データ更新量を考慮した 非同期型レプリケーション方策のモデル化と解析

Modelling and Analysis of an Asynchronous Replication System Considering the Amount of Data Update

木村充位　　今泉充啓 †　　中川覃夫 ‡

Mitsutaka Kimura　Mitsuhiro Imaizumi　Toshio Nakagawa

愛知学泉大学 †　愛知工業大学 ‡

Abstract

This paper considers the problem of reliability in a server system with asynchronous replication considering the amount of data update. We have formulated a model of a server system with asynchronous replication that the server transmits the database content to a backup site after a constant time. In this paper, we formulate a stochastic model of a server system with asynchronous replication considering the amount of data update. That is, the server transmits the database content from a main site to a backup site after a constant number of updates. We derive the expected number of data replication and the expected number of data update until system down.

Keywords; Server System, Asynchronous Replication, Main Site, Backup Site, Disaster

1　はじめに

近年，災害によるデータ損失へ対応し，レプリケーションという手法を使ってサーバのデータが保護されている．これは被災に備えてバックアップデータを保存するだけでなく，被災後に業務再開できるサーバを備えたバックアップサイトをもつシステムで行われる手法であり，メインサイトのデータの内容とバックアップサイトのデータの内容をネットワークを介して同期させることをレプリケーションという [1]．レプリケーションは同期型と非同期型の二種類に分けられる．同期型では，メインサイトとバックアップサイトが同期をとりながらストレージへデータ更新を行う方法である．一方，非同期型は，メインサイトのストレージへデータ更新が行われた後，任意のタイミングでバックアップサイトのストレージへデータ更新が行われる方法である [2] ～ [5]．これまでに，同期型のレプリケーションを適用した確率モデルの提案 [7] やある一定時間間隔でレプリケーションが行われる一般的な非同期型の確率モデルの提案を行い，さらに信頼性に関する諸種の信頼性解析を行った [8]．また，同期型モデルと非同期型モデルのコスト有効性を比較検討し，種々の考察を行った．

ここでは，データ更新量を考慮した非同期型のレプリケーションについて，ある一定回数のデータ更新が行われたときに，レプリケーションを行うモデルを提案する．

2　モデルの設定

システムの概念図を，図 1 に示す.

監視サーバとメインサイト及びバックアップサイトで構成される遠隔距離通信のネットワークシステムを考える．メインサイトとバックアップサイトは，それぞれサーバとストレージで構成されており，バックアップサイトは待機系として常駐する．メインサイトのサーバは通常業務を行い，クライアントの要求によりストレージへデータ更新が行われる．また，監視サーバはメインサイトのストレージについて，ある間隔でバックアップサイトのストレージにレプリケーションの指示を行う．

ここでは，監視サーバの動作に注目し，メインサイトにおけるストレージのデータ更新に着目してモデル化を行う．すなわち，監視サーバは，メインサイトの状態を常時監視し，クライアントの要求によって k 回のデータ更新が行われたならば，ネットワークを介してバックアップサイトのストレージへレプリケーションを行う．なお，メインサイトでは広域災害が発生した場合，通常業務は停止するものとし，バックアップサイトでは災害が発生した場合，ただちに復旧処理が開始されることを仮定する．

(1) 広域災害の発生時間々隔は，ランダム性を考慮して指数分布に従うものとし，メインサイトにおける確率分布を，$F_1(t) = 1 - e^{-\lambda_1 t}$ とする．

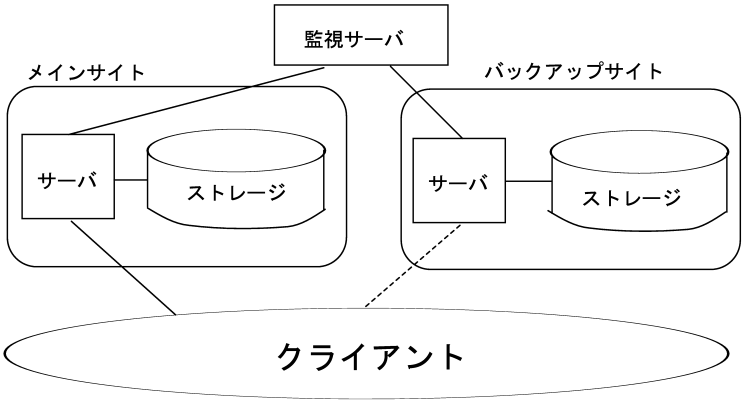

図 1. システムの概要図.
Fig.1 Outline of a system.

