# 大規模コンピュータシステムに対する
# 次世代信頼性評価技術の実用化に関する研究

科学研究費補助金基盤研究（B）（1）研究成果報告書
Grant No. 10558059

平成13年3月

研究代表者　尾崎　俊治
広島大学工学部第二類（電気系）教授

# 目次

## 研究目的

近年，情報通信ネットワークの発達と超並列コンピュータの実用化にともない，航空機の予約システムや銀行のキャッシュディスペンサーに代表される OLTP （オンライン・トランザクション）システムは，大規模化・複雑化の一途をたどっていると言っても過言ではない．OLTP システムのように，システムダウンによる社会的影響が極めて大きいコンピュータシステムの信頼性および保全性技術では，フォールトトレラント（耐故障）の観点からシステム全体の動特性を解析・評価することが必要とされるため，電子部品，情報伝送ソフトウェアといった個々のモジュール単位で信頼性解析を行った後，ボトムアップ的にコンピュータシステム全体の信頼性を向上させるという手法が従来までに行われてきた．しかしながら，先にも述べたように，並列かつ非同期的に情報の伝達を行うことが今後益々必要とされるコンピュータネットワークにおいては，ボトムアップ的信頼性評価技術はシステムの稼働初期段階においてのみ効力を有するが，定常段階に移行するにつれて個々のモジュールシステムの稼働特性が大幅に変化するため，システム全体が初期状態と比較して全く異なる様相を呈する．つまり並列かつ非同期的に信頼性が変化するシステムに対しては，トップダウン形式でシステム全体の信頼性を評価することが重要となり，各モジュールシステムに要求される保全技術にもダイナミックな時間変化を考慮する必要がある．本研究の目的は，コンピュータシステムを構成する (i) コンピュータハードウェア（電子材料），(ii) コンピュータソフトウェア，(iii) データ伝送技術，(iv) ネットワーク設計，(v) コンピュータ保全技術，(vi) 信頼性基礎数理，(vii) 統計的信頼性試験の 7 つの専門家グループによる共同研究を行い，大規模コンピュータシステムにおける新しい総合的な信頼性評価技術を確立することである．このような共同研究の必要性は，トップダウン的な信頼性評価を実現するためだけではなく，種々の個別領域で現在までに達成された研究成果を総括し，それらの結果を有機的に結合させ，全く新しい耐故障コンピュータシステムの設計法を実現するためには不可欠である．

## 研究方針

コンピュータシステムを構築する 7 つの異なる研究専門グループによって多角的に新しい信頼性技術を模索すると同時に，広く最新の研究動向を把握する．これらを実現するために，大規模なシンポジウムを開催し異なる分野の幅広い研究者層から意見を求める．本研究では，専門グループによって協議された指針やシンポジウムにおいて抽出される信頼性評価技術に関する成果を民間企業に所属する技術者にフィードバックすることによって，次世代コンピュータシステムへの実用化を検討する．

我が国における信頼性の基礎数理に関する研究は，主に日本オペレーションズ・リサーチ学会を母体とする研究組織において活発に行われてきた．また，信頼性技術は土木建築物，海洋構造物，メカトロニクス，コンピュータシステムといった個々の領域において独立に発展してきたという歴史的な経緯がある．本研究グループは主にオペレーションズ・リサーチ学会と電子情報通信学会信頼性研究会を活躍の場としているが，本研究計画においては日本材料学会，日本機械学会，情報処理学会，ソフトウェア科学会，日本品質管理学会において活躍する研究者を各専門小委員会の代表者（分担者）としている．本研究グループによって行われた研究業績は本研究企画以外にも「信頼性理論における確率モデルシンポジウム」（名古屋，1984），「工学・技術管理における日豪確率モデルワークショップ」（ゴールドコースト，1993，1996），「革新的生産技術における日英確率モデルワークショップ」（ケンブリッジ，1995）などがある．これらは当該研究グループが海外において運営した信頼性評価技術に関するシンポジウムであり，国内外から高い評価を得ている．

## 研究組織


| | | |
|---|---|---|
| 研究代表者 | 尾崎俊治 | 広島大学・工学部・助教授 |
| 研究分担者 | 土肥正 | 広島大学・工学部・助教授 |
| | 河合一 | 鳥取大学・工学部・教授 |
| | 山田茂 | 鳥取大学・工学部・教授 |
| | 中川覃夫 | 愛知工業大学・経営工学科・教授 |
| | 菅澤善男 | 日本大学・生産工学部・教授 |
| | 田中泰明 | 京都大学大学院・工学研究科・助教授 |
| | 海生直人 | 広島修道大学・経済科学部・教授 |
| | 三道弘明 | 流通科学大学・情報学部・教授 |
| | 木島正明 | 東京都立大学・経済学部・教授 |
| | 木村俊一 | 北海道大学・経済学部・教授 |
| | 紀一誠 | 神奈川大学・理学部・教授 |
| | 福岡浩平 | (株) アライドテレシス・システム本部・本部長 |


## 研究経費


平成 10 年度助成金額　2,400 千円
平成 11 年度助成金額　1,600 千円
平成 12 年度助成金額　1,500 千円

計　**5,500** 千円

## 活動報告

大規模コンピュータシステムに対する次世代信頼性評価技術の体系化を行った．コンピュータシステムを構築する 7 つの異なる研究専門グループによって多角的に新しい信頼性技術を検討した．コンピュータ技術については福岡，紀，菅澤，山田，データ保全技術については中川，三道，ネットワーク設計技術については木村，コンピュータ保全技術および基礎数理については河合，海生，木島，統計的信頼性については土肥，田中がぞれぞれ担当となり，全体の取りまとめは尾崎が担当した．また，大規模なシンポジウムを開催し，異なる分野の幅広い研究者層から意見を収集した．さらに，新しい信頼性評価技術を実現するために，多くの民間企業の協力も得られた．特に，専門グループによって協議された指針やシンポジウムにおいて抽出された信頼性評価技術に関する成果を民間企業に所属する技術者にフィードバックし，次世代コンピュータシステムの実用化を検討した．

### 平成 10 年度

(1) 第 1 回研究準備委員会の開催：(i) コンピュータハードウェア（電子材料）(ii) コンピュータソフトウェア (iii) データ電送技術 (iv) ネットワーク設計 (v) コンピュータ保全技術 (vi) 信頼性基礎数理 (vii) 統計的信頼性試験の 7 つの研究グループによる専門委員会を開催した．特に，各専門領域において現在までに得られた理論的かつ実証的知見を整理し，信頼性理論の歴史的変遷とコンピュータテクノロジーとの関連について資料をまとめた．また，(iv) ネットワーク設計委員会と (vi) 信頼性基礎数理委員会を軸に大規模コンピュータの信頼性評価に関する基本方針をまとめた．

(2) 第 2 回研究準備委員会の開催：(i) で決定された基本方針並びに現在収集しているデータを用いて，大規模コンピュータシステムのトップダウン的信頼性評価技術の草案を作成した．この段階においては，シミュレーションテストを除いたすべての理論的モデルの解析を完了した．各委員間の調整は E-mail も使って行った．各専門小委員会で決定された研究方針をもとに，7 つの研究グループによって独自に研究を実施した．各専門委員会に参加する研究者はおおむね 5〜6 名とし，各専門委員会代表者（分担者）がその調整を行った．

(3) 平成 10 年度信頼性技術シンポジウムの開催：7 つの専門委員会による個別領域における研究調査の報告ならびに研究準備委員会全体でまとめられたトップダウン的信頼性評価技法についての報告を行った．シンポジウムへの参加資格は特に定めないものとし，学会，大学関連企業からの参加を呼びかけた．最終的に本報告会でまとめられた成果は印刷製本され，シンポジウムへの参加者および学会などに無料送付した．また，国際共同研究の企画として，英語版の報告書を作成し，世界各国に配送した．シンポジウムでの講演者は，各専門小委員会において委託された大学研究者ならびに民間の研究者であった．また，シンポジウムでの発表に対する自由応募も並列して行い，我が国における信頼性評価技術の発展に寄与するような学会研究集会が開催できたものと考えている．

### 平成 11 年度

(4) 第 3,4 回研究準備委員会の開催：平成 10 年度に開催された準備委員会とシンポジウムで議論された内容を吟味することによって，平成 11 年度に開始されるべき研究分野の内容を調整し，研究組織を再構築した．また，各専門分野における現在までの研究成果を整理・統合することにより，信頼性技術シンポジウムの準備を行った．

(5) 平成 11 年度信頼性技術研究シンポジウムの開催：前年度のシンポジウムに引き続き，大規模コンピュータシステム信頼性に対する学術研究会を実施した．ここでは，前年度までの問題を克服し，次世代信頼性技術に対する具体的なイメージを定量化することに主眼を置き，意見交換を行った．

平成 12 年度

(6) 平成 10, 11 年度の研究成果の発表の場として平成 12 年 3 月 29, 30 日，京都ガーデンパレス (京都市) において国際会議 ASSM2000: International Conference on Applied Stochastic System Modeling を開催した．この国際会議には，ほぼ全員の研究分担者と海外から 7 名の有名研究者を中心に 50 名ほどの参加者があった．コンピュータシステムの次世代信頼性技術に関する数々の発表とそれに対する活発な討論があった．そして，その報告集として Proceedings of ASSM2000 を出版した．

(7) 平成 10, 11, 12 年の 3 年間にわたる科学研究費基盤研究の総括として平成 12 年 12 月 7, 8, 9 日の 3 日間のシンポジウムが KKR 鳥羽いそぶえ荘 (鳥羽市) で開催された．研究分担者のほぼ全員とその他の有志も参加した．3 年間の研究成果のまとめと今後の研究動向について討論を行った．

# 研究成果

I. 研究論文

1. H. Sandoh, H. Hirakoshi and T. Nakagawa, A new modified discrete preventive maintenace policy and its application to hard disk management, *Journal of Quality in Maintenance Engineering*, **4**(4), pp. 284–290, 1998.
2. T. Shibuya, T. Dohi and S. Osaki, Spare part inventory models with stochastic lead times, *International Journal of Production Economics*, **55**, pp. 257–271, 1998.
3. Y. Shinohara, Y. Nishio, T. Dohi and S. Osaki, An optimal software release problem under cost rate criterion: artificial neural network approach, *Journal of Quality in Maintenance Engineering*, **4**(4), pp. 236–247, 1998.
4. 岡村寛之，土肥正，尾崎俊治，コンピュータシステムの自動スリープ機能における省電力効果 I — 再生過程によるモデル化，情報処理学会論文誌，**39**(6), pp. 1858–1869, 1998.
5. 土肥正，西尾泰彦，篠原康秀，尾崎俊治，ニューラルネットワークを用いたソフトウェア最適リリース問題の幾何学的解法，電子情報通信学会, **J81-A**(1), pp. 110–118, 1998.
6. 土肥正，永井秀治，尾崎俊治，動径基底関数ニューラルネットワークを適用した再生関数の計算手続き，日本応用数理学会論文誌，**8**(2), pp. 13–29, 1998.
7. 藤広敏幸，土肥正，尾崎俊治，屋内不点故障発生件数の推定に関する事例研究，電子情報通信学会, **J81-A**(9), pp. 1316–1319, 1998.
8. 山田茂，影山高章，木村光宏，高橋宗雄，コードレビューにおける人的エラーと人的要因に関する考察，電子情報通信学会論文誌 A，**J81-A**(9), pp. 1238–1246, 1998.
9. T. Kurasugi and I. Kino, Approximation method for two-layer queueing models, *Performance Evaluation*, **36–37**, pp. 55–77, 1999.
10. K. Okuhara, S. Osaki and M. Kijima, Learning to design synergetic computers with an extended symmetric diffusion network, *Neural Computation*, **11**, pp. 1475–1491, 1999.
11. S. Yamada and M. Kimura, Software reliability assessment tool based on object-oriented analysis and its application, *Annals of Software Engineering*, **8**, pp. 223–238, 1999.
12. 田中泰明, 角野泰臣, 土肥正, 尾崎俊治，株価指数オプションの価格評価と相関分析，システム制御情報学会論文誌, **12**(7), pp. 379–389, 1999.
13. 林坂弘一郎，三道弘明, ソフトウェアの保守サービス契約に関する一考察, 電子情報通信学会論文誌，**J82-A**(12), pp. 1819–1829, 1999.
14. T. Dohi, N. Kaio and S. Osaki, The optimal age-dependent checkpoint strategy for a stochastic system subject to general failure mode, *Journal of Mathematical Analysis and Applications*, **249**, pp. 80–94, 2000.
15. T. Dohi, K. Takeita and S. Osaki, Graphical methods for determining/estimating optimal repair-limit replacement policies, *International Journal of Reliability, Quality and Safety Engineering*, **7**(1), pp. 43–60, 2000.
16. H. Hirakoshi and H. Sandoh, An optimal time to sleep for an auto-sleep system considering multi-usage states, *Mathematical and Computer Modelling*, **31**(10-12), pp. 157–164, 2000.
17. B. P. Iskandar, B. Klefsjo and H. Sandoh, An opportunity-based age replacement policy with warranty analyzed by using TTT-transforms, *International Journal of Reliability and Application*, **1**(1), pp. 27–38, 2000.

