岐阜市立女子短期大学研究紀要第 57 輯（平成 20 年 3 月）

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

Modelling of a Synchronous Replication System for Disaster Recovery

木村充位　　今泉充啓 †　　安井一民 ‡
Mitsutaka Kimura　Mitsuhiro Imaizumi　Kazumi Yasui
愛知学泉大学 †　愛知工業大学 ‡

Abstract

As the Information technology has remarkably developed, business data has computerized. It is fatal for business to lose the data by a disaster. In order to avoid data loss and stagnation of business activity by a disaster, remote replication by using network link has been adopted. Remote replication enables data protection and faster business restart in the event of a disaster. This paper considers the problem of reliability in a server system with a synchronous replication for disaster recovery. We formulate a stochastic model of a server system which consists of a monitor server, a main site and a backup site. A client data in main site is replicated between main site and backup site. We derive the probabilities of switching backup site and system failure, the expected number of data replication.

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

1　はじめに

企業活動のＩＴ化に伴い，企業の重要なデータが電子データとしてサーバに蓄積されている．これらのデータは日々生産され続けており，災害などで失われると，ハードウェアやソフトウェアと異なって再度入手することはできない．失われたデータの内容によってはその損失が企業にとって致命的になる場合すらある．

地震，台風，洪水，火災，停電など様々な災害によるデータ損失への対応が重要であり，近年，レプリケーションという手法を使ってデータが保護されている．これは，被災後の業務再開を考慮して，バックアップデータを保存するだけでなく，被災後にサーバを備えたバックアップサイトを使用するシステムが構築されている．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．すなわち，メインサイトが被災した場合，バックアップサイトにレプリケーションされたデータを用いて，バックアップサイトがメインサイトの業務を引き継ぐことになる [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].

ここでは, あるメインサイトに広域災害が発生した場合に，バックアップサイトとの間で同期型のレプリケーションを適用した確率モデルを構築する．また，メインサイトが被災したとき，システムダウンする確率，バックアップサイトに業務が引き継がれる確率，システムダウンするまでのレプリケーション回数などを解析的に導出する．

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．なお，メインサイトでは広域災害が発生した場合，メインサイトでの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．

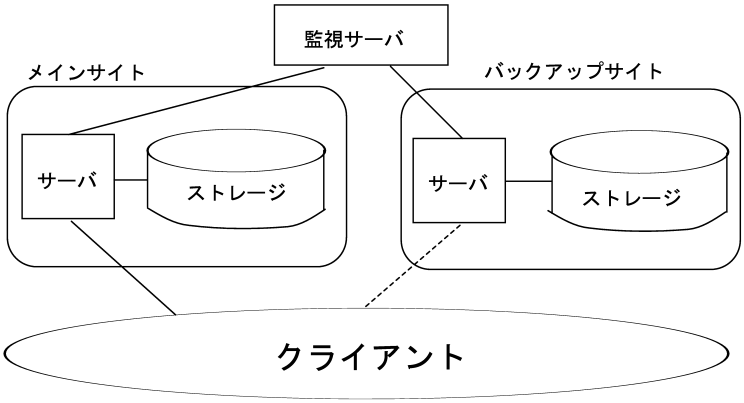

図 1. システムの概要図.
Fig.1 Outline of a system.

(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．

(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) メインサイトがデータ更新処理を終了したとき，

(i) バックアップサイトが正常であれば，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開始する．このレプリケーションに要する時間分布は更新データの多寡によることが考えられるので指数分布 $W(t) = 1 - e^{-wt}$ 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，ストレージの一貫性を考慮してシステムダウンに至ると仮定する．なお，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．

(ii) バックアップサイトが障害状態にあれば，その復旧を待ち，復旧完了後にレプリケーションを行う．ただし，バックアップサイトの障害復旧中に，データ更新要求があれば，メインサイトはその更新処理を行い，データ更新処理終了後にレプリケーションを行う．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．

(4) メインサイトに障害が発生したとき，

(i) 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．

(ii) 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．

以上の仮定のもとで，まずバックアップサイトの状態確率を求めよう．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，各状態間の推移は，2 状態をもつマルコフ再生過程 [6] を形成し，その推移は図 2 のように表される．
このとき，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 にある確率を $P_{ij}(t)(i, j = 0, 1)$ とおくと，$P_{00}(0) = 1$, $P_{01}(0) = 0$, $P_{10}(0) = 0$, $P_{11}(0) = 1$ の初期条件のもとで，次のような状態確率を得る [6].

$$P_{00}(t) \quad = \quad \frac{\gamma}{\lambda_2 + \gamma} + \frac{\lambda_2}{\lambda_2 + \gamma} e^{-(\lambda_2 + \gamma)t},$$

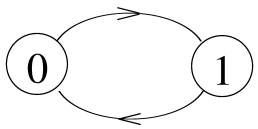

図 2. バックアップサイトの状態推移図.
Fig2. A state transition diagram of a backup site.

