ドメインリーダによる複製管理プロトコルの提案‡

2E－3

中村 健二†　宮西 洋太郎††　佐藤 文明†　水野 忠則†
†静岡大学工学部　　††三菱電機（株）

1．はじめに

分散データベースシステムにおいて複製技術は，性能低下につながる問題がある．一方，可用性と信頼性を改善する大きなメリットがある．

複製の正当性を保証するために複製管理アルゴリズムがあり，もっとも知られているのはプライマリコピー方式[1]と定数合意（Quorum Consensus）方式アルゴリズムである[2]．また，この２つの方式を基盤とした拡張も幾つかある．しかし，これらのアルゴリズムは，データへのアクセス形態によって性能が変化し，完全とはいえない．

本論文では，リードライトの要求数が頻繁に変動しても安定な応答性能を保ち，種々のシステムにも対応できる柔軟な複製管理アルゴリズムとして，ドメインリーダアルゴリズムを提案する．

2．目的と動機

複製の導入は，可用性とフォールトトレランスを保証するためには，必要不可欠である．しかし，従来のアルゴリズムは，少数の複製しか管理できない問題がある．すなわち，データの正当性を保つために，オーバヘッドが生じることから，複製数は２，３個に限定されるケースが多い．

従来のアルゴリズムは少数の複製を管理する目的で構成されたため，複製の数が増えると，性能は急激に低下する．

現在，ネットワークの高速化，資源のコストダウンにもたらす技術の発展の影響で，複製の数を多く利用するようになってきた．しかし，高速化することで複製は必要ではないという疑問も出て来る．しかし，ネットワークの信頼度が向上しても，分断故障が生じないとは言い切れない．故障の場合，複製が存在しないと，システムは無制御状態となる恐れがある．

ドメインリーダアルゴリズムは比較的多数の複製を対象とし，性能と可用性を保ち，正当性と一貫性を保証するアルゴリズムである．

3．ドメインリーダアルゴリズム

ドメインリーダアルゴリズムは従来のアルゴリズムと違って，全複製とのトランザクション処理は同時には行わない．全複製をグループ（ドメイン）に分け，各ドメインには１つのリーダが割り当てられる．リーダはドメイン内の複製（サイト）を管理し，他のリーダと通信する．

リーダやグループの選択方法は，幾つかの研究があるので，本論文では説明しない．

アルゴリズムは図１で示すようなツリー構造になっており，以下に述べるようなステップで構成されている．

(1) サイトへ更新要求が到着した場合

サイトはリーダがロック状態であるかしらべる．そうであれば，リーダの応答を待つ．否ならば，つまり更新可能であれば，２ＰＣ（２相コミットメントプロトコル）でリーダへ新しいデータを送る．データ送信後，コミット（更新された）ならば，自分のサイトを更新し，そうでなければ，アボートする．

(2) リーダへ更新要求が到着した場合

サイト（リーダ側を含む）からの要求でリーダがロック状態ならば，サイトの要求を登録し，ロック開放後，サイトと通信する．逆にアンロック状態であれば，２ＰＣでサイトとデータ交換する．一方，他のリーダからの要求であれば，リーダをロックし，２ＰＣで他のリーダと通信する．

(3) サイトへ読み込み要求が到着した場合

サイトはリーダへ自分のデータが最新であるか聞きにいく．もしそうであれば，サイトのデータをユーザに送信する．もしサイト側のデータが古ければ，リーダから新しいデータが送られ，自分のサイトを更新し，データをユーザに送る．

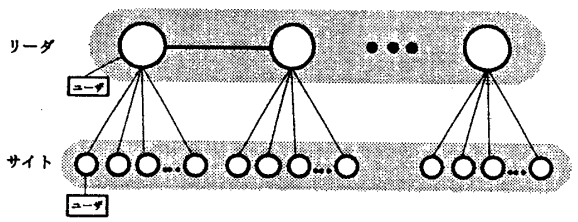

図１ドメインリーダ構造図

---

‡ Replicated Control Protocol Using Concept of Domain Leader

Kenji Nakamura†, Yohtaro Miyanishi††, Fumiaki Sato†, Tadanori Mizuno†
† Shizuoka University, 3-5-1 Johoku, Hamamatsu, Shizuoka 432, Japan
†† Mitsubishi Electric Co., 7-10-4 Nishigotanda, Shinagawa, Tokyo 141, Japan

(4) リーダへ読み込み要求が到着した場合

サイトからの要求の場合はリーダとサイトのデータを比較し，もし同じならば，サイト側からデータを取るようにメッセージを送る．同一でなければ，サイトへ最新データを送る．リーダ側からの要求であれば，ただ送る．