18. B. P. Iskandar and H. Sandoh, An extended opportunity-based age replacement policy, *RAIRO Operations Research*, **34**(2), pp. 145–154, 2000.
19. B. P. Iskandar and H. Sandoh, An opportunity-based age replacement policy considering warranty, *International Journal of Reliability, Quality and Safery Engineering*, **6**(3), pp. 229–236, 2000.
20. K. Ito and T. Nakagawa, Optimal inspection policies for a storage system with degradation at periodic tests, *Mathematical and Computer Modelling*, **31**(10-12), pp. 191–195, 2000.
21. M. Kijima, Valuation of a credit swap of the basket type, *Review of Derivatives Research*, **4**(1), pp. 81–97, 2000.
22. M. Kijima and Y. Muromachi, Credit events and the valuation of credit derivatives of the basket type, *Review of Derivatives Research*, **4**(1), pp. 55–79, 2000.
23. M. Kijima, K. Nakagawa and T. Namatame, Competitive price equilibrium when consumers have a category reservation utility, *Computational and Mathematical Organization Theory*, **6**, pp. 7–27, 2000.
24. K. Sawada and H. Sandoh, Continuous model for software reliability demonstration testing considering damage size of software failures, *Mathematical and Computer Modelling*, **31**(10-12), pp. 321–327, 2000.
25. K. Tokuno and S. Yamada, Markovian software availability measurement based on the number of restoration actions, *IEICE Transactions on Fundamentals*, **E83-A**(5), pp. 835–841, 2000.
26. 今泉充啓，安井一民，中川覃夫, シグネチャを用いたジョブ実行過程の高信頼化方策, 電子情報通信学会論文誌, **J83-A**(9), pp. 1125–1128, 2000.
27. 三道弘明, 秤の点検政策に関する研究, 電子情報通信学会論文誌, **J83-A**(3), pp. 302–308, 2000.
28. 三道弘明，中川覃夫，太田俊彦, 化学製品に対する最適計り直し量に関する一考察, オペレーションズ・リサーチ, **45**(2), pp. 76–80, 2000.
29. 田中泰明, システム信頼性解析における効率的シミュレーション解法，日本応用数理学会誌, **10**(3), pp. 229–239, 2000.
30. 永井秀治, 土肥正, 尾崎俊治, 動径基底関数ニューラルネットワークを適用した再生関数のノンパラメトリック推定, 日本応用数理学会論文誌, **10**(3), pp. 17–30, 2000.
31. 中村正治，福本聡，中川覃夫, 差分バックアップ方式における最適フルバックアップ間隔, 電子情報通信学会論文誌, **J83-D-I**(10), pp. 1087–1096, 2000.
32. 藤原隆次，山田茂，テスト習熟性を考慮したソフトウェア信頼度成長モデルとその適合性評価に関する考察，電子情報通信学会論文誌，**J83-A**(2), pp. 188–195, 2000.
33. 江崎和博，山田茂，高橋宗雄，設計レビューにおけるソフトウェア信頼性に影響を及ぼす人的要因の品質工学的解析，電子情報通信学会論文誌，**J84-A**(2), pp. 218–228, 2001.
34. 銭存華，中村正治，中川覃夫, 差分バックアップ運用を伴うデータベースシステムにおける最適バックアップ方策, 電子情報通信学会論文誌, **J84-A**(2), pp. 208–217, 2001.

II. 図書

1. R. J. Wilson, S. Osaki and M. J. Faddy (eds.), *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, Centre for Statistics Department of Mathematics, The University of Queensland, Brisbane, 1999.

2. S. Osaki (ed.), *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, Department of Industrial and Systems Engineering, Hiroshima University, Higashi-Hiroshima, 2000.

III. 口頭発表

1. T. Dohi, A. Ashioka, N. Kaio and S. Osaki, The optimal repair-time limit replacement policy with imperfect repair: Lorenz transform approach, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 91–100, 1999.
2. T. Dohi, K. Yasui, and S. Osaki, Software reliability assessment models based on cumulative Bernoulli trial processes, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 101–116, 1999.
3. M. Imaizumi, K. Yasui and T. Nakagawa, Reliability of a job execution process using signatures, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 176–181, 1999.
4. K. Ito and T. Nakagawa, Optimal self-diagnosis policy for FADEC of gas turbine engines, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 214–220, 1999.
5. M. Kimura and S. Yamada, A statistical estimation method of mean-shift for population fraction defective based on hidden-Markov modelling, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 285–294, 1999.
6. M. Kimura, K. Yasui and T. Nakagawa, Optimal checkpointing interval of a communication system with rollback recovery, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 295–304, 1999.
7. T. Kimura, A consistent diffusion approximation for finite-capacity multi-server queues, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 305–314, 1999.
8. J. Koyanagi and H. Kawai, An optimal age maintenance for an M/G/1 queueing system, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 325–331, 1999.
9. T. Nakagawa and K. Yasui, Note on reliability of a system complexity, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 377–386, 1999.
10. S. Nakamura, C. Qian, S. Fukumoto, I. Hayashi and T. Nakagawa, Optimal buckup policy for a database system with incremental and full backups, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 397–405, 1999.
11. H. Okamura, T. Dohi and S. Osaki, The phase type approximation for the optimal auto-sleep scheduling, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 426–435, 1999.
12. H. Sandoh and K. Rinsaka, Maintenance service contract model for software, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 466–475, 1999.

13. T. Satow and S. Osaki, Opportunity-based age replacement with different intensity rates, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 476–485, 1999.

14. H. Tanaka, Importance sampling simulation for a stochastic fatigue crack growth model, *Applications of Statistics and Probability (Proc. ICASP8*1999)*, edited by R. E. Melchers and M. G. Stewart, **2**, pp. 907–914, 1999.

15. K. Teramoto, T. Usami and T. Nakagawa, Startified analysis of preferences on market share of perfume, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 520–527, 1999.

16. K. Tokuno and S. Yamada, A Markovian software reliability model with a decreasing perfect debugging rate, *Proc. First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 528–538, 1999.

17. I. Kino, Stack distance distribution in LRU, *Proc. Informs-KORMS*, 2000.

18. 蔵杉俊康，紀一誠，2層型待ち行列網モデルの近似等価流量法による解析，シンポジウム報文集（情報通信ネットワークの新しい性能評価法に関する総合研究），pp. 122–131，2000.

IV. ASSM2000 講演論文

1. M. Abdel-Hameed, Optimal control of dams using $P^M_{\lambda,\tau}$ policies and penalty cost, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 1–6

2. K. Ano, A Poisson arrival selection problem for gamma prior density with parameter $r = 2$, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 7–14

3. E. Çinlar, On conditional Levy processes, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 15–20

4. R. Dekker, Marginal cost analysis for opportunity maintenance models, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, p. 21.

5. T. Dohi, N. Kaio and S. Osaki, A new graphical method to estimate the optimal repair–time limit with incomplete repair and discounting, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 22–30

6. V. Girardin, Relative entropy and covariance type constraints yielding ARMA models, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 31–40

7. K. Goševa-Popstojanova and K. S. Trivedi, Architecture based software reliability, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 41–52

8. K. Hanada and T. Kimura, Numerical valuation of a switched knockout option, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 53–60

9. S. L. Ho, M. Xie and T. N. Goh, A study of the connectionist models for software reliability prediction, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 61–68

10. H. Hohjo and Y. Teraoka, On some competing inventory problem, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 69–74

11. M. Imaizumi, K. Yasui and T. Nakagawa, Optimal reset number of a microprocessor system with network processing, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 75–82

12. K. Ito and T. Nakagawa, Optimal self-diagnosis policy for dual redundant FADEC of gas turbine engines, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 83–87

13. H. Iwaki, M. Kijima and K. Komoribayashi, Optimal life insurance and portfolio choice in a life cycle, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 88–96

14. H. Katagiri and H. Ishii, Linear programming problem under fuzziness and randomness, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 97–106

15. H. Kawai and H. Sandoh, An efficient backup warning policy for a hard disk, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 107–117

16. M. Kawai and M. Tamaki, Choosing either the best or the second best when the number of applicants is random, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 118–125

17. M. Kimura, K. Yasui, T. Nakagawa and N. Ishii, Reliability of a communication system with limited number of rollback, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 126–133

18. V. S. Korolyuk and N. Limnios, Poisson approximation of increment process with Markov switching, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 134–139

19. J. Koyanagi and H. Kawai, An optimal join policy to the queue in processing two kinds of jobs, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 140–147

20. M. Kurano, M. Yasuda, J. Nakagami and Y. Yoshida, A fuzzy treatment of uncertain Markov decision processes, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 148–157

21. S. Nakagawa, S. Fukumoto and N. Ishii, Optimal checkpointing intervals for a double modular redundancy with signatures, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 158–164

22. S. Nakamura, C. Qian, I. Hayashi and T. Nakagawa, An optimal maintenance time of automatic monitoring system of ATM with two kinds of breakdowns, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 165–171

23. T. Namekata and T. S. H. Driessen, Bargaining property of nucleolus and $\tau$-value in a class of TU-games, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 172–179

24. H. Okamura, T. Dohi and S. Osaki, A structural approximation method to generate the optimal auto-sleep schedule for a computer system, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 180–187

25. C. Qian, S. Nakamura and T. Nakagawa, Replacement policies for a shock model with maintenance and minimal repair, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 188–195

26. H. Sandoh and N. Igaki, Optimal inspection policies for a scale, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 196–205

27. T. Satow and S. Osaki, Optimal replacement policies for a two-unit system with shock damage interaction, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 206–215

28. K. Sawaki, Optimal policies in continuous time inventory control models with limited supply, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 216–223

29. K. Teramoto, T. Usami, H. Yamada and T. Nakagawa, Perceptual positioning of soft drinks on the Japan market, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 224–231

30. Y. Teraoka, S. Osumi and H. Hohjo, On some Stackelberg type location game, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 232–239

31. K. Tokuno and S. Yamada, On relationship between software availability measurement and the number of restorations, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 240–250

32. M. Toyama, S. Tomita and Y. Yoshitomi, GA approach optimizing development of thin-walled objects, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 251–258

33. Y. Yoshida, M. Yasuda, J. Nakagami and M. Kurano, A multi-objective optimal stopping problem in a stochastic and fuzzy environmnet, *Proc. ASSM2000: International Conference on Applied Stochastic System Modeling*, pp. 259–266

資料
(ASSM2000 講演論文)

# A NEW GRAPHICAL METHOD TO ESTIMATE THE OPTIMAL REPAIR–TIME LIMIT WITH INCOMPLETE REPAIR AND DISCOUNTING

Tadashi Dohi[†], Naoto Kaio[‡] and Shunji Osaki[†]

[†]Department of Industrial and Systems Engineering, Hiroshima University,
4-1 Kagamiyama 1 Chome, Higashi-Hiroshima, 739-8527, Japan

[‡]Department of Economic Informatics, Hiroshima Shudo University,
1-1-1 Ozukahigashi, Asaminami-ku, Hiroshima 731-3195, Japan

dohi@gal.sys.hiroshima-u.ac.jp / kaio@shudo-u.ac.jp /
osaki@gal.sys.hiroshima-u.ac.jp

**Abstract**—In this paper, we consider a repair-time limit replacement problem with imperfect repair and discounting, and focus on the problem to determine the optimal repair-time limit which minimizes the expected total discounted cost over an infinite time horizon. Based upon a sophisticated graphical idea, we develop a non-parametric method to estimate the optimal repair-time limit from the empirical repair-time data. Numerical examples are devoted to estimate the optimal policy and to examine the asymptotic properties of the estimator.

**Keywords**—maintenance optimization, repair limit replacement policy, incomplete repair, discounting, non-parametric estimation

## 1. INTRODUCTION

Since the seminal contribution by Hastings [1], a large number of repair limit replacement problems were analyzed in the literature. This paper concerns a different type of repair-time limit replacement problem with imperfect repair from Nguyen and Murthy [2]. More specifically, consider a single-unit system where each spare is provided only by an order after a lead time and each failed unit is repairable. When the unit fails, one estimates the completion time of repair, which may be a possibly subjective one. If one estimates that the repair is completed up to a prespecified time-limit at the failure point of time, then the repair is started immediately, otherwise, the spare unit is ordered with a lead time. Since the repair is imperfect, the repaired unit or the ordered one fails again during a finite time horizon. Nakagawa and Osaki [3] considered the similar problem to determine the optimal repair-time limit which minimizes the expected cost per unit time in the steady-state, though they did not take account of the imperfect repair.

Since the knowledge on the repair-time distribution is incomplete in general, any statistical estimation method for the optimal repair-time limit will be needed in practical situations. Dohi *et al.* [4] developed a non-parametric method to estimate the optimal repair-time limit applying the total time on test (TTT) statistics for the problem with imperfect repair by Nguyen and Murthy [2]. Also, Dohi *et al.* [5] showed that the TTT method is applicable to a repair-cost limit replacement problem with imperfect repair. Recently, Dohi *et al.* [6] proposed a new graphical method based on the Lorenz transform [7, 8], for the repair-time limit replacement problem with imperfect repair under the expected cost criterion per unit time in the steady-state.

However, if the maintenance operation is performed for a sufficiently large planning horizon, then it is important to take account of an effect of discount factor in estimating the operating cost.

In other words, it will be useful under a fluctuating economic circumstance that the repair-time limit schedule is designed so as to minimize the expected total discounted cost over an infinite time horizon. In fact, for the repair limit replacement problem with discounting, the Lorenz method in [6] can not be applied directly to determine the optimal repair limit policy. Main purpose of this paper is to develop a new statistical method for the optimal repair-time limit under the expected total discounted cost criterion, from the complete sample of repair-time data.

The paper is organized as follows. In Section 2, we describe the repair-time limit replacement problem under consideration and define the notation and assumptions. In Section 3, the optimal repair-time limit is analytically derived under the assumption that the knowledge on the repair-time distribution is complete. In Section 4, the underlying problem to seek the optimal repair limit replacement policy is translated to a graphical one. Then, the similar but somewhat different geometrical idea from [6] is introduced and plays an important role for the translation. Next, we develop a non-parametric statistical estimation method for the optimal repair-time limit from the empirical data. It is found that the repair limit problem with discounting shows quite different aspects from the non-discounting problem. Numerical examples are devoted to illustrate the asymptotic behaviour of estimates for the optimal repair-time limit and the corresponding minimum expected cost in Section 5. Finally, the paper is concluded with some remarks in Section 6.

## 2. MODEL DESCRIPTION

The repair time $X$ for each unit is a non-negative i.i.d. random variable. The decision maker has a *subjective* probability distribution function $\Pr\{X \leq t\} = G(t)$ on the repair time, with density $g(t)$ $(> 0)$ and finite mean $1/\mu$ $(> 0)$. Suppose that the distribution function $G(t) \in [0, 1]$ is arbitrary, continuous and strictly increasing in $t \in [0, \infty)$, and in addition has an inverse function, *i.e.* $G^{-1}(\cdot)$. Suppose that the time to failure for a repaired unit, $Y_1$, is a non-negative i.i.d. random variable having the probability distribution function $F_1(t)$ with density function $f_1(t)$ and finite mean $1/\lambda_1$ $(> 0)$. Also, the time to failure for a new (spare) unit, $Y_2$, is a non-negative i.i.d. random variable having the probability distribution function $F_2(t)$ with density function $f_2(t)$ and finite mean $1/\lambda_2$ $(> 0)$. Further, we define:

$t_0 \in [0, \infty)$: repair-time limit (decision variable)

$k_f$ $(> 0)$: penalty cost per unit time when the system is in down state

$k_r$ $(> 0)$: repair cost per unit time

$c$ $(> 0)$: fixed cost associated with the ordering of a new unit

$L$ $(> 0)$: lead time for delivery of a new unit

$\beta$ $(> 0)$: discount factor

Consider a single-unit repairable system, where each spare is provided only by an order after a lead time $L$ and each failed unit is repairable. When the unit has failed at time $t = 0$, the decision maker wishes to determine whether he or she should repair it or should order a new spare. If the decision maker estimates that the repair is completed within a prespecified time limit $t_0 \in [0, \infty)$, then the repair is started immediately at $t = 0$ and completes at time $t = X$. After the completion of repair, the unit is started to operate again, but fails again for a finite time span since the repair is imperfect.