(a) バックアップサイトが正常ならば，メインサイトを瞬時にグループから切り離し，バックアップサイトへ切り替える．

(b) バックアップサイトが障害状態にあれば，システムダウンに至る．ここで，バックアップサイトの障害発生間隔は，指数分布 $F_2(t) = 1 - e^{-\lambda_2 t}$ に従うものとする．

(2) メインサイトでは，クライアントからのデータ更新要求が指数分布 $A(t) = 1 - e^{-\alpha t}$ に従ってランダムに発生し，そのデータ更新処理は分布 $B(t) = 1 - e^{-\beta t}$ に従って完了するものとする．

(3) 監視サーバは，メインサイトの状態を常時監視する．ここで，メインサイトにおいて，クライアントの要求によって k 回 $(k = 1, 2, \cdots)$ のデータ更新が行われたとき，バックアップサイトの状態を確認し，ネットワークを介してバックアップサイトのストレージへレプリケーションを開始する．

(a) もしバックアップサイトが正常状態のとき，ネットワークを介してバックアップサイトのストレージへレプリケーションを開始する．このレプリケーションに要する時間分布は指数分布 $W_1(t) = 1 - e^{-w_1 t}$ とする．ここで，もしレプリケーション中にメインサイトとバックアップサイトのいずれかに障害が発生した場合は，システムダウンに至ると仮定する．

(b) バックアップサイトが障害状態にあれば，その復旧を待つ．バックアップサイトの復旧に要する時間分布は $G(t) = 1 - e^{-\gamma t}$ に従うものとする．

(i) もしバックアップサイトが復旧完了前に，メインサイトに広域災害が発生すれば，システムダウンに至る．

(ii) もしバックアップサイトが復旧完了前に，メインサイトにデータ更新要求があれば，データ更新を行う．

(iii) 復旧完了後はレプリケーションを行う．このレプリケーションに要する時間分布は指数分布 $W_2(t) = 1 - e^{-w_2 t}$ とする．なお，このときのレプリケーションに要する平均時間 $1/w_2$ はメインサイトでのデータ更新量が k 回以上となるため，$1/w_2 \geq 1/w_1$ とする．ここで，もしレプリケーション中にメインサイトとバックアップサイトのいずれかに障害が発生した場合は，システムダウンに至ると仮定する．

まず，バックアップサイトの状態推移確率を求める．文献 [7] により，バックアップサイトの状態を，

状態 0 : 正常状態.

状態 1 : 障害発生.

と定義すると，バックアップサイトが時刻 0 で状態 i にあり，時刻 t で状態 j にある確率 $P_{ij}(t)(i, j = 0, 1)$ は以下のようになる [6].

$$
\begin{aligned}
P_{00}(t) &= \frac{\gamma}{\lambda_2+\gamma} + \frac{\lambda_2}{\lambda_2+\gamma}e^{-(\lambda_2+\gamma)t}, \\
P_{11}(t) &= \frac{\lambda_2}{\lambda_2+\gamma} + \frac{\gamma}{\lambda_2+\gamma}e^{-(\lambda_2+\gamma)t}, \\
P_{01}(t) &= 1 - P_{00}(t), \\
P_{10}(t) &= 1 - P_{11}(t).
\end{aligned}
$$

さらに，監視サーバを含む非同期型ネットワークシステムの状態を次のように定義する．

状態 2 : システムの開始または再開始.

状態 3 : メインサイトのデータ更新処理を開始.

状態 R_1 : メインサイトが k 回目のデータ更新処理を完了後，レプリケーション動作の開始.

状態 4 : メインサイトが k 回目のデータ更新処理を完了後，バックアップサイトが障害状態.

状態 5 : バックアップサイトが障害状態で，メインサイトがデータ更新処理開始.

状態 R_2 : メインサイトがアイドリング状態で，バックアップサイトの復旧が完了し，レプリケーション動作の開始.

状態 F : システムダウン.

状態 S_W : バックアップサイトへシステム業務の切り替え.