表１では，完全グラフネットワークでプライマリコピー，定数合意，ドメインリーダのリードとライトの応答コストメッセージ数を比較した結果を示す．ドメインリーダはｎ個の複製に対して，リードのコストはプライマリよりやや悪いが，ライトのコストはリーダ数ｍ個で，他の方式より良い結果となる．

方式	リード	ライト
プライマリコピー	0	n－1
定数合意 (リード定数＝ライト定数)	$\frac{n}{2}$	$\frac{n}{2}$
ドメインリーダ (m＜2n)	1	m

表１応答コスト比較

４．性能評価

ドメインリーダの性能評価は，シミュレーションで完全グラフトポロジーの上で検討し，応答時間の平均値を求めた．また，読み込みと書き込み要求頻度を変動して詳細に検討した．それから，シミュレーションでは複製の数を１００とし（リーダが１０個で，各リーダは，１０個のサイトを持っている），要求の到着率はポアソン分布を用いて計算した．

図２では，リード要求の到着率が９０％のグラフで，表１のようにリードの要求が多いと，差はあまりない．一方，図３では，ライト要求が８０％のグラフである．

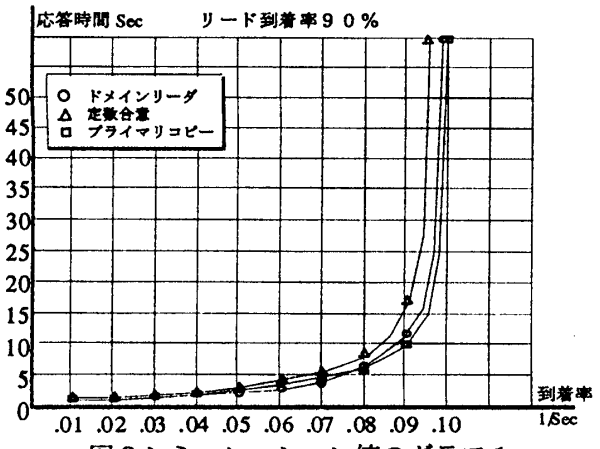

図２シミュレーション値のグラフ１

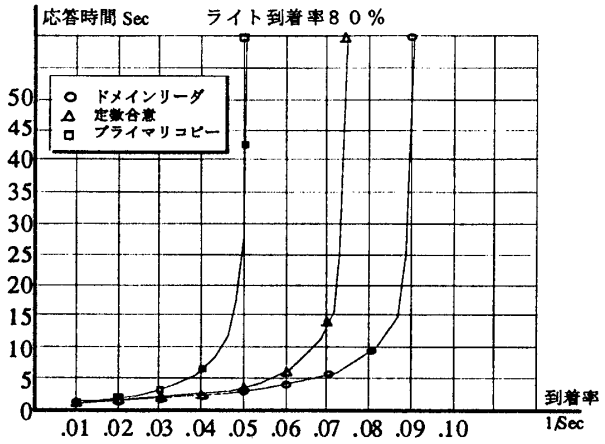

図３シミュレーション値のグラフ２

本研究では，複製数が多いシステムではドメインリーダを他のアルゴリズムと比較した結果，良い成績を残した．また逆に複製の数が少ない場合にも，同等な結果を得た．

５．おわりに

計算機技術の進歩により，多数の複製を用いた分散データベースシステムは増大する．また，複製をマルチデータベース環境に応用することも考えられる．すなわち，多数の複製を管理することは，重要な課題に成りつつあり，急速にニーズに答えなければならない．

本論文は多数の複製を対象とした複製管理アルゴリズムの提案であり，柔軟性，可用性，信頼性，性能などを考慮した研究である．

また，これからの課題として以下の項目があげられる．サブーリーダを用いたネットワーク分断時の複製管理の考察．マルチデータベース環境への対応．それから，実際のネットワーク構成に近似した性能評価シミューレションを検討する予定である．

参考文献

[1]M.Stonebraker,"Concurrency Control and Consistency of Multiple Copies of Data in Distributed INGRES", IEEE Trans. Software Eng.,Vol.S-E5, No.3, May1979, pp. 188-194.

[2]H.Gifford,"Weigthed Voting for Replicated Data", Proc.Seventh Symp. Operating Systems, ACM Press, New York, N.Y., 1979, pp. 150-162.

[3]中川路 哲男，”OSIとUNIX分散ﾄﾗﾝｻﾞｸｼｮﾝ処理技術解説”，ＳＲＣﾊﾝﾄﾞﾌﾞｯｸ，1994，PP.82-109.

[4]A.El Abbadi and S.Toueg, "Availability in Partioned Replicated Databases", ACM Trans.Database Syst.1989.