On the other hand, if the decision maker estimates that the repair time exceeds the time limit $t_0$, then the failed unit is scrapped at time $t = 0$ and a new spare unit is ordered immediately. A

new unit is delivered after the lead time $L$. Further, the new unit also fails for a finite time span. Without any loss of generality, it is assumed that the time required for replacement is negligible. Under these model setting, we define the interval from the failure point of time to the following failure time as one cycle. Figure 1 depicts the configuration of the repair limit repacement model under consideration.

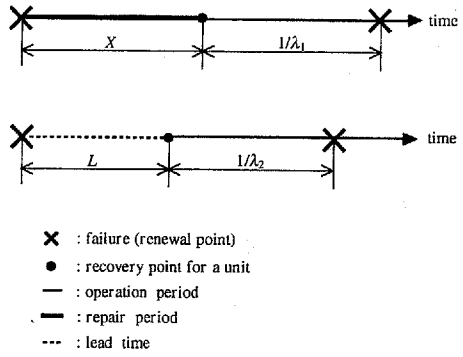


Figure 1: Configuration of the repair limit replacement model.

We make the following additional assumptions:

**(A-1)** $(k_f + k_r)\Big[\mathcal{L}\{f_2(\beta)\}\exp(-\beta L) - \mathcal{L}\{f_1(\beta)\}\Big] + \mathcal{L}\{f_1(\beta)\}\Big[k_f\{1-\exp(-\beta L)\} + c\beta\exp(-\beta L)\Big]$
$> 0,$

**(A-2)** $(k_f + k_r)\Big[1 - \mathcal{L}\{f_2(\beta)\}\exp(-\beta L)\Big] > \beta\Big[k_f\{1-\exp(-\beta L)\}/\beta + c\exp(-\beta L)\Big],$

where $\mathcal{L}\{f_i(\beta)\} = \int_0^\infty \exp(-\beta x)dF_i(x)$ $(i = 1, 2)$ is the Laplace-Stieltzes transform of $F_i(t)$. These assumptions might be somewhat technical, but will be needed to prove the unique optimal repair-time limit.

## 3. EXPECTED TOTAL DISCOUNTED COST

Let us formulate the expected total discounted cost over an infinite time horizon. If the decision maker judges that a new spare unit should be ordered, then the ordering cost for one cycle is $\int_{t_0}^\infty c\exp(-\beta L)dG(t)$. In this case, the expected penalty cost for one cycle is $\int_{t_0}^\infty \int_0^L k_f \exp(-\beta x)dx\, dG(t)$. On the other hand, if he or she selects the repair option, the expected penalty cost for one cycle is $\int_0^{t_0}\int_0^t k_f\exp(-\beta x)dxdG(t)$ and the expected repair cost for one cycle is $\int_0^{t_0}\int_0^t k_r\exp(-\beta x)dx\, dG(t)$. Thus, the expected discounted cost during one cycle is

$$V(t_0) = \frac{(k_r + k_f)}{\beta}\int_0^{t_0}\{1-\exp(-\beta t)\}dG(t) + \Big[\frac{k_f\{1-\exp(-\beta L)\}}{\beta} + c\exp(-\beta L)\Big]\overline{G}(t_0), \qquad (1)$$

where in general $\overline{\psi}(\cdot) = 1 - \psi(\cdot)$. Also, the discounted value of unit cost after one cycle becomes

$$\begin{aligned}\delta(t_0) &= \int_0^{t_0}\int_0^\infty \exp(-\beta[t+x])dF_1(x)dG(t) + \int_{t_0}^\infty\int_0^\infty \exp(-\beta[L+x])dF_2(x)dG(t)\\ &= \mathcal{L}\{f_1(\beta)\}\int_0^{t_0}\exp(-\beta t)dG(t) + \mathcal{L}\{f_2(\beta)\}\exp(-\beta L)\overline{G}(t_0). \qquad (2)\end{aligned}$$

Then the expected total discounted cost over an infinite time horizon is

$$TC(t_0) = \sum_{n=0}^{\infty} V(t_0)\delta(t_0)^n = V(t_0)/\overline{\delta}(t_0), \tag{3}$$

and the problem is to determine the optimal repair-time limit $t_0^* \in [0, \infty)$ satisfying

$$TC(t_0^*) = \min_{0 \le t_0 < \infty} TC(t_0). \tag{4}$$

Differentiating $TC(t_0)$ with respect to $t_0$ yields

$$\frac{d}{dt_0} TC(t_0) = \frac{g(t_0)}{\overline{\delta}(t_0)^2} \cdot q_0(t_0), \tag{5}$$

where

$$\begin{aligned} q_0(t_0) &= \Big[\frac{(k_r + k_f)}{\beta}\Big\{1 - \exp(-\beta t_0)\Big\} - \frac{k_f\{1 - \exp(-\beta L)\}}{\beta} - c\exp(-\beta L)\Big]\overline{\delta}(t_0) \\ &\quad -\Big[\mathcal{L}\{f_2(\beta)\}\exp(-\beta L) - \mathcal{L}\{f_1(\beta)\}\exp(-\beta t_0)\Big]V(t_0). \end{aligned} \tag{6}$$

We have the following result to guarantee the existence of the optimal repair-time limit analytically.

**Theorem 1:** Under the assumptions (A-1) and (A-2), there exists a finite and unique optimal repair-time limit $t_0^*$ $(0 < t_0^* < \infty)$ which satisfies the non-linear equation $q_0(t_0^*) = 0$, and the minimum expected total discounted cost becomes

$$TC(t_0^*) = \frac{(k_r + k_f)\{1 - \exp(-\beta t_0^*)\}/\beta - k_f\{1 - \exp(-\beta L)\}/\beta - c\exp(-\beta L)}{\mathcal{L}\{f_2(\beta)\}\exp(-\beta L) - \mathcal{L}\{f_1(\beta)\}\exp(-\beta t_0^*)}. \tag{7}$$

**Proof:** Differentiating $q_0(t_0)$ with respect to $t_0$ yields

$$\frac{d}{dt_0} q_0(t_0) = \exp(-\beta t_0) Z(t_0), \tag{8}$$

where

$$Z(t_0) = (k_r + k_f)\overline{\delta}(t_0) - \beta\mathcal{L}\{f_1(\beta)\}V(t_0). \tag{9}$$

The further differentiation yields

$$\begin{aligned} \frac{d}{dt_0} Z(t_0) &= g(t_0)\Big[(k_f + k_r)\Big\{\mathcal{L}\{f_2(\beta)\}\exp(-\beta L) - \mathcal{L}\{f_1(\beta)\}\Big\} \\ &\quad +\mathcal{L}\{f_1(\beta)\}\Big\{k_f\{1 - \exp(-\beta L)\} + c\beta\exp(-\beta L)\Big\}\Big] > 0, \end{aligned} \tag{10}$$

which is due to (A-1). Also, since

$$\begin{aligned} Z(0) &= (k_f + k_r)\Big\{1 - \mathcal{L}\{f_2(\beta)\}\exp(-\beta L)\Big\} \\ &\quad -\beta\mathcal{L}\{f_1(\beta)\}\Big\{k_f\{1 - \exp(-\beta L)\}/\beta + c\exp(-\beta L)\Big\} > 0 \end{aligned} \tag{11}$$

from (A-2), we have $Z(t_0) > 0$. From this result, it is seen that the function $TC(t_0)$ is strictly convex in $t_0$. Further, it is straightforward to confirm

$$q_0(0) = -\Big[k_f\{1-\exp(-\beta L)\}/\beta + c\exp(-\beta L)\Big]\Big[1-\mathcal{L}\{f_1(\beta)\}\Big] < 0 \tag{12}$$

and

$$\begin{aligned} q_0(\infty) &= \frac{\mathcal{L}\{g(\beta)\}}{\beta}\Big[(k_f+k_r)\Big\{\mathcal{L}\{f_2(\beta)\}\exp(-\beta L) - \mathcal{L}\{f_1(\beta)\}\Big\} + \mathcal{L}\{f_1(\beta)\}\Big\{k_f\{1-\exp(-\beta L)\} \\ &\quad + c\beta\exp(-\beta L)\Big\}\Big] + \frac{1}{\beta}\Big[(k_f+k_r)\Big\{1-\mathcal{L}\{f_2(\beta)\}\exp(-\beta L)\Big\} - \beta\Big\{k_f\{1-\exp(-\beta L)\}/\beta \\ &\quad + c\exp(-\beta L)\Big\}\Big] > 0 \end{aligned} \tag{13}$$

from both (A-1) and (A-2), where $\mathcal{L}\{g(\beta)\} = \int_0^\infty \exp(-\beta x)dG(x)$. The proof is completed.

In the following section, we develop a graphical method for the repair-time limit replacement problem, applying the concept of the similar idea to the Lorenz curve [7, 8]. The result is applied directly to a statistical non-parametric problem to estimate the optimal repair-time limit from the empirical repair-time data.

# 4. GRAPHICAL METHOD

Define the following transform;

$$\phi_\beta(p) \equiv 1 - \int_0^{G^{-1}(p)} \exp(-\beta x)dG(x), \tag{14}$$

where

$$G^{-1}(p) = \inf\{t_0 \mid G(t_0) \geq p\}, \quad 0 \leq p \leq 1. \tag{15}$$

From a few algebraic manipulation, we obtain

$$TC(t_0) = TC(p) = \frac{a_1\phi_\beta(p) + a_2(p-1)}{a_3\phi_\beta(p) + a_4 p + a_5}, \tag{16}$$

where $a_1 = (k_r+k_f)/\beta$ $(>0)$, $a_2 = k_r/\beta + \{k_f/\beta - c\}\exp(-\beta L)$ $(>0)$, $a_3 = \mathcal{L}\{f_1(\beta)\}$ $(>0)$, $a_4 = \mathcal{L}\{f_2(\beta)\}\exp(-\beta L)$ $(>0)$ and $a_5 = 1 - \mathcal{L}\{f_1(\beta)\} - \mathcal{L}\{f_2(\beta)\}\exp(-\beta L)$. Hence, the optimization problem in Eq.(4) can be rewritten by $\min_{0\leq p\leq 1} TC(p)$.

**Lemma 1:**

$$a_2 - a_1 a_4/a_3 < 0. \tag{17}$$

**Theorem 2:** Under the assumptions (A-1) and (A-2), the minimization problem in Eq.(4) is equivalent to

$$\max_{0\leq p\leq 1} : \frac{\xi_\beta(p) + \alpha}{p+\zeta}, \tag{18}$$

where $\xi_\beta(p) = \{1 - \phi_\beta(p)\}/\mathcal{L}\{g(\beta)\}$,

$$\alpha = -\frac{1}{\mathcal{L}\{g(\beta)\}}\left\{1 + \frac{a_2 a_4 + a_1 a_4 a_5/a_3}{a_2 a_3 - a_1 a_4} + \frac{a_5}{a_3}\right\} \tag{19}$$

and

$$\zeta = -\frac{a_2 + a_1 a_5/a_3}{a_2 - a_1 a_4/a_3}. \tag{20}$$

For the proof, see Dohi, *et al.* [4]. The theorem above means as follows. In the two-dimensional plane $(x, y) \in \mathcal{R}^2$, plot the curve $(p, \xi_\beta(p)) \in [0,1] \times [0,1]$ and the point $\mathrm{B}(-\zeta, -\alpha)$, where $\alpha > 0$ and $\zeta > 0$ from the assumptions. Then the problem is the determination of $p^*$ to give the maximum slope from the point B to the curve $(p, \xi_\beta(p))$. Since there exists the inverse function $G^{-1}(\cdot)$, the optimal repair-time limit is given by $t_0^* = G^{-1}(p^*)$. This result is essentially same as Theorem 1, but it is interesting that one can graphically obtain the optimal repair-time limit when the repair-time distribution is completely known. In other words, this graphical idea becomes an important hint to develop a non-parametric method to estimate the optimal repair-time limit replacement policy from the empirical repair-time data.

Next, we propose a statistical method to estimate the optimal repair-time limit using an empirical curve from complete samples on the repair time. Suppose that the optimal repair-time limit has to be estimated from an ordered complete sample $0 = x_0 \le x_1 \le x_2 \le \cdots \le x_n$ of repair times from an absolutely continuous repair-time distribution $G$, which is unknown. The estimator of the repair-time distribution should be the following empirical distribution;

$$G_{in}(x) \equiv \begin{cases} i/n & \text{for } x_i \le x < x_{i+1}, \\ 1 & \text{for } x_n \le x, \end{cases} \tag{21}$$

where $i = 0, 1, 2, \ldots, n-1$. On the other hand, the non-parametric estimator of $\xi_\beta(p)$ in Eq.(18) is defined by

$$\xi_{in}^\beta = (1 - \phi_{in}^\beta)/\left\{1 - \beta \sum_{j=1}^{n}\left[1 - \frac{j-1}{n}\right](x_j - x_{j-1})\exp(-\beta x_j)\right\} \tag{22}$$

where

$$\phi_{in}^\beta = \exp(-\beta x_i)\left\{1 - i/n + \beta \sum_{j=1}^{i}\left[1 - \frac{j-1}{n}\right](x_j - x_{j-1})\right\}. \tag{23}$$

By plotting the point $(i/n, \xi_{in}^\beta)$, $i = 0, 1, 2, \cdots, n$, and connecting them by line segments, we obtain the sample curve.

The following result is the empirical counterpart of Theorem 2.

**Theorem 3:** The estimator of the optimal repair-time limit which minimizes the expected total discounted cost over an infinite time horizon is $\hat{t}_0^* = x_i^*$, where

$$\left\{x_i^* \middle| \max_{0 \le i \le n} \frac{\xi_{in}^\beta + \alpha}{i/n + \zeta}\right\}. \tag{24}$$

Of our next interest is the convergence speed of the estimators $\hat{t}_0^*$ and $C(\hat{t}_0^*)$. We examine numerically the strong consistent property of the estimator $\hat{t}_0^*$ in the following section.