各状態間の推移は，マルコフ再生過程 [6] を形成し，その推移は図 3 のように表される．

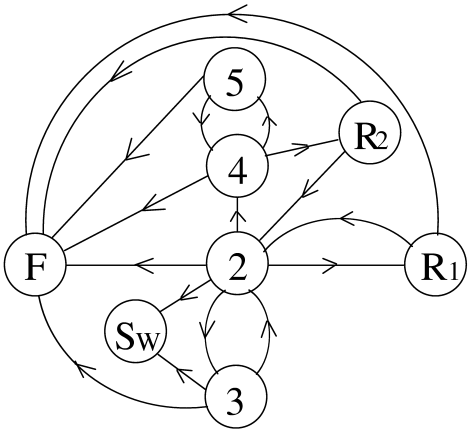

図 3. システムの状態推移図.
Fig3. A state transition diagram of a system.

各状態間の推移確率時間分布 $Q_{i,j}(t)(i = 2,3,4,5,R_1,R_2; j = 2,3,4,5,R_1,R_2,F,S_W)$ とおくと次式を得る．

$$Q_{2,3}(t) = \int_0^t \overline{F_1}(t)dA(t), \tag{1}$$

$$Q_{2,F}(t) = \int_0^t P_{01}(t)\overline{A}(t)dF_1(t), \tag{2}$$

$$Q_{2,S_W}(t) = \int_0^t P_{00}(t)\overline{A}(t)dF_1(t), \tag{3}$$

$$Q_{3,2}(t) = \int_0^t \overline{F_1}(t)dB(t), \tag{4}$$

$$Q_{3,F}(t) = \int_0^t P_{01}(t)\overline{B}(t)dF_1(t), \tag{5}$$

$$Q_{3,S_W}(t) = \int_0^t P_{00}(t)\overline{B}(t)dF_1(t), \tag{6}$$

$$Q_{4,5}(t) = \int_0^t \overline{G}(t)\overline{F_1}(t)dA(t), \tag{7}$$

$$Q_{4,R_2}(t) = \int_0^t \overline{A}(t)\overline{F_1}(t)dG(t), \tag{8}$$

$$Q_{4,F}(t) = \int_0^t \overline{G}(t)\overline{A}(t)dF_1(t), \tag{9}$$

$$Q_{5,4}(t) = \int_0^t \overline{F_1}(t)dB(t), \tag{10}$$

$$Q_{5,F}(t) = \int_0^t \overline{B}(t)dF_1(t), \tag{11}$$

$$Q_{R_i,2}(t) = \int_0^t \overline{F_1}(t)\overline{F_2}(t)dW_i(t), \qquad (i=1,2) \tag{12}$$

$$Q_{R_i,F}(t) = \int_0^t \overline{W_i}(t)\overline{F_2}(t)d[F_1(t)] + \int_0^t \overline{W_i}(t)\overline{F_1}(t)d[F_2(t)], \qquad (i=1,2). \tag{13}$$

さて，システムが時刻 0 で動作を開始してからバックアップサイトへシステム業務が切り替えられるか，または，システムダウンに至るまでの間にレプリケーションが行われる平均回数を求める．状態 2 から状態 R_1 までの経過時間分布を $H_{2,R_1}(t)$, 状態 2 から状態 4 までの経過時間分布を $H_{2,4}(t)$, 状態 4 から状態 R_2 までの経過時間分布を $H_{4,R_2}(t)$ とすると，

$$H_{2,R_1}(t) = [Q_{2,3}(t) * Q_{3,2}(t)]^{(k-1)} * Q_{2,3}(t) * \int_0^t P_{00}(t)\overline{F_1}(t)dB(t), \tag{14}$$

$$H_{2,4}(t) = [Q_{2,3}(t) * Q_{3,2}(t)]^{(k-1)} * Q_{2,3}(t) * \int_0^t P_{01}(t)\overline{F_1}(t)dB(t), \tag{15}$$

$$H_{4,R_2}(t) = \sum_{i=1}^{\infty}[Q_{4,5}(t) * Q_{5,4}(t)]^{(i-1)} * Q_{4,R_2}(t), \tag{16}$$

$$H_{2,R_2}(t) = H_{2,4}(t) * H_{4,R_2}(t), \tag{17}$$

を得る．よって，システムが時刻 0 で状態 2 を出発し，時刻 t までに状態 R_1 を訪問する回数の分布を $M_{2,R_1}(t)$, システムが時刻 0 で状態 2 を出発し，時刻 t までに状態 R_2 を訪問する回数の分布を $M_{2,R_2}(t)$ とすると，

$$M_{2,R_1}(t) = H_{2,R_1}(t) + H_{2,R_1}(t) * Q_{R_1,2}(t) * M_{2,R_1}(t), \tag{18}$$

$$M_{2,R_2}(t) = H_{2,R_2}(t) + H_{2,R_2}(t) * Q_{R_2,2}(t) * M_{2,R_2}(t). \tag{19}$$

