに等確率で発生する多重化データベースモデルにおいては，site quorum の方が大きくなる．さらに，data quorum と site quorum のネットワーク分割時に実行可能なトランザクションの比率の差を定量的に評価するため，いくつかの多重化データベースモデル上でネットワーク分割のシミュレーションを行ないその比率を求めた．

本稿では以下，２節において多重化データベースについて，３節において data quorum と site quorum の比較について，４節においてネットワーク分割のシミュレーションについて，５節においてまとめを述べる．

2 多重化データベース

2.1 本研究における多重化データベースモデル

本研究では以下，多重化データベースのネットワークは同時に３つ以上の断片に分かれないとし，各データ項目の複製（全データ項目同数とする）がそれぞれ各サイトに等確率で配置され，また各トランザクションが全てのデータ項目に等確率で読み書きし（全トランザクション同データ項目数のデータ項目に読み書きする），全サイトに等確率で発生すると仮定する．

読み書きされるべきデータ項目がすべて読み書き可能であるトランザクションを実行可能トランザクションとし，多重化データベース中の全トランザクション数に対する実行可能トランザクション数の比率を実行可能トランザクション比率と呼ぶ．

2.2 多重化データベース

多重化データベースとは各データ項目の複製が複数のサイトに配置されている分散データベースであり，複数のサイトに複製を配置することで障害に対するデータの可用性が高くなるという特徴がある．しかし，同じデータ項目の複製が複数存在するため，各データ項目の複製間の一貫性を維持する必要がある．このことはデータの一貫性制御と呼ばれる．データ一貫性制御アルゴリズムの一つとして仮想分割が提案されている [1]．

データの一貫性制御の最も簡単な方法は，異なる内容を持つ２つ以上の複製が共に有効にならないようにすることである [1]．仮想分割では，この方法が用いられている．この方法では，データの一貫性制御は次のように行なわれる．各トランザクションによるデータ項目の読み出しは任意の１つの複製から値を読み出し，データ項目への書き込みは，書き込みを行なうデータ項目の全ての複製に対して行なう．

また多重化データベースにおいては，サイト障害，コミュニケーションリンク障害 (link failure) が生じた場合，データベースシステムが２つの断片 (component) に分割され得る．１つの断片に存在する任意の２つのサイト間では通信可能であるが，異なる断片に存在するサイト間では通信不可

能となる．ネットワークが２つの断片に分割されることはネットワーク分割と呼ばれる．上で述べたデータの一貫性制御の方法においては，ネットワーク分割時にはデータの一貫性を保つために各データ項目は高々１つの断片でのみ読み書き可能とする必要がある．そのため，各トランザクションはデータ項目に読み書きする時にそのデータ項目が読み書き可能であることを確認しなければならない．

2.3 data quorum と site quorum

トランザクションが各データ項目について読み書き可能かどうかを調べる方方法として data quorum が提案されている [1] ．data quorum ではトランザクションがデータ項目 i への読み書きが可能であるかどうかを次のように調べる．

1. アクセス可能な i の複製の数を調べる．

2. アクセス可能な i の複製の数が，i の全複製数の過半数の場合のみ読み書き可能となる．

data quorum では，ネットワーク分割時に各データ項目はそのデータ項目の過半数の複製が存在する断片でのみ読み書き可能となる．

別の方法として，site quorum が提案されている [1] ．site quorum では，トランザクションがデータ項目 i への読み書きが可能であるかどうかを次のように調べる．

1. アクセス可能なサイト数を調べる．

2. アクセス可能なサイト数が全サイト数の過半数のときは，i の複製のどれか１つにでもアクセス可能であるなら i への読み書きが可能となる．

   アクセス可能なサイト数が全サイト数の過半数に満たない場合は，i のすべての複製にアクセス可能な場合のみ i への読み書きが可能となる．

site quorum では，ネットワーク分割時に過半数のサイトが存在する断片（サイトの多数派断片）において，その断片に複製が存在するすべてのデータ項目が読み書き可能となる．サイトの少数派断片では，その断片に複製すべてが存在するデータ項目のみ読み書き可能となる．