# 5. NUMERICAL ILLUSTRATIONS

In this section, we present three examples to understand the graphical and statistical methods proposed in the previous sections.

**Example 1:** Suppose that the repair-time distribution $G(t)$ is known and obeys the Weibull distribution;

$$G(t) = 1 - \exp\left\{-\left(\frac{t}{\theta}\right)^{\gamma}\right\} \tag{25}$$

with the shape parameter $\gamma = 1.5$ and the scale parameter $\theta = 1.2$. The other model parameters are $c = 10.0000(\$)$, $L = 5.0000$ (day), $k_f = 3.0000$ (\$/day) $k_r = 1.2000$ (\$/day) and $\beta = 0.0500$. The determination of the optimal repair-time limit is presented in Fig. 2. In this case, we have B $= (-0.8524, -0.8540)$ and the optimal point with maximum slope from B is $(p^*, \xi_\beta(p^*)) = (0.3530, 0.5167)$. Thus, the optimal repair-time limit and the minimum expected cost are $t_0^* = G^{-1}(0.3530) = 13.1971$ (day) and $TC(t_0^*) = 79.2792$ (\$), respectively.

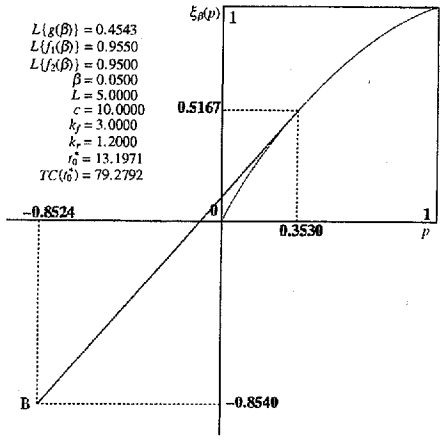


Figure 2: Determination of the optimal repair-time limit.

**Example 2:** The repair-time data are made by the random number following the Weibull distribution with shape parameter $\gamma = 1.5$ and scale parameter $\theta = 1.2$. The other model parameters are same as Example 1 except that $\mathcal{L}\{g(\beta)\} = 0.5369$. The sample curve based on the 20 sample data is shown in Fig. 3. Since B $= (-0.8524, -0.7226)$, the optimal point with maximum slope from B becomes $(i^*/n, \xi_{i^*n}^{\beta}) = (11/20, \xi_{11,20}^{\beta}) = (0.5500, 0.7273)$. Hence, the estimates of the optimal repair-time limit and the minimum expected cost are $\hat{t}_0^* = 17.1523$ (day) and $TC(\hat{t}_0^*) = 81.5756$ (\$), respectively.

**Example 3:** Suppose that the repair-time distribution and model parameters are similar to those in Example 2. Then the real optimal repair-time limit and the minimum expected cost become $t_0^* = 13.1971$ (day) and $TC(t_0^*) = 79.2792$ (\$), respectively. On the other hand, the asymptotic behaviour of estimates for the optimal repair-time limit and the corresponding minimum expected cost are depicted in Figs. 4 and 5, respectively. From these figures, we observe that the estimates converge to the corresponding real optima around where the number of data is 30. In other words, without specifying the repair-time distribution, the proposed non-parametric method may function well to estimate the optimal repair-time limit precisely.

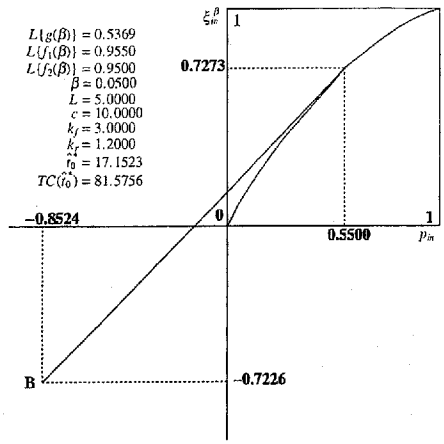


Figure 3: Estimation of the optimal repair-time limit.

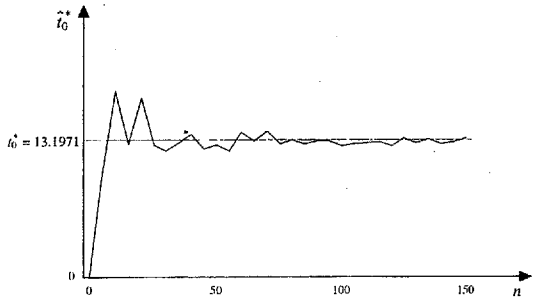


Figure 4: Asymptotic behaviour of the estimate for the optimal repair-time limit.

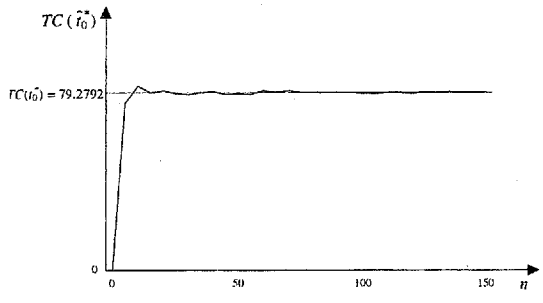


Figure 5: Asymptotic behaviour of the estimate for the minimum expected total discounted cost.

## 6. CONCLUSION

In this paper, we have developed a new graphical device to calculate the optimal repair-time limit replacement policy with imperfect repair and discounting. The basic idea is similar to the classical one by the TTT transform, but it should be noted that the underlying maintenance problem is quite different from the TTT-based problem. In numerical examples, it has been observed that the non-parametric method based on the empirical curve has nice convergence properties, although any estimator related with the empirical distribution does not converge to the real optimal at earlier phase and more than 50 data are needed to get the satisfactory estimate from our experiences. In that sense, the method proposed here will be useful to estimate the optimal repair-time limit replacement policy in practice.

## REFERENCES


1. N. A. J. Hastings, The repair limit replacement method, *Operational Research Quarterly*, **20**, 337-349 (1969).

2. D. G. Nguyen and D. N. P. Murthy, Optimal repair limit replacement policies with imperfect repair, *Journal of Operational Research Society*, **32**, 409-416 (1981).

3. T. Nakagawa and S. Osaki, Optimum ordering policies with lead time for an operating unit, *R.A.I.R.O. Recherche opérationnelle/Operations Research*, **12**, 383-393 (1978).

4. T. Dohi, N. Matsushima, N. Kaio and S. Osaki, Nonparametric repair limit replacement policies with imperfect repair, *European Journal of Operational Research*, **96**, 260-273 (1996).

5. T. Dohi, N. Kaio and S. Osaki, A graphical method to repair-cost limit replacement policies with imperfect repair, to appear in *Mathematical and Computer Modelling*, (1999).

6. T. Dohi, A. Ashioka, N. Kaio and S. Osaki, The optimal repair-time limit replacement policy with imperfect repair: Lorenz transform approach, *Proceedings of the First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, R. J. Wilson, S. Osaki and M. J. Faddy (eds.), 91-100 (1999).

7. M. O. Lorenz, Methods of measuring the concentration of wealth, *Journal of American Statistical Association*, **9**, 209-219 (1893).

8. J. L. Gastwirth, A general definition of the Lorenz curve, *Econometrica*, **39**, 1037-1039 (1971).

# NUMERICAL VALUATION OF A SWITCHED KNOCKOUT OPTION

KUNIO HANADA and TOSHIKAZU KIMURA
Graduate School of Economics, Hokkaido University
Sapporo 060-0809, Japan
hanada@pop.econ.hokudai.ac.jp / kimura@econ.hokudai.ac.jp

**Abstract**—Knockout options are kinds of exotic contingent claims whose right to exercise is nullified when the underlying asset price hits a knockout boundary. Beginning with a mathematical model of Merton (1973), some extended models have been developed for the knockout options, under a common assumption that the knockout boundary exists in the whole trading interval. In this paper, however, we consider a new European knockout option whose knockout boundary exists only in a certain part of the trading interval, so that we call it a *switched knockout option*. Extensive numerical experiments show that the switched knockout options have quite different properties from the ordinary knockout as well as vanilla options, especially on the sensitivity with volatility.

**Keywords**—switched knockout options; incomplete knockout boundary; European call/put; numerical valuation; Crank-Nicolson method

## 1. INTRODUCTION

For a vanilla European option, the payoff at exercise can be determined by the spot price of the underlying asset, independently on its past history in the trading interval. The so-called *exotic* or *path-dependent* options have values that depend on the history of the asset price in some non-trivial way. Among various exotic options, we focus on a knockout option with an incomplete boundary in this paper.

Knockout options are contingent claims whose right to exercise is nullified when the undelying asset value crosses a certain value. The set of those values over the trading interval is called a knockout boundary. Knockout options are classified as either *up-and-out* or *down-and-out* options by the relative position between initial values of the asset price and the knockout boundary. Of course, they are classified into two basic types, *i.e,* call *or* put. Hence, there are totally four different types in knockout options: When the initial price is below the knockout boundary, there are up-and-out calls and puts. On the other hand, when the initial price is above the knockout boundary, there are down-and-out calls and puts.

In Merton [1], he has first studied a basic mathematical model of down-and-out European knockout options to obtain closed pricing formulas under an assumption that the knockout boundary is an exponential function of remaining time to maturity. Rubinstein and Reiner [2] and Rich [3] developed pricing formulas for all types of the basic knockout options. Rich also examined comparative statistics for these formulas. In addition, more general knockout options have been proposed by many researchers: Cox and Rubinstein [4] dealt with a down-and-out European knockout option with a rebate, whose holder can receive a specified amount of money if the boundary is crossed. Kunitomo and Ikeda [5] and Geman and Yor [6] obtained pricing formulas for knockout options with two seperate boundaries that are located above and below the asset price at the initial time.

Roberts and Shortland [7] analyzed the option price under an assumption that both of the drift and volatility parameters are functions of time. Linetsky [8] proposed a new-type knockout option called a step option, which is not instantaneously nullified when the asset price hits the knockout boundary. These basic and generalized knockout options above have exponential knockout boundaries. Recently, Hanada and Kimura [9] developed an approximate pricing formula for a knockout option with a general class of non-exponential knockout boundaries.

All of the previous results are based on a common assumption that the knockout boundary exists in the whole trading interval from initial time to maturity. In this paper, however, we consider an incomplete knockout boundary that exists only in a certain part of the trading interval. In other words, there is a toggled *switch* in the knockout boundary; this option is equivalent to a vanilla or an ordinary knockout option according as the switch is *off* or *on*. Hence, we call it a *switched knockout option* in this paper. Obviously, the vanilla and ordinary knockout options are special cases of our switched knockout option.

This paper is organized as follows: In Section 2, we mathematically specify the switched knockout option to show that its price at arbitrary time satisfies a partial differential equation together with some boundary conditions. In Section 3, we numerically solved this equation by the Crank-Nicolson method to examine general properties of switched knockout options. To avoid redundancy, we are mainly concerned with the analysis of the up-and-out call option, but we also refer to some general properties of other three types shortly. Finally, in Section 4, we give a few concluding remarks.

## 2. MATHEMATICAL FORMULATION

We use the same assumptions as those in the Black-Scholes model [10] except for knockout boundaries: Assume that the capital market is well-defined and follows the efficient market hypothesis. Let $S(t)$ denote the underlying asset price at time $t$ and let $T$ ($\geq 0$) be the maturity. Then, the process $\{S(t);\ 0 \leq t \leq T\}$ satisfies the stochastic differential equation

$$\frac{dS(t)}{S(t)} = \mu dt + \sigma dW(t), \qquad 0 \leq t \leq T, \tag{1}$$

where $\mu$ ($\sigma$) is the drift (volatility) of the process $S(\cdot)$ and $r$ is the risk-free interest rate, all of which are assumed to be positive constants. In (1), $\{W(t);\ 0 \leq t \leq T\}$ is the standard Brownian motion process, so that the process $S(\cdot)$ becomes a geometric Brownian motion. Also, assume that the option price written on $S(t)$, say $V$, is a function of $S(t)$ and $t$, *i.e.*, $V \equiv V(S(t), t)$ for $S(t) > 0$ and $0 \leq t \leq T$. From these assumptions and Itô's lemma, we have the partial differential equation

$$\frac{1}{2}\sigma^2 S(t)^2 \frac{\partial^2 V(S(t),t)}{\partial S(t)^2} + rS(t)\frac{\partial V(S(t),t)}{\partial S(t)} - rV(S(t),t) + \frac{\partial V(S(t),t)}{\partial t} = 0, \qquad S(t) > 0, \quad 0 \leq t \leq T; \tag{2}$$

see Harrison and Pliska [11] or Øksendal [12].