一般に $\phi(s) \equiv \int_0^\infty e^{-st}d\Phi(t)$ とおき，式（18）と式（19）を LS 変換し，再生方程式を解くことによって，次式を得る．

$$m_{2,R_1}(s) = \frac{h_{2,R_1}(s)}{1 - h_{2,R_1}(s)q_{R_1,2}(s)}, \tag{20}$$

$$m_{2,R_2}(s) = \frac{h_{2,R_2}(s)}{1 - h_{2,R_2}(s)q_{R_2,2}(s)}. \tag{21}$$

従って，システムが状態 F または状態 SW に至るまでのレプリケーションの平均回数 M_R は次のように求めることができる．

$$
\begin{aligned}
M_R &\equiv \lim_{s\to 0}[m_{2,R_1}(s)+m_{2,R_2}(s)] \\
&= \frac{h_{2,R_1}(0)}{1-h_{2,R_1}(0)q_{R_1,2}(0)}+\frac{h_{2,R_2}(0)}{1-h_{2,R_2}(0)q_{R_2,2}(0)}, \qquad (22)
\end{aligned}
$$

ここで,

$$
\begin{aligned}
h_{2,R_1}(0) &\equiv \left[\frac{\alpha\beta}{(\alpha+\lambda_1)(\beta+\lambda_1)}\right]^k\left(\frac{\lambda_1+\beta+\gamma}{\lambda_1+\lambda_2+\beta+\gamma}\right), \\
h_{2,R_2}(0) &\equiv \left[\frac{\alpha\beta}{(\alpha+\lambda_1)(\beta+\lambda_1)}\right]^k\left[\frac{\gamma\lambda_2(\beta+\lambda_1)}{[\alpha\lambda_1+(\gamma+\lambda_1)(\beta+\lambda_1)][\lambda_1+\lambda_2+\beta+\gamma]}\right], \\
q_{R_i,2}(0) &\equiv \frac{w_i}{\lambda_1+\lambda_2+w_i}, \quad (i=1,2).
\end{aligned}
$$

次に，1 回もレプリケーションできないままバックアップサイトへシステム業務が切り替えられるか，または，システムダウンに至るまでのデータ更新の平均回数を求める．システムが時刻 0 で状態 2 を出発し，時刻 t までに状態 R_1, R_2 を訪問しないで状態 3 を訪問する回数の分布を $M_{D_1}(t)$, 状態 R_1, R_2 を訪問しないでデータ更新する回数の分布を $M_{D_2}(t)$ とすると，

$$
M_{D_1}(t) = \sum_{i=1}^{k}(i-1)[Q_{2,3}(t)*Q_{3,2}(t)]^{(i-1)}*\{Q_{2,S_W}(t)+Q_{2,F}(t)+Q_{2,3}(t)*[Q_{3,S_W}(t)+Q_{3,F}(t)]\}, \qquad (23)
$$

$$
M_{D_2}(t) = \sum_{i=1}^{\infty}\{k+(i-1)\}H_{2,4}(t)*[Q_{4,5}(t)*Q_{5,4}(t)]^{(i-1)}*[Q_{4,F}(t)+Q_{4,5}(t)*Q_{5,F}(t)], \qquad (24)
$$

を得る．さらに，式（23）と式（24）を LS 変換すると，

$$
\begin{aligned}
m_{D_1}(s) &= \left(\frac{q_{2,S_W}(s)+q_{2,F}(s)+q_{2,3}(s)[q_{3,S_W}(s)+q_{3,F}(s)]}{1-q_{2,3}(s)q_{3,2}(s)}\right) \\
&\quad\times\left[\frac{q_{2,3}(s)q_{3,2}(s)}{1-q_{2,3}(s)q_{3,2}(s)}-\left(\frac{q_{2,3}(s)q_{3,2}(s)}{1-q_{2,3}(s)q_{3,2}(s)}+k\right)[q_{2,3}(s)q_{3,2}(s)]^k\right], \qquad (25) \\
m_{D_2}(s) &= \left(\frac{h_{2,4}(s)[q_{4,F}(s)+q_{4,5}(s)q_{5,F}(s)]}{1-q_{4,5}(s)q_{5,4}(s)}\right)\left(\frac{q_{4,5}(s)q_{5,4}(s)}{1-q_{4,5}(s)q_{5,4}(s)}+k\right). \qquad (26)
\end{aligned}
$$