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

さて，監視サーバを含む同期型の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 5 : システムの開始または再開始.

状態 6 : バックアップサイトが正常状態で，メインサイトがデータ更新処理開始.

状態 7 : バックアップサイトが障害状態で，メインサイトのデータ更新処理完了.

状態 8 : メインサイトがデータ更新処理中にバックアップサイトの障害発生.

状態 R : 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.

システムの各状態を上のように定義すると，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.

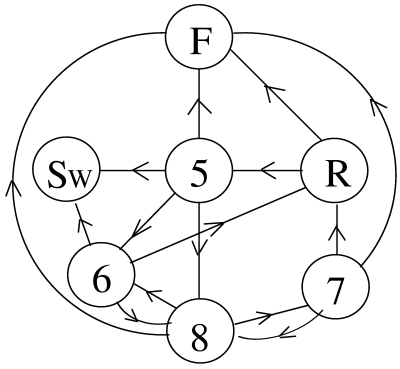

図 3. システムの状態推移図.
Fig3. A state transition diagram of a system.

各状態間の推移確率時間分布 $Q_{i,j}(t)(i = 5, 6, 7, 8, R; j = 5, 6, 7, 8, R, F, S_W)$ とおくと次式を得る.

$$
\begin{aligned}
Q_{5,6}(t) &= \int_0^t P_{00}(t)\overline{F}_1(t)dA(t), && (1)\\
Q_{5,8}(t) &= \int_0^t P_{01}(t)\overline{F}_1(t)dA(t), && (2)\\
Q_{5,S_W}(t) &= \int_0^t P_{00}(t)\overline{A}(t)dF_1(t), && (3)\\
Q_{5,F}(t) &= \int_0^t P_{01}(t)\overline{A}(t)dF_1(t), && (4)\\
Q_{6,R}(t) &= \int_0^t \overline{F_2}(t)\overline{F_1}(t)dB(t), && (5)\\
Q_{6,8}(t) &= \int_0^t \overline{F_1}(t)\overline{B}(t)dF_2(t), && (6)\\
Q_{6,S_W}(t) &= \int_0^t \overline{F_2}(t)\overline{B}(t)dF_1(t), && (7)\\
Q_{7,R}(t) &= \int_0^t \overline{F_1}(t)\overline{A}(t)dG(t), && (8)\\
Q_{7,8}(t) &= \int_0^t \overline{F_1}(t)\overline{G}(t)dA(t) && (9)\\
Q_{7,F}(t) &= \int_0^t \overline{G}(t)\overline{A}(t)dF_1(t), && (10)\\
Q_{8,7}(t) &= \int_0^t \overline{G}(t)\overline{F_1}(t)dB(t), && (11)\\
Q_{8,6}(t) &= \int_0^t \overline{B}(t)\overline{F_1}(t)dG(t) && (12)\\
Q_{8,F}(t) &= \int_0^t \overline{G}(t)\overline{B}(t)dF_1(t), && (13)\\
Q_{R,5}(t) &= \int_0^t \overline{F_1}(t)\overline{F_2}(t)dW(t), && (14)\\
Q_{R,F}(t) &= \int_0^t \overline{W}(t)\overline{F_2}(t)d[F_1(t)] + \int_0^t \overline{W}(t)\overline{F_1}(t)d[F_2(t)]. && (15)
\end{aligned}
$$

次に，状態 R を訪問する平均回数 M_R と状態 5 を訪問する平均回数 M_5 を求める．システムが時刻 0 で状態 5 を出発し，時刻 t までに状態 R を訪問する回数の分布を $M_{i,R}(t)(i=5,6,7,8)$ とすると，次のような再生形方程式を得る．

$$
\begin{aligned}
M_{5,R}(t) &= Q_{5,6}(t) * M_{6,R}(t) + Q_{5,8}(t) * M_{8,R}(t), && (16)\\
M_{6,R}(t) &= Q_{6,R}(t) + Q_{6,R}(t) * Q_{R,5}(t) * M_{5,R}(t) + Q_{6,8}(t) * M_{8,R}(t), && (17)\\
M_{7,R}(t) &= Q_{7,R}(t) + Q_{7,R}(t) * Q_{R,5}(t) * M_{5,R}(t) + Q_{7,8}(t) * M_{8,R}(t), && (18)\\
M_{8,R}(t) &= Q_{8,6}(t) * M_{6,R}(t) + Q_{8,7}(t) * M_{7,R}(t). && (19)
\end{aligned}
$$