For a vanilla call option with the exercise price $K$ ($> 0$), the option price $V$ satisfies the terminate condition

$$V(S(T), T) = \max(S(T) - K, 0), \tag{3}$$

together with the boundary conditions

$$\lim_{\xi \to \infty} \frac{V(\xi, t)}{\xi - Ke^{-r(T-t)}} = 1, \qquad 0 \leq t \leq T \tag{4}$$

and

$$\lim_{\xi \to 0} V(\xi, t) = 0, \qquad 0 \leq t \leq T. \tag{5}$$

For a switched knockout option, however, these boundary conditions should be modified as follows: Let $\mathcal{I}_{\text{on}}$ be the set of time intervals where the nullified switch is on, and let $\mathcal{I}_{\text{off}} \equiv [0,\ T] \backslash \mathcal{I}_{\text{on}}$. Let $B(t)$ be the value of knockout boundary at time $t$ and assume that $B(t) > 0$ for $t \in [0, T]$. Then, for the up-and-out call type, the boundary conditions should be

$$\begin{aligned} &\lim_{\xi \to \infty} \frac{V(\xi, t)}{\xi - Ke^{-r(T-t)}} = 1, \quad t \in \mathcal{I}_{\text{off}}, \\ &V(\xi, t) = 0, \qquad\qquad (\xi,\ t) \in [B(t),\ \infty) \times \mathcal{I}_{\text{on}}, \\ &\lim_{\xi \to 0} V(\xi, t) = 0, \qquad 0 \leq t \leq T, \end{aligned} \tag{6}$$

whereas, for the down-and-out call type, the boundary conditions are given by

$$\begin{aligned} &\lim_{\xi \to \infty} \frac{V(\xi, t)}{\xi - Ke^{-r(T-t)}} = 1, \quad 0 \leq t \leq T, \\ &\lim_{\xi \to 0} V(\xi, t) = 0, \qquad t \in \mathcal{I}_{\text{off}}, \\ &V(\xi, t) = 0, \qquad\qquad (\xi,\ t) \in [0,\ B(t)] \times \mathcal{I}_{\text{on}}. \end{aligned} \tag{7}$$

Similarly, we can formulate the price of the switched knockout puts with the exercise price $K$: The terminate condition at time $t = T$ is given by

$$V(S(T), T) = \max(K - S(T),\ 0). \tag{8}$$

The boundary conditions are, for the up-and-out put type,

$$\begin{aligned} &\lim_{\xi \to \infty} V(\xi, t) = 0, \qquad t \in \mathcal{I}_{\text{off}}, \\ &V(\xi, t) = 0, \qquad\qquad (\xi,\ t) \in [B(t),\ \infty) \times \mathcal{I}_{\text{on}}, \\ &\lim_{\xi \to 0} V(\xi, t) = Ke^{-r(T-t)}, \quad 0 \leq t \leq T, \end{aligned} \tag{9}$$

and for the down-and-out put type,

$$\begin{aligned} &\lim_{\xi \to \infty} V(\xi, t) = 0, \qquad 0 \leq t \leq T, \\ &\lim_{\xi \to 0} V(\xi, t) = Ke^{-r(T-t)}, \quad t \in \mathcal{I}_{\text{off}}, \\ &V(\xi, t) = 0, \qquad\qquad (\xi,\ t) \in [0,\ B(t)] \times \mathcal{I}_{\text{on}}. \end{aligned} \tag{10}$$

# 3. GENERAL PROPERTIES

## 3.1 PRELIMINARIES FOR NUMERICAL EXPERIMENTS

In general, it is quite difficult to obtain an analytical solution of the partial differential equation (2) together with such complex conditions as described in Section 2. The purpose of this paper is, however, no to obtain closed-form pricing formulas, but to examine general properties of the switched knockout options, in particular, the differences from the associated options without the nullified switch. Hence, we use a numerical method for the examination. In our numerical experiments, we used the Crank-Nicolson method for solving (2) with the terminate and boundary conditions. The Crank-Nicolson method has been known as a most accurate implicit finite-difference method;

see Courtadon [13] for details. Also, see Hull [14] and Wilmott *et al.* [15] for the general theory of finite-difference methods for option pricing.

To keep the original form of the knockout boundary as it is and to avoid the complication, we directly apply the Crank-Nicolson method to (2) without using any transformation of variables in the calculation. For convenience, we set the initial time to be $t = 0$ and the maturity to be $t = T = 1$. As a computational requirement, we restrict the state space of $S(t)$ for all $t$ in an interval $[0, S_{\max}]$ with $S_{\max} \equiv 1,000$ and divide this interval into 10,000 fragments with equal widths. Also, the time interval [0, 1] is divided into 500 fragments. For the option parameters, we use $K = 100$ and $r = 0.05$ in all cases, and $\sigma = 0.3$ if not clearly mentioned. For the knockout boundary function, we use a constant-valued boundary

$$B(t) = \begin{cases} B, & t \in \mathcal{I}_{\text{on}} \\ S_{\max}, & t \in \mathcal{I}_{\text{off}}, \end{cases} \tag{11}$$

where both $\mathcal{I}_{\text{on}}$ and $\mathcal{I}_{\text{off}}$ are compact sets in [0, 1] and $B \equiv 180$ in all cases.

## 3.2 THE UP-AND-OUT CALL TYPE

Figures 1 and 2 illustrate the curves of the up-and-out call price $V(S(0), 0)$ as a function of $S(0)$ for several knockout boundaries, where the intervals $\mathcal{I}_{\text{on}} = \emptyset$ (*i.e.* empty set) and $\mathcal{I}_{\text{on}} = [0, 1]$ are added for comparisons, which represent the vanilla and ordinary knockout options, respectively. Clearly, the prices of these extreme cases give upper and lower bounds for $V$ of the switched knockout options. In Figure 1, the knockout boundaries exist in latter parts of the trading interval, whereas in Figure 2 they exist in former parts. From these figures, we see that there are significant differences between these two cases: The option prices for the former-part cases are higher and more sensitive to the length of $\mathcal{I}_{\text{on}}$ than those for the latter-part cases. No doubt, this result is due to the assumption that the process $S(\cdot)$ follows a geometric Brownian motion with continuous sample paths. In actual markets, it is reasonable to place a knockout boundary at a latter part of the trading interval for hedging risk in future. In this sense, switched knockout options with latter-part boundaries can be attractive alternatives to the vanilla option. Another marked difference is the value of each option price when $S(0) \geq B = 180$. That is, the option prices for the latter-part cases have positive values, while those for the former-part cases are always 0.

To see the effects of volatility to option prices, we compute the prices of switched knockout options with $\sigma = 0.2, 0.3, 0.4$. Figures 3 and 4 illustrate the curves of $V(S(0), 0)$ as a function $S(0)$ when $\mathcal{I}_{\text{on}} = [0.5, 1]$ and $\mathcal{I}_{\text{on}} = [0, 0.5]$, respectively. For the vanilla option, it is well known that the price is monotonously increasing with $\sigma$, *i.e.*, $\partial V/\partial\sigma > 0$ for all $\sigma > 0$. However, we see from Figures 3 and 4 that this property does not hold for switched knockout options: Roughly speaking, for all $\sigma > 0$, $\partial V/\partial\sigma > 0$ when $S(0) << K$ and $\partial V/\partial\sigma < 0$ when $S(0) >> K$. This result indicates that a new scheme for risk hedging should be invented for switched knockout options. For more numerical results, see Hanada [16].

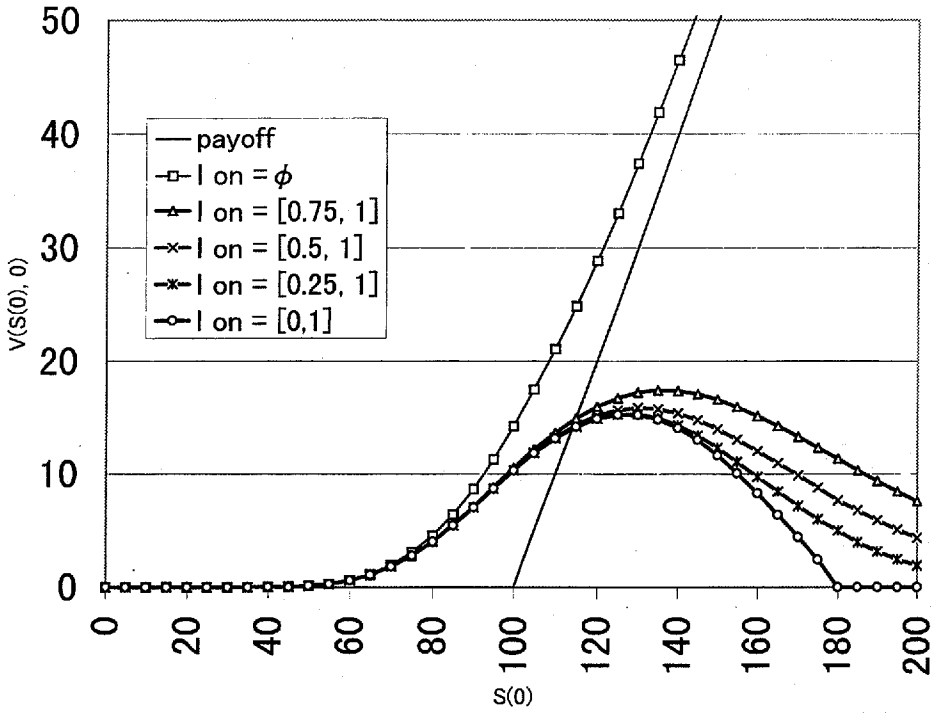


Figure 1: Prices of the Up-and-Out Calls: Latter-Part Cases

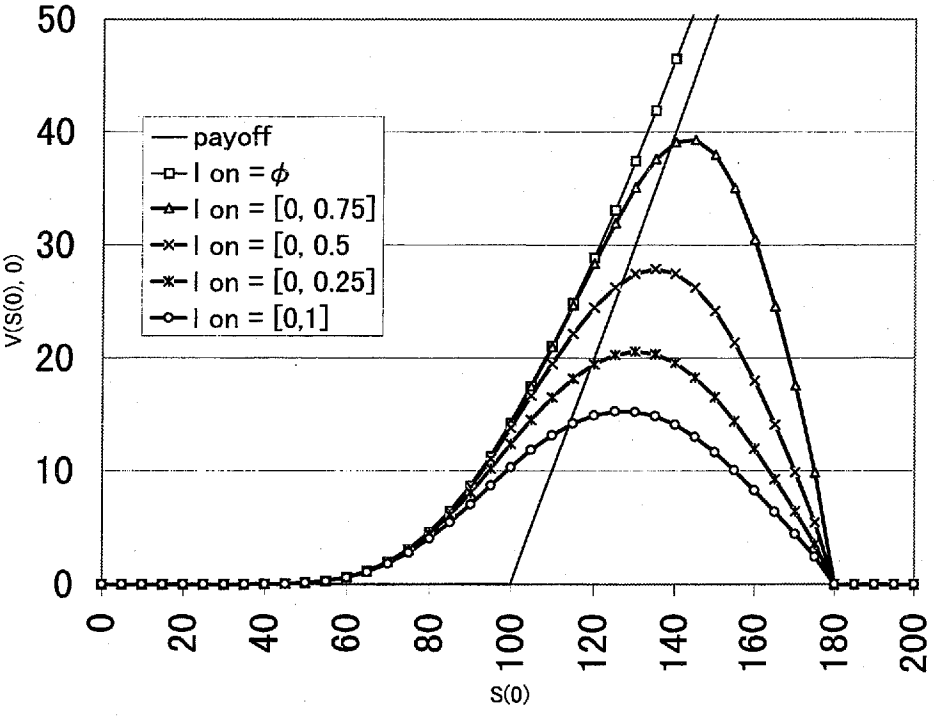


Figure 2: Prices of the Up-and-Out Calls: Former-Part Cases

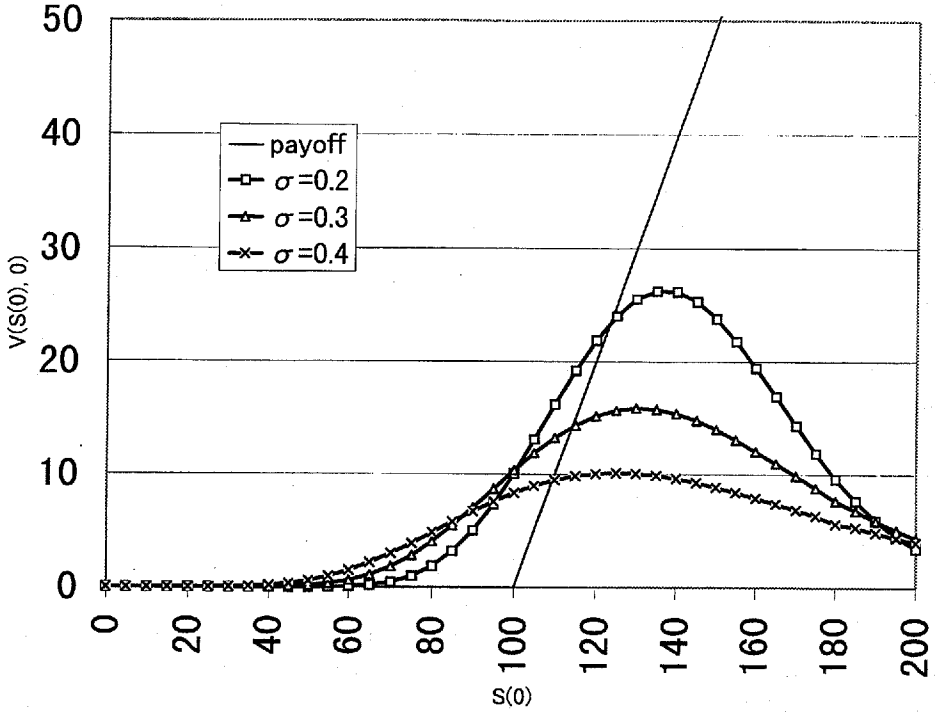


Figure 3: Prices of the Up-and-Out Calls: $\mathcal{I}_{\text{on}} = [0.5, 1]$

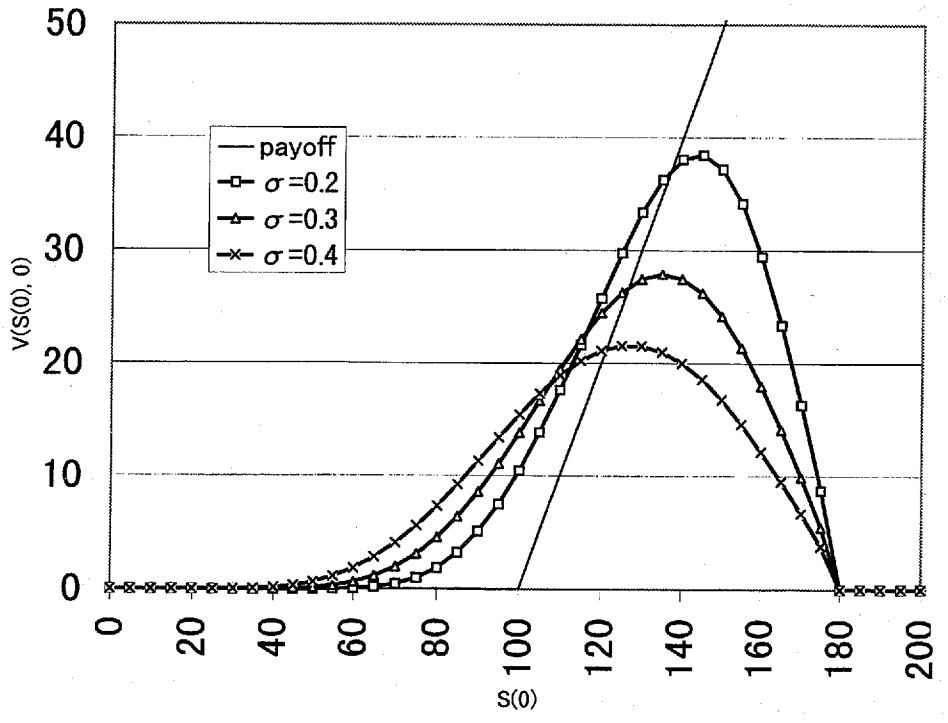


Figure 4: Prices of the Up-and-Out Calls: $\mathcal{I}_{\text{on}} = [0, 0.5]$

### 3.3 THE OTHER THREE TYPES

Some general properties of the other three types of switched knockout options can be shortly discussed by using similarity or symmetry: For the down-and-out call, it has properties similar to those for the associated vanila option. This result is, in some sense, reasonable because of the similarity on the boundary position. That is, both down-and-out and vanilla calls have the knockout boundaries in the direction that the option value is decreasing. Unlike the up-and-out call, the price for the down-and-out call is an increasing function of volatility, just as in the vanilla call.

