分散システムにおけるロック動作の解析と性能改善案

6U－1

宮西 洋太郎　中村 健二　渡辺 尚　水野 忠則
三菱電機　静岡大学工学部　静岡大学工学部　静岡大学工学部

１．はじめに

分散システムにおけるデータベース更新において、ロック操作を用いる場合が多い。複製を有する場合には、複製へのロック操作はデータ更新応答時間性能低下の原因の一つとなりうる。

ここでは、ロック動作が応答時間に与える影響について、待行列解析により近似的に応答時間を求める方法を示し、シミュレーションの結果と対比する。また、応答時間性能の低下を避ける一つのデータ更新方法を提案する。

２．関連する研究

分散システムの性能解析は待行列モデルを用いて行われているが[1)]、ここではロック操作を伴う分散データ更新応答時間性能に焦点を絞って待行列モデルによる性能解析を行う。

分散データ更新については、データの並行処理制御や一貫性制御のためにロック操作を用いた、2PL方式、2PC方式、タイムスタンプ方式、楽観的並行処理制御方式が研究されている[2)]。またネットワーク分断時の複製維持方式に投票-定数方式も提案されている[2)]。またデータの内容を利用する方式も研究されている[2)]。

３．対象データのモデル

(1) 対象データのアクセス

データのアクセスには読込み操作と書込み操作がある。ここでは書込み操作(以下更新と称する)を検討する。また一般的には書込みデータを生成するために、事前に読込みを行うことが必要であるが、ここでは簡単のため読込み時間と書込みの時間の和が一つの確率変数として指数分布に従うものとする。またアクセス要求の発生はポアソン分布に従うものとする。

対象となるデータは、データアイテム(j)単位に更新され、また 更新要求は地域的に分散した複数サイトの内の一つに到来し、そこで管理されるものとする。また、サイト(i)には、複製存在変数$X_{ij}=1$ならば複製データを持つものとする。

(2) 対象データの特性

データの複製を配置する際、対象データの統計的特性を利用するが、後半で提案する更新アルゴリズムにおいては、さらに、そのデータの持つ意味も利用することを考える。

４．待ち行列モデル

図1 及び図2 に待ち行列モデルを示す。

(1) サイトiでの更新処理時間

サイトiでの更新処理時間は、図2での並行する複数の指数分布に従う処理が全て完了するまでの時間である。以下に、ある一つのアイテムj のみについて考える。アイテムjの複製が存在するサイトを1,2,...,k,..,Kと番号を再付与し、サイト間での通信時間およびサイト待ち時間は無視できるものとすると、サイトiでのアイテムjの更新処理時間t の確率密度関数 $f_{ij}(t)$ は、

$$f_{ij}(t)=\sum_{r=1}^{K}(-1)^r Rr'$$

ただし、

$$Rr'=\sum_{k_1=1}^{K-r+1}\cdots\sum_{k_r>k_{r-1}}^{K}-(\mu_{k1}+\cdots+\mu_{kr})\cdot\exp(-(\mu_{k1}+\cdots+\mu_{kr})t)$$

$1/\mu_k$：サイトkの複製への単独の平均更新時間

以下、K＝3の場合について検討する、

$$f_{ij}(t)=\mu_1\exp(-\mu_1 t)+\mu_2\exp(-\mu_2 t)+\mu_3\exp(-\mu_3 t)-(\mu_1+\mu_2)\exp(-(\mu_1+\mu_2)t)-(\mu_2+\mu_3)\exp(-(\mu_2+\mu_3)t)-(\mu_3+\mu_1)\exp(-(\mu_3+\mu_1)t)+(\mu_1+\mu_2+\mu_3)\exp(-(\mu_1+\mu_2+\mu_3)t)$$

この関数$f_{ij}(t)$から更新処理時間の平均値$\overline{t}=S_{ij}$、2次モーメント$\overline{t^2}=M_{ij}$が求められる。

Analysis of Lock Operations in Distributed Systems and A Proposal of a Data Update Argorithm
Yohtaro Miyanishi*, Kenji Nakamura#, Takashi Watanabe#, Tadanori Mizuno#
*Mitsubishi Electric Corp., 7-10-4 Nishigotanda Shinagawa,Tokyo,141
#Shizuoka Unversity, Faculty of Engineering 3-5-1 Johoku,Hamamatsu,432,Japan

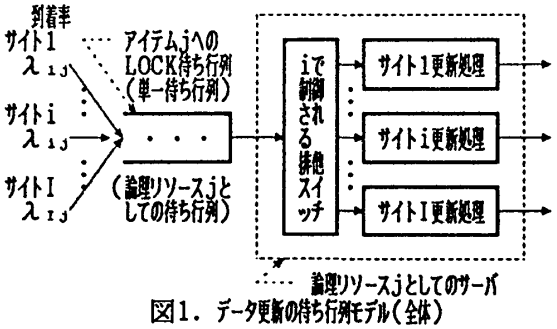

図1．データ更新の待ち行列モデル(全体)