3 data quorum と site quorum の比較

本研究における仮定の下ではネットワーク分割時にある１つのデータ項目の全複製がサイト少数派断片に存在する確率は，そのデータ項目の過半数の複製がサイト少数派断片に存在する確率に比べて小さくなる．つまり，サイト多数派断片において読み書き可能となるデータ項目は site quorum を用いた方が data quorum を用いるより多くなる．従ってサイト多数派断片では読み書き可能なデータ項目数に大きな差が生じる．サイトの少数派断片においては，読み書き可能なデータ項目は data quorum の方が多くなるが，data quorum, site quorum 両方ともサイトの多数派断片に比べ少なく，読み書き不可能なデータ項目に読み書きを試みるトランザクションの比率は data quorum, site quorum ともに大きくなり，それらの比率にあまり差は生じない．

これらのことより，データベース全体で読み書き不可能なデータ項目に読み書きを試みるトランザクションの比率が data quorum の方が大きくなり，実行可能トランザクション比率は site quorum の方が data quorum より大きくなる．

また，多くのデータ項目に読み書きするトランザクションの比率が増加すると，data quorum では全断片において読み書き不可能なデータ項目に読み書きするトランザクションの比率が大きくなり全断片で実行不可能なトランザクションの比率が増加する．しかし，site quorum ではサイト多数派断片では実行不可能なトランザクションの比率はあまり増加せず，データベース全体に対する実行可能トランザクション比率は data quorum の方が site quorum よりも，より小さくなる．

4 ネットワーク分割のシミュレーション

data quorum と site quorum の差を定量評価するために，いくつかの多重化データベースモデルの上でネットワーク分割のシミュレーションを行なった．一回のシミュレーションで同じ多重化データベースモデルに対しさまざまなネットワーク分割をランダムに生じさせ，それぞれのネットワーク分割に対して site quorum と data quorum のそれぞれを用いた場合に対して実行可能なトランザクションの比率を求め，その平均を計算し site quorum と data quorum の比較を行なった．

図 1: 階層構造

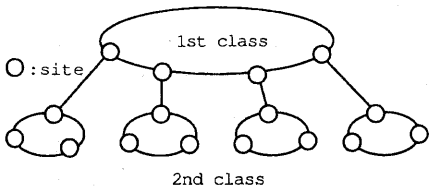
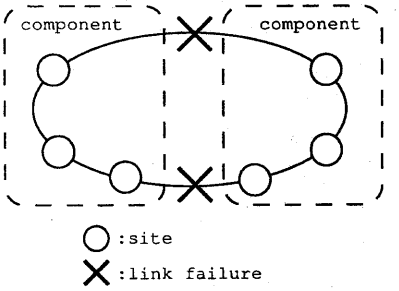


図 2: ネットワーク分割



4.1 データベース

多重化データベースのネットワークトポロジはリング構造，階層構造，格子構造，ランダムツリー構造の４つの構造を用いた．全トポロジーに対してデータ項目数，サイト数を設定する．各データ項目の複製数，各トランザクションのアクセスするデータ項目数はパラメータとする．

階層構造は２階層からなる構造に限定し，各階層はリング構造で構成されていると仮定し，第１階層のリング構造内サイト数，第２階層の各リング構造内サイト数を設定する．サイト数１６，第１階層のリング構造内サイト数４，第２階層の各リング構造内サイト数３の階層構造の例を図1に示す．格子構造の場合は，行数，列数を設定する．

4.2 シミュレーションの手順

多重化データベースのネットワークトポロジを決定し，次に各データ項目を全サイトに等確率で配置する．このように決定したデータベースに対してネットワーク分割を生じさせる．

データベース内の通信リンクを１つランダムに選び，その通信リンク上に障害が発生しその通信リンクは通信不可能となったとする．その通信リンク障害によりネットワーク分割が生じたかどうかを調べる．ネットワーク分割が生じなかった場合は，ネットワーク分割が生じるまで障害が発生する通信リンクをランダムに選び続ける．ネットワーク分割の例を図２に示す．