Except for some trivial differences, general properties of the up-and-out put and the down-and-out call are almost symmetric about the line $S(0) = K$. This result clearly reflects the symmetry of the payoff lines for call and put options. We can observe a similar symmetric relation between the prices of down-and-out put and up-and-out call switched knockout options; see Hanada [16] for detailed numerical data.

## 4. CONCLUSION

In this paper, we have introduced the switched knockout option whose boundary feature is in the middle of the vanilla and ordinary knockout options. From extensive numerical experiments, we saw that the position of $\mathcal{I}_{\text{on}}$ in the trading interval significantly affects the option price, and that the sign of the hedge parameter $\partial V/\partial\sigma$ varies depending on $S(0)$. In addition, we saw that there are some similar and symmetric relations among the four types of switched knockout options.

A future direction of this research is to examine the cases that

- $\mathcal{I}_{\text{on}}$ contains many disjoint intervals,
- two knockout boundaries are located above and below $S(0)$,
- the knockout boundary is either a certain function of time $t$ and $S(t)$ or a random variable.

Another future direction is to develop an approximate pricing formula for the switched knockout option; see Hanada and Kimura [9] for a related research.

## REFERENCES


1. R. Merton, The theory of rational option pricing, *Bell Journal of Economics and Management Science*, **4**, 141–183 (1973).
2. M. Rubinstein and E. Reiner, Breaking down the barriers, *Risk*, **4**, 28–35 (1991).
3. D.R. Rich, The mathematical foundations of barrier option-pricing theory, *Advances in Futures and Options Research*, **7**, 267–311 (1994).
4. J.C. Cox and W. Rubinstein, *Options Markets*, Prentice Hall, Englewood Cliffs, N.J. (1985).
5. N. Kunitomo and M. Ikeda, Pricing options with curved boundaries, *Mathematical Finance*, **2**, 257–297 (1992).
6. H. Geman and M. Yor, Pricing and hedging double-barrier options: A probabilistic approach, *Mathematical Finance*, **6**, 365–378 (1996).
7. G.O. Roberts and C.F. Shortland, Pricing barrier options with time-dependent coefficients, *Mathematical Finance*, **7**, 83–93 (1997).

8. V. Linetsky, Step options, *Mathematical Finance*, **9**, 55–96 (1999).

9. K. Hanada and T. Kimura, Pricing knockout options with a general boundary, Proceedings of APORS'2000: The Fifth Conference of the Asia-Pacific Operations Research Societies within IFORS, Singapore (2000), to appear.

10. F. Black and M. Scholes, The pricing of options and corporate liabilities, *Journal of Political Economy*, **81**, 637–654 (1973).

11. J.M. Harrison and S.R. Pliska, Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications*, **11**, 215–260 (1981).

12. B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 5th ed., Springer, Berlin (1998).

13. G. Courtadon, A more accurate finite difference approximation for the valuation of options, *Journal of Financial and Quantitative Analysis*, **17**, 697–703 (1982).

14. J.C. Hull, *Options, Futures, and Other Derivative Securities*, 3rd ed., Prentice Hall, Englewood Cliffs, N.J. (1997).

15. P. Wilmott, J. Dewynne and S. Howison, *Option Pricing*, Oxford Financial Press, Oxford (1993).

16. K. Hanada, Numerical valuation of a switched knockout option, [in Japanese], *Economic Studies*, Hokkaido University, **49** (1999), to appear.

# OPTIMAL RESET NUMBER OF A MICROPROCESSOR SYSTEM WITH NETWORK PROCESSING

MITSUHIRO IMAIZUMI†, KAZUMI YASUI†† and TOSHIO NAKAGAWA††

†School of Business Management, Aichi Gakusen University,
1 Shiotori, Ooike-cho, Toyota, 471-8532, Japan
††Department of Industrial Engineering, Aichi Institute of Technology,
1247 Yachigusa, Yagusa-cho, Toyota, 470-0392, Japan
vzv03107@nifty.ne.jp / yasuilab@ie.aitech.ac.jp / nakagawa@ie.aitech.ac.jp

**Abstract**—As a computer network technology has remarkably developed, microcomputers which form a data terminal equipment (DTE) in a communication network have been used in many practical fields and the demand for improvement of their reliabilities has greatly increased. In fact, a microprocessor ($\mu P$) which is one of vital devices of a communication network often fails through some faults due to noise and changes in the environment and programming bugs. Therefore, it is necessary to take preventive measures for occurrences of such errors. This paper considers the maintenance problem for improving the reliability of a $\mu P$ system with network processing. After the system has made a stand-alone processing, it executes successively communication procedures of a network processing. When either $\mu P$ failures or application software errors in the system have occurred, a $\mu P$ is reset to the beginning of its initial state and restarts again. The reliability quantities such as the mean time to the success of a network processing and the expected reset number, using the theory of Markov renewal processes, are derived. An optimal reset number which minimizes the expected cost until a network processing is successful, is analytically discussed. A numerical example is finally given.

**Keywords**—Microprocessor, Network processing, Mean time, Expected cost, Reset number.

## 1. INTRODUCTION

As a computer network technology has remarkably developed, microcomputers which form a data terminal equipment (DTE) in a communication network have been used in many practical fields. Recently, a new communication network combining the information processing and communication plays an important role as the infrastructure in the information society. Therefore, the demand for improvement of reliabilities and functions for devices of a communication network has greatly increased [1].

In fact, a microprocessor ($\mu P$) which is one of vital devices of a communication network often fails through some faults due to noise and changes in the environment and programming bugs. Hence, it is necessary to take preventive measures for occurrences of such errors. Generally, when we consider the reliability of the system on an operational stage, we should regard the cause of error occurrences of a $\mu P$ as faults of software, such as mistakes of operational control and memory access, rather than faults of hardware. That is, when errors of a $\mu P$ have occurred, it is effective to recover the system by the operation of reset [2].

This paper considers the maintenance problem for improving the reliability of a $\mu P$ system with network processing: After the system has made a stand-alone processing, it executes successively

communication procedures of a network processing. When either $\mu P$ failures or application software errors in the system have occurred, a $\mu P$ is reset to the beginning of its initial state and restarts again. Most reliability evaluation models of a $\mu P$ system until now have assumed that both errors of a $\mu P$ and failures of the data transmission occur unlimitedly [3],[4],[5],[6]. This paper assumes that if the reset due to errors has occurred $N$ times intermittently, then a $\mu P$ interrupts its processing and restarts again from the beginning of its initial state after a constant time. That is, if the reset has occurred frequently, the system has latent faults, and takes preventive maintenances to check the environment and to eliminate errors.

We derive the reliability quantities such as the mean time and the expected reset number until a network processing is successful. Further, we regard the losses and times for the reset and the interruption of processing and for the maintenance to restart the system as expected costs, and discuss optimal policies which minimize them. A numerical example is finally given.

## 2. MODEL AND ANALYSIS

We pay attention to only a certain DTE which consists of a workstation or a personal computer and connects with some networks, and consider the problem for improving its reliability.

Suppose that errors of a $\mu P$ system occur according to an exponential distribution $F(t)$ with mean $1/\lambda$. If errors of a $\mu P$ have occurred, a $\mu P$ is reset to the beginning of its initial state and restarts again. It is assumed that any reset times are neglected.

**(1)** After a $\mu P$ begins to operate, it executes an initial processing immediately and a stand-alone processing.

**(2)** The times for an initial processing and a stand-alone processing have a general distribution $V(t)$ with mean $1/v$ and an exponential distribution $A(t)$ with $1/\alpha$, respectively.

**(3)** After a $\mu P$ completes a stand-alone processing, it begins to execute a network connection processing:

- **(a)** A connection processing needs the time according to a general distribution $B(t)$ with mean $1/\beta$, and fails with probability $\gamma$ $(0 \leq \gamma < 1)$.
- **(b)** If a connection processing has failed, a $\mu P$ executes the same processing again after a constant time $w$ where $W(t) \equiv 0$ for $t < w$ and 1 for $t \geq w$.

**(4)** After a connection processing has been successful, a $\mu P$ executes a network processing.

- **(c)** A network processing needs the time according to a general distribution $U(t)$ with mean $1/u$, and is successful with probability 1 if it has not failed.

**(5)** If the $N$-th reset has occurred since a $\mu$P begins to operate, once it interrupts the processing, and restarts again from the beginning after a constant time $\mu$, where $G(t) \equiv 0$ for $t < \mu$ and 1 for $t \geq \mu$.

Under the above assumptions, we define the following states of the system:

**State 0:** An initial processing begins.

**State 1:** A stand-alone processing begins.

**State 2:** A stand-alone processing is completed and a network connection processing begins.

**State 3:** A network connection processing succeeds and a network processing begins.

**State $F$:** A network processing is interrupted.

**State $S$:** A network processing succeeds.

The system states defined above form a Markov renewal process [7] where state $S$ is an absorbing state.

Let $Q_{i,j}(t)$ $(i = 0, 1, 2, 3; j = 0, 1, 2, 3, S)$ be one-step transition probabilities of a Markov renewal process. Then, mass functions $Q_{i,j}(t)$ from state $i$ at time 0 to state $j$ at time $t$ are:

$$Q_{0,0}(t) = \int_0^t \overline{V(t)} dF(t), \tag{1}$$

$$Q_{0,1}(t) = \int_0^t \overline{F(t)} dV(t), \tag{2}$$

$$Q_{1,0}(t) = \int_0^t \overline{A(t)} dF(t), \tag{3}$$

$$Q_{1,2}(t) = \int_0^t \overline{F(t)} dA(t), \tag{4}$$

$$Q_{2,0}(t) = \sum_{j=1}^{\infty} X^{(j-1)}(t) * \int_0^t [\overline{B(t)} + \gamma B(t) * \overline{W(t)}] dF(t), \tag{5}$$

$$Q_{2,3}(t) = \sum_{j=1}^{\infty} X^{(j-1)}(t) * [(1-\gamma) \int_0^t \overline{F(t)} dB(t)], \tag{6}$$

$$Q_{3,0}(t) = \int_0^t \overline{U(t)} dF(t), \tag{7}$$

$$Q_{3,S}(t) = \int_0^t \overline{F(t)} dU(t), \tag{8}$$

where

$$X(t) \equiv \gamma \int_0^t \overline{F(t)} dB(t) * \int_0^t \overline{F(t)} dW(t), \tag{9}$$

the asterisk mark denotes the Stieltjes convolution and $a^{(n)}(t)$ denotes the $n-$fold Stieltjes convolution of a distribution $a(t)$ with itself, i.e., $a^{(n)}(t) \equiv a^{(n-1)}(t) * a(t), a(t) * b(t) \equiv \int_0^t b(t-u) da(u)$.

We derive the mean time $\ell_S$ from the beginning of system operation until a network processing is successful. Let $H_{0,S}(t)$ be the first-passage time distribution from state 0 to state $S$. Then, we have

$$H_{0,S}(t) = \sum_{j=1}^{N} D^{(j-1)}(t) * Z(t), \tag{10}$$

where

$$D(t) \equiv Q_{0,0}(t) + Q_{0,1}(t) * Q_{1,0}(t) + Q_{0,1}(t) * Q_{1,2}(t) * Q_{2,0}(t)$$

$$+ Q_{0,1}(t) * Q_{1,2}(t) * Q_{2,3}(t) * Q_{3,0}(t), \tag{11}$$

$$Z(t) \equiv Q_{0,1}(t) * Q_{1,2}(t) * Q_{2,3}(t) * Q_{3,S}(t). \tag{12}$$

It is noted that $D(t)$ is the distribution function which a $\mu P$ is reset by occurrences of errors and $Z(t)$ is the distribution function which the system moves from state 0 to state $F$ directly without being reset. Further, the first-passage time distribution $H_{0,F}(t)$ from state 0 to state $F$ by a $\mu P$ the $N$-th reset is given by

$$H_{0,F}(t) \equiv D^{(N)}(t). \tag{13}$$

Therefore, the first-passage time distribution $L_S(t)$ until a network processing is successful is given by the following renewal equation:

$$L_S(t) = H_{0,S}(t) + H_{0,F}(t) * G(t) * L_S(t). \tag{14}$$

Let $\phi(s)$ be the Laplace-Stieltjes (LS) transform of any function $\Phi(t)$, i.e., $\phi(s) \equiv \int_0^\infty e^{-st} d\Phi(t)$. Taking the LS transforms on both sides of (14) and arranging them, we have

$$l_S(s) = \frac{h_{0,S}(s)}{1 - h_{0,F}(s)g(s)}. \tag{15}$$

Hence, the mean time $\ell_S$ is given by

$$\ell_S \equiv \int_0^\infty t dL_S(t) = \lim_{s \to 0} \{ - \frac{dl_S(s)}{ds} \} = - \frac{z'(0) + d'(0)}{1 - d(0)} + \frac{\mu d(0)^N}{1 - d(0)^N}. \tag{16}$$

where $\phi'(s)$ is the differential function of $\phi(s)$, i.e., $\phi'(s) \equiv d\phi(s)/ds$. From equation (16), $\ell_S$ is strictly decreasing in $N$ and is minimized when $N = \infty$.

Next, we derive the expected reset number $M_R$ from the start of system operation or the restart by the reset until a network processing is successful. Let $M_R(t)$ be the expected reset number until a network processing is successful in an interval $(0, t]$. Then, we have

$$M_R(t) = \sum_{j=1}^{N-1} j D^{(j)}(t) * Z(t). \tag{17}$$

Thus, the expected reset number is given by

$$M_R \equiv \lim_{t \to \infty} M_R(t) = \lim_{s \to 0} \sum_{j=1}^{N-1} j[d(s)]^j z(s) = \frac{d(0)}{1 - d(0)} [1 - N d(0)^{N-1} + (N-1) d(0)^N], \tag{18}$$

where it is noted that $z(0) = 1 - d(0)$.