一般に $\phi(s) \equiv \int_0^\infty e^{-st}d\Phi(t)$ とおき，式（16）～式（19）を LS 変換し， $m_{5,R}(s)$ について再生方程式を解くことによって，次式を得る．

$$
\begin{aligned}
m_{5,R}(s) &= \frac{y_1(s)}{1 - y_1(s)q_{R,5}(s)}, && (20)\\
x_1(s) &\equiv \frac{q_{5,6}(s)q_{6,8}(s) + q_{5,8}(s)}{1 - q_{8,6}(s)q_{6,8}(s) - q_{8,7}(s)q_{7,8}(s)},\\
y_1(s) &\equiv q_{5,6}(s)q_{6,R}(s) + x_1(s)[q_{8,6}(s)q_{6,R}(s) + q_{8,7}(s)q_{7,R}(s)].
\end{aligned}
$$

ここで，$x_1(s)$ は状態 5 から状態 R に推移せず，状態 8 に推移する経過時間分布の LS 変換形を表し，$y_1(s)$ は状態 5 から状態 R に初めて推移する経過時間分布の LS 変換形を表す．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．

$$\begin{aligned} M_R &\equiv \lim_{s\to 0} m_{5,R}(s) \\ &= \frac{y_1(0)}{1-y_1(0)q_{R,5}(0)}, \end{aligned} \tag{21}$$

ここで，

$$\begin{aligned} x_1(0) &\equiv \frac{q_{5,6}(0)q_{6,8}(0)+q_{5,8}(0)}{1-q_{8,6}(0)q_{6,8}(0)-q_{8,7}(0)q_{7,8}(0)}, \\ y_1(0) &\equiv q_{5,6}(0)q_{6,R}(0)+x_1(0)[q_{8,6}(0)q_{6,R}(0)+q_{8,7}(0)q_{7,R}(0)], \\ q_{5,6}(0) &= \frac{\alpha(\lambda_1+\alpha+\gamma)}{(\lambda_1+\alpha)(\lambda_1+\lambda_2+\alpha+\gamma)} \\ q_{5,8}(0) &= \frac{\alpha\lambda_2}{(\lambda_1+\alpha)(\lambda_1+\lambda_2+\alpha+\gamma)} \\ q_{6,R}(0) &= \frac{\beta}{\beta+\lambda_1+\lambda_2}, \\ q_{7,R}(0) &= \frac{\gamma}{\alpha+\lambda_1+\gamma}, \\ q_{6,8}(0) &= \frac{\lambda_2}{\beta+\lambda_1+\lambda_2}, \\ q_{7,8}(0) &= \frac{\alpha}{\alpha+\lambda_1+\gamma}, \\ q_{8,6}(0) &= \frac{\gamma}{\beta+\gamma+\lambda_1}, \\ q_{8,7}(0) &= \frac{\beta}{\beta+\gamma+\lambda_1}, \\ q_{R,5}(0) &= \frac{w}{\lambda_1+\lambda_2+w}. \end{aligned}$$

システムが定常状態で状態 F にある確率 P_F と状態 SW にある確率 P_{SW} を求める．システムが時刻 0 で状態 $i(i=5,6,7,8,R)$ にあり，時刻 t で状態 F にある確率分布を $P_{i,F}(t)$ とすると，次のような再生形方程式を得る．

$$P_{5,F}(t) = Q_{5,F}(t)+Q_{5,6}(t)*P_{6,F}(t)+Q_{5,8}(t)*P_{8,F}(t), \tag{22}$$
$$P_{6,F}(t) = Q_{6,8}(t)*P_{8,F}(t)+Q_{6,R}(t)*P_{R,F}(t), \tag{23}$$
$$P_{7,F}(t) = Q_{7,F}(t)+Q_{7,8}(t)*P_{8,F}(t)+Q_{7,R}(t)*P_{R,F}(t), \tag{24}$$
$$P_{8,F}(t) = Q_{8,F}(t)+Q_{8,6}(t)*P_{6,F}(t)+Q_{8,7}(t)*P_{7,F}(t), \tag{25}$$
$$P_{R,F}(t) = Q_{R,F}(t)+Q_{R,5}(t)*P_{5,F}(t), \tag{26}$$