ネットワーク分割を生じさせた後，data quorum について断片１で読み書き可能となるデータ項目数 (d_dq_c1) を求める．全データ項目数を N として data quorum の断片１において読み書き可能となるデータ項目の比率 ($dratio\_dq\_c1$) を次のように求める．

$$dratio\_dq\_c1 = d\_dq\_c1/N$$

断片１に存在するサイトの比率を $sratio\_c1$，各トランザクションが読み書きするデータ項目数を Mとして data quorum の断片１において実行可能なトランザクションの比率 ($tratio\_dq\_c1$) を次のように求める．

$$tratio\_dq\_c1 = sratio\_c1 \times dratio\_dq\_c1^{M}$$

(１つのデータ項目に読み書きを試み，成功する確率は $dratio\_dq\_c1$ であるので M個のデータ項目に読み書きを試み，成功する確率は $dratio\_dq\_c1^{M}$ となる.)

断片２ ($tration\_dq\_c2$) についても同様に求め，data quorum に対する実行可能トランザクション比率 ($tratio\_dq$) を次のように求める．

$$tratio\_dq = tratio\_dq\_c1 + tratio\_dq\_c2$$

site quorum に対する実行可能トランザクション比率 ($tratio\_sq$) は data quorum と同様に求める．

設定した繰り返し回数だけネットワーク分割を発生させ，各ネットワーク分割について以上のように data quorum と site quorum のそれぞれについて求めた実行可能トランザクション比率から全ネットワーク分割における data quorum と site quorum それぞれについての実行可能トランザクション比率の平均値を計算し求める．

4.3 シミュレーションについて

各トポロジについて複製数をパラメータとして実行可能トランザクション比率を data quorum と site quorum 間で比較した．全トポロジにおいて，データ項目数を１００，各トランザクションのアクセスするデータ項目数を２とした．サイト数は表２に示す．各条件で１０００回実験し平均をとった．

結果は，どのトポロジーにおいても実行可能トランザクション比率は site quorum の方が data

表 1: 各トポロジにおけるサイト数（*1 は階層構造における第 1 階層のリング内サイト数または格子構造における行数，*2 は階層構造における第 2 階層の各リング内サイト数または格子構造における列数）

トポロジ	サイト数	*1	*2
リング	100		
階層	99	9	10
格子	100	10	10
ランダムツリー	100		

図 3: リング構造

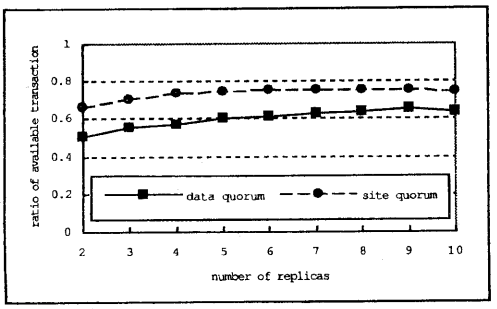


図 4: 階層構造

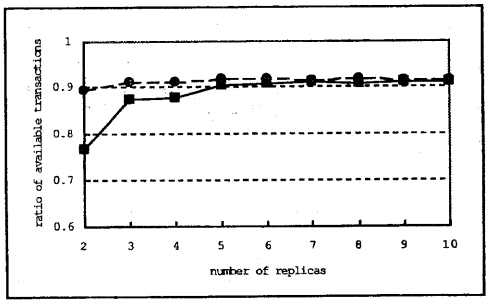
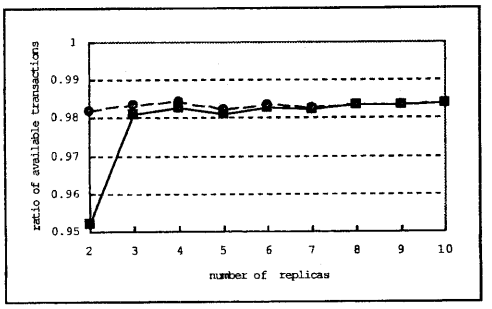


図 5: 格子構造



quorum より大きい値となった．また複製数が少ないほど，site quorum と data quorum の差大きいくなることが分かった．結果を図 3から図 6に示す．図 4から図 6の凡例は図 3と同じとする．