従って，システムが状態 F または状態 SW に至るまでに状態 R_1, R_2 を訪問しない場合のデータの平均更新回数 M_D は次のように求めることができる．

$$
\begin{aligned}
M_D &\equiv \lim_{s\to 0}[m_{D_1}(s)+m_{D_2}(s)] \\
&= m_{D_1}(0)+m_{D_2}(0), \qquad (27)
\end{aligned}
$$

ここで,

$$
\begin{aligned}
m_{D_1}(0) &= \frac{\alpha\beta}{\lambda_1(\alpha+\beta+\lambda_1)}-\left(\frac{\alpha\beta}{\lambda_1(\alpha+\beta+\lambda_1)}+k\right)\left(\frac{\alpha\beta}{(\alpha+\lambda_1)(\beta+\lambda_1)}\right)^k, \\
m_{D_2}(0) &= \left[\frac{\lambda_1\lambda_2(\alpha+\beta+\lambda_1)}{[\alpha\lambda_1+(\gamma+\lambda_1)(\beta+\lambda_1)][\lambda_1+\lambda_2+\beta+\gamma]}\right] \\
&\quad\times\left(\frac{\alpha\beta}{\alpha\lambda_1+(\gamma+\lambda_1)(\beta+\lambda_1)}+k\right)\left(\frac{\alpha\beta}{(\alpha+\lambda_1)(\beta+\lambda_1)}\right)^k.
\end{aligned}
$$

3 おわりに

メインサイトにおいて，ある一定回数のデータ更新が行われたときに，バックアップサイトへレプリケーションを行う非同期型レプリケーションを適用したサーバシステムについてモデル化を行った．これまでに非同期型レプリケーションについて，ある一定時間間隔でレプリケーションが行われる一般的な方式をモデル化し考察してきた [8]．ここではデータ更新の頻度を考慮した非同期型のレプリケーションを適用し，確率モデルを設定した．さらにシステムダウンするまでのレプリケーションの平均回数やデータ更新の平均回数を解析的に導出した．

今後は導出したレプリケーションの平均回数やデータ更新の平均回数を用いて期待費用を解析的に導出し，その期待費用を最小にするレプリケーション時期を考察する予定である．このようなサーバシステムの高信頼化の問題は，今後ますます重要な課題となることが考えられ，この方面に対する多くの研究が期待される．

参考文献

[1] 大和純一，菅真樹，菊池芳秀, “非常災害に向けた高度情報通信ネットワークの構成と制御小特集５．広域災害に対するストレージによるデータ保護”, 電子情報通信学会誌, vol.89 No.9 pp.801-805, 2006.

[2] Oracle Corporation, “Oracle Database 10g の Oracle Data Guard 企業のための障害時リカバリ”, http://otndnld.oracle.co.jp/products/database/oracle10g/pdf/DataGuardTechOverview_10gR1.pdf.

[3] VERITAS Software Corporation, “VERITAS volume replication for UNIX datasheet”, http://eval.veritas.com/mktginfo/products/Datasheets/High_Availability/vvr_datasheet_unix.pdf.

[4] 今井哲郎, 荒木荘一郎，菅原智義，藤田範人，末村剛彦，“広域分散データセンサ間でのサービス無停止障害回避方式”， 信学技報，NS2003-293, IN2003-248, pp.199-202, March 2004.

[5] EMC Corporation, “Using asynchronous replication for business continuity between two or more sites”, http://www.emc.com/products/software/srdf_a/pdf/C1058_srdf_asynchronous_mode_wp_ldv.pdf.

[6] S. Osaki, “Applied Stochastic System Modeling”, *Springer-Verlag*, Berlin, 1992.

[7] 木村充位，今泉充啓，安井一民，“広域災害を伴うシステムの同期型レプリケーションのモデル化”, 岐阜市立女子短期大学研究紀要第 57 輯, pp. 41-47 , 2008.

[8] Mitsutaka Kimura, Mitsuhiro Imaizumi, Toshio Nakagawa ; Reliability Consideration of a Server System with Asynchronous Replication For Disaster Recovery, Proceedings of the 2008 Asian International Workshop on Advanced Reliability Modeling. pp. 686-693, 2008

（提出期日　平成 20 年 11 月 28 日）