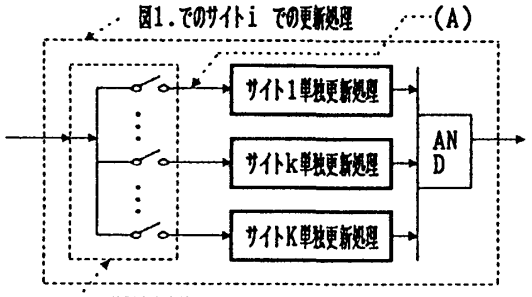

図2．データ更新の待ち行列モデル(サイトiでの更新処理)

(2)ロック待ち時間

各アイテムについてのロック及び解除は複製を持っている各サイトに対して同期して行われる(サイト間の通信時間、CPU処理時間はここでは無視する)ので、論理的に一つのリソースに対してロック及び解除を行っているものとみなせる。

図1において、「論理リソースjとしてのサーバ」と記した部分の処理時間の確率密度分布は、(1)で求めた「サイトiでの更新処理」確率密度分布をアクセス到着率λ_{ij}でウエイト付けて合成されたものである。この分布は指数分布ではないが、待ち時間はM/G/1のポランチェック-ヒンチンの公式により求めることができる。

サイト共通の平均待ち時間W_j

$$W_j = \frac{\rho_j \cdot S_j}{2\cdot(1-\rho_j)}\left(1 + \left(\frac{\sigma_j}{S_j}\right)^2\right)$$

ただし、

$$S_j = \sum_{i=1}^{I} \frac{\lambda_{ij}}{\lambda_j} \cdot S_{ij} \quad , \quad \lambda_j = \sum_{i=1}^{I} \lambda_{ij}$$

$$\rho_j = \sum_{i=1}^{I} \lambda_{ij} \cdot S_{ij}$$

$$\sigma_j^2 = \sum_{i=1}^{I} \frac{\lambda_{ij}}{\lambda_j} \cdot M_{ij} - (S_j)^2$$

サイトi の平均応答時間T_{ij}は次式で求まる。

$$T_{ij} = W_{ij} + S_{ij}$$

(3)シミュレーションとの比較

シミュレーションとの比較の例を以下に記す。

I=3, J=1, λ_{11}=0.2, λ_{21}=0.1, λ_{31}=0.05

$1/\mu_1$=2.0, $1/\mu_2$=1.0, $1/\mu_3$=0.5の場合

	解析値	シミュレーション結果		
	i=1,2,3	i=1	i=2	i=3
S_{ij}	2.386	2.365	2.379	2.375
W_{ij}	9.792	9.370	9.508	9.622
T_{ij}	12.177	11.735	11.888	11.997
ρ_1	0.700	0.703 (サイト1使用率)		
ρ_2	0.350	0.353 (サイト2使用率)		
ρ_3	0.175	0.176 (サイト3使用率)		
ρ_L	0.835	0.840 (ロック中率)		

5．データ更新アルゴリズム案

ここに提案する方式は「権限分割方式」とも呼べる方式で、各サイト毎にデータ更新の限界値を設定し、その範囲内で更新を許容する方式である。動作はmodestであるので、Modestly Optimistic Concurrency Control (MOCC)(制限付楽観的制御方式)と呼ぶことにする。

この方式が有効である適用分野は預金口座残高、在庫量、座席予約用座席数等のように共通の量を取り崩すような分野である。

通常のアクセス時は、上限値の範囲内で更新を行ない、夜間等のアクセスの閑散時に全体の一貫性回復するという方法である。またネットワーク分断時にも上限値範囲内での更新が可能である。

6．おわりに

(1) ロックを用いて複数の複製を更新する場合の応答時間性能についての近似的な解析的評価方法を示した。またシミュレーションとの対比によりその妥当性を示した。

(2) ロックに起因する性能低下を回避する更新アルゴリズムの概要を示した。

今後の課題は

・複数データアイテムの並列更新時の解析
(この場合、図2.(A)に待ち行列が発生する。)
・提案アルゴリズムの性能評価
・性能評価と複製配置の関連の検討等である。

7．参考文献

1)Kleinrock L.,"On the Modeling and Analysis of Computer Network," Proc.IEEE.V.81,No.8.
2)Reading in DISTRIBUTED COMPUTING SYSTEMS Chapter 9,10,11 IEEE Comp.Soc.Press,1994