Further, let $M_F(t)$ be the distribution of the expected interruption number of processing from the start of system operation until a network processing is successful. Then, we have the following renewal equation:

$$M_F(t) = H_{0,F}(t) * [1 + G(t) * M_F(t)]. \tag{19}$$

Similar to equation (18), the expected interruption number $M_F$ until a network processing is successful is given by

$$M_F = \frac{d(0)^N}{1 - d(0)^N}. \tag{20}$$

# 3. OPTIMAL POLICIES

We obtain two objective functions which are the total expected cost $C(N)$ and the expected cost $\hat{C}(N)$ per unit of time until a network processing is successful, and discuss optimal policies which minimize them, respectively.

## 3.1 POLICY 1

Let $c_1$ be the cost for the reset and $c_2$ be the cost for an interruption of processing. Then, we define the total expected cost $C(N)$ until a network processing is successful as the following equation:

$$C(N) \equiv c_1 M_R + c_2 M_F = c_1[\frac{D(1-D^N)}{1-D} - ND^N] + \frac{c_2 D^N}{1-D^N} \qquad (N = 1, 2, \cdots), \tag{21}$$

where $D \equiv d(0)$ which is the probability that a $\mu P$ is reset.

We seek an optimal number $N^*$ which minimizes $C(N)$. From the inequality $C(N+1) - C(N) \geq 0$, we have

$$N(1-D^N)(1-D^{N+1}) \geq \frac{c_2}{c_1}. \tag{22}$$

Denoting the left-hand side of (22) by $L(N)$, we have

$$L(1) = (1-D)(1-D^2), \tag{23}$$

$$L(\infty) = \infty. \tag{24}$$

Hence, $L(N)$ is strictly increasing in $N$ from $L(1)$ to $\infty$. Thus, we have the following optimal policy:

**(i)** If $L(1) < c_2/c_1$, then there exists a finite and unique minimum $N^*(> 1)$ which satisfies (22).

**(ii)** If $L(1) \geq c_2/c_1$, then $N^* = 1$ and the total expected cost is $C(1) = (c_2 D)/(1-D)$.

In this model, $c_1$ is the cost for the increase of system resources such as spaces of memory and times by the reset, and $c_2$ is for the increase of system resources by the preventive maintenance to eliminate the cause of errors. It could be generally estimated that $c_2$ is greater than $c_1$, i.e., $c_2 \geq c_1$. Thus, we have $L(1) < c_2/c_1$, and hence, $N^* > 1$. Further, it is easily shown that $N^*$ increases with $c_2/c_1$.

## 3.2 POLICY 2

In the policy 1, we have considered the total expected cost as an objective function. However, it would be more practical to introduce the measure of the time until a network processing is successful. Next, we consider an optimal policy which minimizes the expected cost per unit of time until a network processing is successful. That is, from equations (16) and (21), we define the expected cost $\hat{C}(N)$ per unit of time as the following equation:

$$\hat{C}(N) \equiv \frac{C(N)}{\ell_S} = \frac{c_1 \sum_{j=1}^{N-1} jD^j(1-D) - \frac{A}{\mu}c_2}{A + \frac{\mu D^N}{1-D^N}} + \frac{c_2}{\mu} \qquad (N = 1, 2, \cdots), \tag{25}$$

where

$$A \equiv -\frac{z'(0)+d'(0)}{1-D} > 0. \tag{26}$$

We seek an optimal number $N_1^*$ which minimizes $\hat{C}(N)$. From the inequality $\hat{C}(N+1)-\hat{C}(N) \geq 0$, we have

$$N(1-D^N)(1-D^{N+1}) + \frac{\mu}{A}[ND^N(1-D^{N+1}) + (1-D)\sum_{j=1}^{N-1} jD^j] \geq \frac{c_2}{c_1}. \tag{27}$$

Denoting the left-hand side of (27) by $L_1(N)$,

$$L_1(1) = (1-D^2)(1-D+\frac{\mu}{A}D), \tag{28}$$

$$L_1(\infty) = \infty. \tag{29}$$

Putting the second term on the bracket of the left-hand side of (27) by

$$L_2(N) \equiv ND^N(1-D^{N+1}) + (1-D)\sum_{j=1}^{N-1} jD^j, \tag{30}$$

we have

$$L_2(1) = (1-D^2)D, \tag{31}$$

$$L_2(\infty) = \frac{D}{1-D}, \tag{32}$$

$$L_2(N+1) - L_2(N) = D^{N+1}[1-D^{N+2}+ND^N(1-D^2)] > 0. \tag{33}$$

Hence, $L_2(N)$ is strictly increasing in $N$. Further, since $N(1-D^N)(1-D^{N+1})$ in (27) is also strictly increasing in $N$, $L_1(N)$ is strictly increasing in $N$ from $L_1(1)$ to $\infty$. Thus, we have the following optimal policy:

**(i)** If $L_1(1) < c_2/c_1$, then there exists a finite and unique minimum $N_1^*(>1)$ which satisfies (27).

**(ii)** If $L_1(1) \geq c_2/c_1$, then $N_1^* = 1$, and the resulting cost is

$$\hat{C}(1) = \frac{c_2 D}{A(1-D)+\mu D}. \tag{34}$$

Further, we compare the optimal policy 2 to the optimal policy 1. Since from equations (22) and (27),

$$L_1(N) - L(N) = \frac{\mu}{A}[ND^N(1-D^{N+1}) + (1-D)\sum_{j=1}^{N-1} jD^j] > 0 \qquad (N = 1, 2, \cdots), \tag{35}$$

and hence, $N^* \geq N_1^*$.

This means that when the number $N$ of reset is small, the mean time until a network processing is large, since $\ell_S$ strictly decreases in $N$. Thus, it would be better to adopt the policy 2 where $N$ is small when we consider only the cost of the system on the whole. On the other hand, if we want a processing time to be small, we should adopt the policy 1.

# 4. NUMERICAL EXAMPLE

We compute numerically the optimal number $N_1^*$ which minimizes $\hat{C}(N)$ for Policy 2. Suppose that the mean initial processing time $1/v$ of $\mu P$ is a unit of time and the mean time to error occurrences is $(1/\lambda)/(1/v) = 30 \sim 60$. Further, the mean stand-alone processing time is $(1/\alpha)/(1/v) = 5 \sim 20$, the mean network connection processing time is $(1/\beta)/(1/v) = 1$, the mean waiting time when a network connection processing fails is $w/(1/v) = 1 \sim 4$, the mean network processing time is $(1/u)/(1/v) = 10$, the mean maintenance time after an interruption of processing is $(1/\mu)/(1/v) = 10$, the probability that a network connection processing fails is $\gamma = 0.2, 0.4, 0.6$, and the cost $c_1$ for the reset is a unit of cost and the cost rate of an interruption of processing is $c_2/c_1 = 1 \sim 3$.

Table 1 gives the optimal reset number $N_1^*$ which minimizes the expected cost $\hat{C}(N)$. For example, when $(1/\lambda)/(1/v) = 60$, $wv = 2$, $\gamma = 0.2$, $(1/\alpha)/(1/v) = 10$ and $c_2/c_1 = 2$, the optimal number is $N_1^* = 3$.

Table 1: Optimal reset number $N_1^*$ to minimize $\hat{C}(N)$.

| $(1/\lambda)/(1/v)$ | $wv$ | $\gamma$ | $(1/\alpha)/(1/v) = 5$ | | | | | $(1/\alpha)/(1/v) = 10$ | | | | | $(1/\alpha)/(1/v) = 20$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $c_2/c_1$ | | | | | $c_2/c_1$ | | | | | $c_2/c_1$ | | | | |
| | | | 1 | 1.5 | 2 | 2.5 | 3 | 1 | 1.5 | 2 | 2.5 | 3 | 1 | 1.5 | 2 | 2.5 | 3 |
| 30 | 1 | 0.2 | 2 | 2 | 3 | 3 | 4 | 2 | 3 | 3 | 3 | 4 | 2 | 3 | 3 | 4 | 4 |
| | | 0.4 | 2 | 2 | 3 | 3 | 4 | 2 | 3 | 3 | 4 | 4 | 2 | 3 | 3 | 4 | 4 |
| | | 0.6 | 2 | 3 | 3 | 3 | 4 | 2 | 3 | 3 | 4 | 4 | 3 | 3 | 4 | 4 | 4 |
| | 2 | 0.2 | 2 | 2 | 3 | 3 | 4 | 2 | 3 | 3 | 3 | 4 | 2 | 3 | 3 | 4 | 4 |
| | | 0.4 | 2 | 3 | 3 | 3 | 4 | 2 | 3 | 3 | 4 | 4 | 3 | 3 | 3 | 4 | 4 |
| | | 0.6 | 2 | 3 | 3 | 3 | 4 | 2 | 3 | 3 | 4 | 4 | 3 | 3 | 4 | 4 | 4 |
| | 4 | 0.2 | 2 | 2 | 3 | 3 | 4 | 2 | 3 | 3 | 4 | 4 | 2 | 3 | 3 | 4 | 4 |
| | | 0.4 | 2 | 3 | 3 | 3 | 4 | 2 | 3 | 3 | 4 | 4 | 3 | 3 | 4 | 4 | 4 |
| | | 0.6 | 2 | 3 | 3 | 4 | 4 | 2 | 3 | 3 | 4 | 4 | 3 | 3 | 4 | 4 | 5 |
| 60 | 1 | 0.2 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 |
| | | 0.4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 |
| | | 0.6 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 |
| | 2 | 0.2 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 |
| | | 0.4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 |
| | | 0.6 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 |
| | 4 | 0.2 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 |
| | | 0.4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 |
| | | 0.6 | 2 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 2 | 3 | 3 | 3 | 4 |

This shows that the optimal number $N_1^*$ decreases with $(1/\lambda)/(1/v)$, however, increases with $wv$, $\gamma$, $(1/\alpha)/(1/v)$ and $c_2/c_1$. This can be interpreted that when the cost for an interruption of processing is large, $N_1^*$ increases with $c_2/c_1$, and so, the processing should not be excessively interrupted. That is, we should keep on executing the processing as long as possible by the reset. Table 1 also shows that $N_1^*$ depends on each parameter when $(1/\lambda)/(1/v)$ is small, i.e., when errors of a $\mu P$ occur frequently, however, $N_1^*$ depends little on $wv$, $\gamma$ and $(1/\alpha)/(1/v)$ when $(1/\lambda)/(1/v) \geq 60$, and $N_1^*$ is almost determined by $c_2/c_1$.

# 5. CONCLUSIONS

We have investigated the problem for improving the reliability of a $\mu P$ system with network processing, and have derived the mean time and mean reset numbers until a network processing is successful. Further, we have discussed the optimal reset numbers which minimize the total expected cost and the expected cost per unit of time.

It has been shown from the mathematical analysis that the optimal reset number which minimizes the total cost is larger than that which minimizes the expected cost per unit of time. It has been also shown from the numerical example that the optimal reset number which minimizes the expected cost decreases with the mean time to error occurrences of a $\mu P$, however, increases with the mean stand-alone processing time, the probability that a network processing fails and the cost for an interruption of processing. Further, when the mean time to error occurrences is large, the optimal reset number depends little on each parameter and is almost determined by the cost for an interruption of processing.

It would be very important to evaluate the reliability of a $\mu P$ system with network processing. Further studies for such subjects would be expected.

# REFERENCES


1. K. Ono, *Computer Communication*, Ohmsha (1996) (in Japanese).
2. T. Nanya, *Fault-Tolerant Computer*, Ohmsha (1991) (in Japanese).
3. K. Yasui, T. Nakagawa and M. Motoori, A Two-Stage Error Control Policy for a Data Transmission System with Intermittent Faults, *The Transactions of the Institute of Electronics, Information and Communication Engineers of Japan*, Vol.J75-A, No.5, pp. 944-948, May (1992) (in Japanese).
4. H. Sandoh, T. Nakagawa and S. Koike, A Bayesian Approach to an Optimal ARQ Number in Data Transmission, *The Transactions of the Institute of Electronics, Information and Communication Engineers of Japan*, Vol.J75-A, No.7, pp. 1198-1192, July (1992) (in Japanese).
5. T. Nakagawa, K. Yasui and H. Sandoh, An Optimal ARQ Policy for a Data Transmission System with Intermittent Faults, *The Transactions of the Institute of Electronics, Information and Communication Engineers of Japan*, Vol.J76-A, No.8, pp. 1201-1206, August (1993) (in Japanese).
6. K. Yasui, T. Nakagawa and H. Sandoh, An ARQ Policy for a Data Transmission System with Three Types of Error Probabilities, *The Transactions of the Institute of Electronics, Information and Communication Engineers of Japan*, Vol.J78-A, No.7, pp. 824-830, July (1995) (in Japanese).
7. S. Osaki, *Applied Stochastic System Modeling*, Springer-Verlag Berlin (1992).

# OPTIMAL SELF-DIAGNOSIS POLICY FOR DUAL REDUNDANT FADEC OF GAS TURBINE ENGINES

KODO ITO[1] and TOSHIO NAKAGAWA[2]

[1] Nagoya Guidance and Propulsion Systems Works,
Mitsubishi Heavy Industries, LTD.,
1200 Higashi-Tanaka, Komaki-shi, Aichi 485-8561, Japan
[2]Department of Industrial Engineering, Aichi Institute of Technology
1247 Yachigusa, Yakusa-cho, Toyota 470-0392, Japan

**Abstract**—FADEC(Full-Authority Digital Engine Control) was developed for aircraft gas turbine engine controllers and now has been widely adopted to industrial ones because of its high performance. Although aircraft FADECs are so expensive because they operate in severe environment, industrial FADECs should be inexpensive considering the severe cost competition in the market. Recently, the recent progress of electronics produces high performance and low price PLCs(Programmable Logic Controllers). Although they were originally developed as substitutive relay logic sequencers, they have been utilized as multi-purpose numerical controllers now. When we adopt them, we shall develop inexpensive gas turbine FADECs. However, the PLC makers do not assure to use them as gas turbine engine controllers. So engine makers should consider adequate measures and assure their reliabilities when they utilize them as FADECs. This paper considers the self-diagnosis policy for dual redundant FADECs. The self-diagnosis is performed at every n-th control calculation cycles. Introducing expected cost per unit time, an optimal n* which minimizes it is considered.

**Keywords**—Advice to authors, Important notice

## 1. INTRODUCTION

The original idea of gas turbine engine was represented by Barber in England at 1791, and the engine was firstly realized in 20-th century. After that, they had advanced greatly during World War II. Today, gas turbine engines have been widely utilized as main engines of airplanes, high performance mechanical pumps, emergency generator and cogeneration systems because they can generate high power compare to their size, their start time is very short and no coolant water is necessary for operation.