式（22）～式（26）を LS 変換し，$p_{5,F}(s)$ について再生方程式を解くことによって，次式を得る．

$$p_{5,F}(s) = \frac{\left[\ q_{5,F}(s)+x_1(s)[q_{8,F}(s)+q_{8,7}(s)q_{7,F}(s)]+y_1(s)q_{R,F}(s)\ \right]}{1-y_1(s)q_{R,5}(s)}. \tag{27}$$

従って，システムが定常状態で状態 F にある確率 P_F は次のように求めることができる．

$$P_F \equiv \lim_{s\to 0} p_{5,F}(s)$$

$$= \frac{\left[\ q_{5,F}(0) + x_1(0)[q_{8,F}(0) + q_{8,7}(0)q_{7,F}(0)] + y_1(0)q_{R,F}(0)\ \right]}{1 - y_1(0)q_{R,5}(0)}, \tag{28}$$

ここで,

$$\begin{aligned} q_{5,F}(0) &= \frac{\lambda_1\lambda_2}{(\lambda_1+\alpha)(\lambda_1+\lambda_2+\alpha+\gamma)}, \\ q_{7,F}(0) &= \frac{\lambda_1}{\alpha+\lambda_1+\gamma}, \\ q_{8,F}(0) &= \frac{\lambda_1}{\beta+\lambda_1+\gamma}, \\ q_{R,F}(0) &= \frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+w}. \end{aligned}$$

同様にして，システムが時刻 0 で状態 $i(i = 5, 6, 7, 8, R)$ にあり，時刻 t で状態 SW にある確率分布を $P_{i,SW}(t)$ とすると，次のような再生形方程式を得る.

$$P_{5,SW}(t) = Q_{5,SW}(t) + Q_{5,6}(t) * P_{6,SW}(t) + Q_{5,8}(t) * P_{8,SW}(t), \tag{29}$$
$$P_{6,SW}(t) = Q_{6,SW}(t) + Q_{6,8}(t) * P_{8,SW}(t) + Q_{6,R}(t) * Q_{R,5}(t) * P_{5,SW}(t), \tag{30}$$
$$P_{7,SW}(t) = Q_{7,8}(t) * P_{8,SW}(t) + Q_{7,R}(t) * Q_{R,5}(t) * P_{5,SW}(t), \tag{31}$$
$$P_{8,SW}(t) = Q_{8,6}(t) * P_{6,SW}(t) + Q_{8,7}(t) * P_{7,SW}(t), \tag{32}$$

式（29）～式（32）を LS 変換し，$p_{5,SW}(s)$ について再生方程式を解くことによって，次式を得る.

$$p_{5,SW}(s) = \frac{q_{5,SW}(s) + [q_{5,6}(s) + x_1(s)q_{8,6}(s)]q_{6,SW}(s)}{1 - y_1(s)q_{R,5}(s)}. \tag{33}$$

従って，システムが，定常状態で状態 SW にある確率 P_{S_W} は次のように求めることができる.

$$\begin{aligned} P_{SW} &\equiv \lim_{s\to 0} p_{5,SW}(s) \\ &= \frac{q_{5,SW}(0) + [q_{5,6}(0) + x_1(0)q_{8,6}(0)]q_{6,SW}(0)}{1 - y_1(0)q_{R,5}(0)}, \end{aligned} \tag{34}$$

ここで,

$$\begin{aligned} q_{5,SW}(0) &= \frac{\lambda_1(\lambda_1+\alpha+\gamma)}{(\lambda_1+\alpha)(\lambda_1+\lambda_2+\alpha+\gamma)}, \\ q_{6,SW}(0) &= \frac{\lambda_1}{\beta+\lambda_1+\lambda_2}, \end{aligned}$$

明らかに，$P_F + P_{SW} = 1$ である.

3　おわりに

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．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が，ここでは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．さらにメインサイトが被災したとき，システムダウンする確率，バックアップサイトに業務が引き継がれる確率，システムダウンするまでのレプリケーション回数などを解析的に導出した.

今後は非同期型レプリケーションのモデル化を行い，期待費用を解析的に導出する．さらに，同期型レプリケーションについても期待費用を導出し，それらの期待費用の比較から，メインサイトのデータ更新の頻度によってどのレプリケーション方式が有効か等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.

（提出期日　平成 19 年 11 月 26 日）