各トポロジについて各トランザクションがアクセスするデータ項目数をパラメータとして，実行可能トランザクション比率を data quorum と site quorum 間で比較した．全トポロジーにおいて各データ項目の複製数は 3 とし，データ項目数，サイト数，実験回数は 1 つ目のシミュレーションと同じとした．

結果は，どのトポロジーにおいてもアクセスするデータ項目数が 1，2，3 と増加するにつれ実行可能トランザクション比率は data quorum の方が大きく減少した．結果を図 7から図 10に示す．図 7から図 10の凡例は図 3と同じとする．

5 まとめ

本研究では，多重化データベースにおいてネットワーク分割時に実行可能なトランザクションの比率について data quorum と site quorum の比較を行なった．各データ項目の複製（全データ項目同数とする）がそれぞれ各サイトに等確率で配置され，また各トランザクションが全てのデータ項目に等確率で読み書きし（全トランザクション同データ項目数のデータ項目に読み書きする），全サイトに等確率で発生する多重化データベースにおいては，site quorum の方が比率が大きくなった．

さらに，site quorum と data quorum の差を定量評価するためにいくつかの多重化データベースモデルの上でネットワーク分割のシミュレーションを行ない，ネットワーク分割時に実行可能となるトランザクションの比率を site quorum を用いた場合と data quorum を用いた場合とで比較を行なった．

比較の結果，リング構造，階層構造，格子構造，ランダムツリー構造のいずれのトポロジにおいてもネットワーク分割時に実行可能なトランザクションの比率は site quorum の方が data quorum より大きくなり，複製数が少ないほどその差が大きく

図 6: ランダムツリー構造

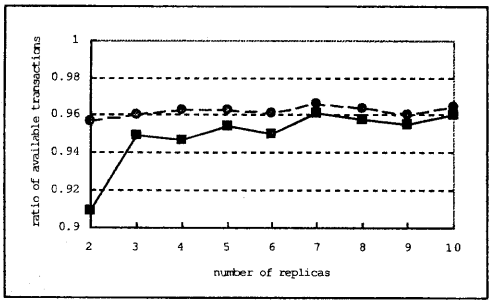
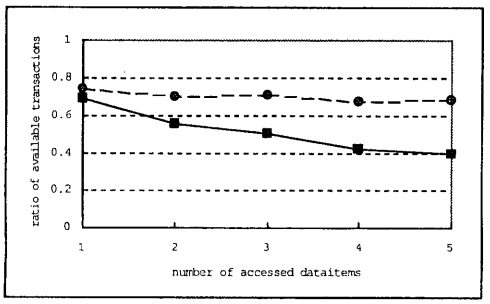


図 7: リング構造



なることが分かった．また，多くのデータ項目数に読み書きするトランザクションの比率が増加した場合，ネットワーク分割時に実行不可能なトランザクションの比率は data quorum の方が site quorum よりも，より大きくなることが分かった．

参考文献

[1] P. A. Bernstein, et al.: *Concurrency Control and Recovery in Database Systems.* Addison-Wesley, 1987.

図 8: 階層構造

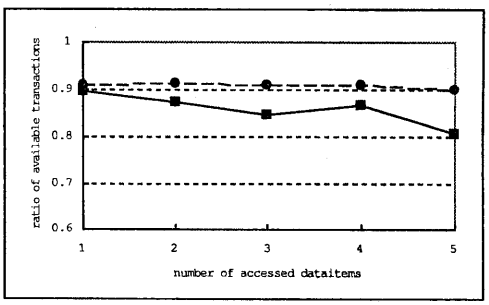


図 9: 格子構造

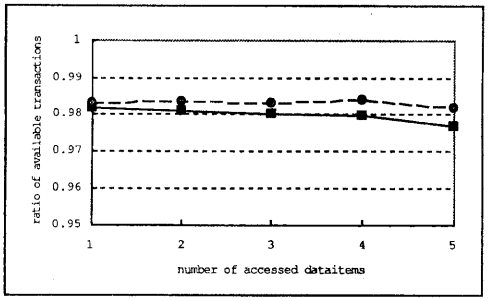


図 10: ランダムツリー構造