Gas turbine engines are mainly constituted with three parts, *i.e.*, compressor, combustor and turbine. The engine control is performed by governing the fuel flow to engine. When gas turbine engines operate, surge, stool and over-temperature of exhaust gas should be paid attention because these phenomena cause serious damage for engine. To prevent such dangerous phenomena, the turbine speed, inlet temperature and pressure, and exhaust gas temperature of gas turbine engine are monitored and engine controller should determine appropriate fuel flow considering these data.

The gas turbine engine operates in serious environment and hydro mechanical controller (HMC) is adopted to engine controller for long period because of its high reliability, durability and excellent responsibility. However, the performance of gas turbine engines have advanced and customers need to decline the operation cost. So HMC could not follow these advanced demands and the engine controller has been electrified. The first electric engine controller which was a support unit of HMS, was adopted for J47-17 turbo jet engine of F86D fighter at the late 1940-th. The change of

device,*i.e.*, from vacuum tube to transistor and transistor to IC, has changed the roll of electric engine controller from the assistant of HMS to full authority controller because of the reliability growth. In 1960-th, the analogue full authority controller could not follow the accuracy demand of engines, and full authority digital engine controller (FADEC) was developed [1, 2, 3].

FADEC is an electric engine controller which can perform the complicated signal processes of digital engine data. Aircraft FADECs, which are expected high mission reliability and are needed to decrease weight, hardware complication and electric consumption, adopt generally a duplicated system [4, 5, 6].

Industrial gas turbine engines have been advanced absorbing key technologies which were established for aircraft gas turbine engines. FADECs which were originally developed for aircrafts, have also been adopted for industrial gas turbine engines. Comparing between general industrial gas turbine FADECs and aircraft FADECs, the following differences are recognized:

1) Aircraft gas turbine FADECs have to perform high speed data processing because the rapid response for aircraft body movement is necessary and inlet pressure and temperature change greatly depending on height. On the other hand, industrial gas turbine FADECs not need such high performance comparing to aircraft ones because they operate at steady speed on ground.

2) Aircraft gas turbine FADECs have to be reliable and fault tolerable, and so, they adopt a duplicate system because their malfunction in operation may cause serious damages for aircrafts and crews. Industrial gas turbine FADECs also have to be reliable and fault tolerable, and still be low cost because they have to be competitive in market.

Depending on the advance of microelectronics, small, high performance, low cost programmable logic controllers (PLC) have distributed in market. Their origins were relay sequencers and they are still utilized as sequencers of industrial automatic systems. Applying numerical calculation ability of microprocessors, these PLCs occupy the analogue-digital and digital-analogue transformer and can perform numerical control. When we use such PLCs, very high cost performance FADEC system can be realized. However, these PLCs are developed as general industrial controllers and PLC makers might not permit them for applying high temperature and pressurized hot gas controllers. Then, gas turbine makers which apply these PLC as FADEC, have to design some protective mechanism and have to assure high reliability of FADEC.

In this paper, we consider a self-diagnosis policy for dual redundant gas turbine engine FADECs.

## 2. ANALYSIS

A dual redundant system is commonly employed for aircraft FADECs. We consider the following dual redundant FADEC:

(a) A system is constituted with two perfectly endependent channels *i.e.*, they have their own engine sensors, fuel control valves, watch dog timers and power sources.

(b) A data communication line called CCDL (Cross Channel Data Link) connects two channels. Each channel exchanges its engine sensor data and calculation results with another channel, and can diagnosis each other (cross-diagnosis). Furthermore, each channel performs self-diagnosis by its watch dog timer. We call the cross-diagnosis between two channels and self-diagnosis of each channel as the self-diagnosis of FADEC.

(c) Although two channels perform the same control calculation at same time, only one channel can conquer a whole system at one time. Initially, one channel has a priority to control a

whole engine (active condition) and another channel is in standby condition. When the active channel fails, the channel changes from active condition to standby one, and another channel changes from standby condition to active one. When two channel fail, a whole engine system stops.

We introduce the expected cost per unit time and derive an optimal diagnosis policy which minimizes it. Consider the following control and self-diagnosis policy for dual redundant FADEC :

(I) The reliability of channel $i$ at time $t$ is $\bar{F}_i(t)$, where $i = 1, 2$.

(II) The control calculation of each channel is performed at time interval $T_0$ and the self-diagnosis and cross-diagnosis are performed synchronously between two channels at every $n$-th calculation. The coverage of these diagnoses is 100%.

(III) Initially, channel 1 is in active condition and channel 2 is in standby condition. When channel 1 fails, channel 1 changes to standby condition and channel 2 changes to active condition with no failure. We assume these elapsed time for changing are negligible.

(IV) When $n$ decreases, the diagnosis calculation per unit time increases and it degrades the quality of control. We assume that the degradation of control is represented as $c_1/(n + T_1)$, where $c_1$ is constant and $T_1$ is the percentage of diagnosis time divided by $T_0$.

(V) When $n$ increases, the time interval from occurrence of failure to its detection is prolonged and it causes the damage of gas turbine engine because of the extraordinary fuel control signal. The damage of engine is represented as $c_2(nT_0 - t)$, where $t$ is the time that failure occurs and $c_2$ is the system loss per unit time.

When channel $i$ fails at time $t_i$, the following two expected time intervals from occurrence of failure to its detection depending on the timing of $t_i$, are considered:

When $t_2 \leq t_1 < t_m$ or $t_{m-1} < t_1 < t_2 \leq t_m$, the expected time interval is

$$\sum_{m=1}^{\infty} F_2(t_m) \int_{t_{m-1}}^{t_m} (t_m - t_1) dF_1(t_1)\,, \tag{1}$$

where $t_m = mnT_0\,(m = 1, 2, 3 \cdots)$.

When $t_1 \leq t_{m-1} < t_2 \leq t_m$, the expected time interval is

$$\sum_{m=2}^{\infty} \sum_{k=1}^{m-1} \int_{t_{k-1}}^{t_k} (t_k - t_1 + t_m - t_2) dF_1(t_1) \int_{t_{m-1}}^{t_m} dF_2(t_2)\,. \tag{2}$$

The total expected time interval is the summation of equations (1) and (2), *i.e.*,

$$\sum_{m=1}^{\infty} F_2(t_m) \int_{t_{m-1}}^{t_m} (t_m - t_1) dF_1(t_1)$$

$$+ \sum_{m=2}^{\infty} \sum_{k=1}^{m-1} \int_{t_{k-1}}^{t_k} (t_k - t_1 + t_m - t_2) dF_1(t_1) \int_{t_{m-1}}^{t_m} dF_2(t_2)$$

$$= nT_0 \sum_{m=0}^{\infty} \bar{F}_1(mnT_0) - \frac{1}{\lambda_1}$$

$$+ \sum_{m=1}^{\infty} F_1(mnT_0) \int_0^{nT_0} [F_2(t + mnT_0) - F_2(mnT_0)] dt\,, \tag{3}$$

where $1/\lambda_1 \equiv \int_0^\infty \bar{F}_1(t)dt$, and $F_1(0) = 0$.

Thus, the expected cost is

$$C(n) = \frac{c_1}{n+T_1} + c_2 \left\{ nT_0 \sum_{m=0}^{\infty} \bar{F}_1(mnT_0) - \frac{1}{\lambda_1} + \sum_{m=1}^{\infty} F_1(mnT_0) \int_0^{nT_0} [F_2(t+mnT_0) - F_2(mnT_0)]dt \right\}. \quad (4)$$

Assuming $F_i(t) = 1 - exp(-\lambda_i t)$ $(i = 1, 2)$, equation(4) is rewritten as

$$C(n) = \frac{c_1}{n+T_1} + c_2 \left\{ nT_0 \left[ \frac{1}{1-e^{-\lambda_1 nT_0}} + \frac{1}{1-e^{-\lambda_2 nT_0}} - \frac{1}{1-e^{-(\lambda_1+\lambda_2)nT_0}} \right] + \frac{1-e^{-\lambda_2 nT_0}}{\lambda_2(1-e^{-(\lambda_1+\lambda_2)nT_0})} - \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right\}. \quad (5)$$

We easily find that

$$C(0) = \frac{c_1}{T_1}, \quad (6)$$

$$C(\infty) = \infty. \quad (7)$$

Therefore, there exist a finite $n^*$ $(< \infty)$ which minimizes $C(n)$.

## 3. NUMERICAL EXAMPLE

It is convenient to introduce $x$ defined as $x = nT_0$ to calculate optimal $n^*$. Using $x$, equation (5) is rewritten as

$$C(x) = \frac{c_1 T_0}{x+T_1T_0} + c_2 \left\{ x \left[ \frac{1}{1-e^{-\lambda_1 x}} + \frac{1}{1-e^{-\lambda_2 x}} - \frac{1}{1-e^{-(\lambda_1+\lambda_2)x}} \right] + \frac{1-e^{-\lambda_2 x}}{\lambda_2(1-e^{-(\lambda_1+\lambda_2)x})} - \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right\}. \quad (8)$$

We obtain the derivative of $C(x)$ as

$$\frac{dC(x)}{dx} = -\frac{c_1 T_0}{(x+T_1T_0)^2} + c_2 \left\{ \frac{1}{1-e^{-\lambda_1 x}} + \frac{1}{1-e^{-\lambda_2 x}} - \frac{1}{1-e^{-(\lambda_1+\lambda_2)x}} \cdot -x \left[ \frac{\lambda_1 e^{-\lambda_1 x}}{(1-e^{-\lambda_1 x})^2} + \frac{\lambda_2 e^{-\lambda_2 x}}{(1-e^{-\lambda_2 x})^2} - \frac{(\lambda_1+\lambda_2)e^{-(\lambda_1+\lambda_2)x}}{(1-e^{-(\lambda_1+\lambda_2)x})^2} \right] + \frac{e^{-\lambda_2 x}}{1-e^{-(\lambda_1+\lambda_2)x}} - \frac{(\lambda_1+\lambda_2)(1-e^{-\lambda_2 x})e^{-(\lambda_1+\lambda_2)x}}{\lambda_2(1-e^{-(\lambda_1+\lambda_2)x})^2} \right\}. \quad (9)$$

To obtain a minimum $C(x)$, we search $x^*$ numerically which satisfies $dC(x)/dx = 0$.

Table 1 gives the optimal self-diagnosis time interval $x^*$ and $n^*$, and the expected cost $C(n^*)$ for $\lambda_1 = 1.0 \times 10^{-7}, 2.0 \times 10^{-7}, 4.0 \times 10^{-7}$ per hour, $c_1 = 1, 5, 10$ and $T_1 = 0.1, 0.5, 1.0$ when $\lambda_2 = 1.0 \times 10^{-7}$ per hour, $c_2 = 10$ and $T_0 = 10^{-2}(10msec.)$. Optimal $n^*$s are integers and are also denoted parenthetically for comparison. When $c_1/c_2$ or $1/T_1$ increases, $x^*$ or $n^*$, and $C(n^*)$ increase. When $\lambda_1/\lambda_2$ increases, $x^*$ or $n^*$ decreases and $C(n^*)$ increases.

**Table 1.** Optimal self-diagnosis interval $x^*$ and $n^*$, which minimize expected cost $C(n^*)$.

| $\lambda_1/\lambda_2$ | $c_1/c_2$ | $T_1$ | $x^*$ | $n^*$ | $C_1(n^*)$ |
|---|---|---|---|---|---|
| 1.0 | 0.1 | 0.1 | 0.036 | 4 (3.6) | 0.54 |
| 2.0 | 0.1 | 0.1 | 0.034 | 3 (3.4) | 0.57 |
| 4.0 | 0.1 | 0.1 | 0.032 | 3 (3.2) | 0.59 |
| 1.0 | 0.5 | 0.1 | 0.081 | 8 (8.1) | 1.22 |
| 1.0 | 1.0 | 0.1 | 0.114 | 11(11.4) | 1.72 |
| 1.0 | 0.1 | 0.5 | 0.032 | 3 (3.2) | 0.51 |
| 1.0 | 0.1 | 1.0 | 0.026 | 3 (2.6) | 0.47 |

# 4. CONCLUSION

We have considered an optimal self-diagnosis policy for dual redundancy FADEC: A FADEC performs the control calculation at time interval $T_0$ and self-diagnosis is performed at every $n$-th calculation. The expected cost is derived and there exists an optimal $n^*$ which minimizes it, when reliability functions of two channels are exponential distributions. Numerical examples have shown optimal $n^*$.

# REFERENCES


1. Robinson,K.(1987) Digital Controls for Gas Turbine Engines (Presented at the Gas Turbine Conference and Exhibition, Anaheim, California - May31-June4).

2. Kendell,R.(1981) Full-Authority Digital Electronic Controls for Civil Aircraft Engines (Presented at Gas Turbine Conference and Products Show, Houston, Texas - March9-12).

3. Scoles,R.J.(1986) FADEC - Every Jet Engine Should Have One (Presented at Aerospace Technology Conference and Exposition, Long Beach, California - October 13-16).

4. Eccles,E.S., Simons,E.D. and Evans,J.F.O.(1980) Redundancy Concepts in Full Authority Electronic Engine Control - Particularly Dual Redundancy (Presented at AGARD Conference).

5. Davies,W.J., Hoelzer,C.A. and Vizzini,R.W.(1983) F-14 Aircraft and Propulsion Control Integration Evaluation, Journal of Engineering for Power, Vol.105, pp.663-668.

6. Cahill,M.J. and Underwood,F.N.(1987) Development of Digital Engine Control System for the Harrier II, (Presented at AIAA/SAE/ASME/ASEE 23rd Joint Propulsion Conference, San Diego, California - June 29 - July 2).

# OPTIMAL LIFE INSURANCE AND PORTFOLIO CHOICE IN A LIFE CYCLE

HIDEKI IWAKI[1], MASAAKI KIJIMA[2] and KATSUYA KOMORIBAYASHI[3]

[1]Graduate School of Systems Management, The University of Tsukuba,
3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
[2]Faculty of Economics, Tokyo Metropolitan University
1-1 Minami-Ohsawa, Hachioji-shi, Tokyo 192-0397, Japan
[3]IBJ-DL Financial Technology Co.,Ltd.
1-5-1 Otemachi, Chiyoda-ku, Tokyo 100-0004, Japan

iwaki@gssm.otsuka.tsukuba.ac.jp / kijima@bcomp.metro-u.ac.jp /
komoribayashi@fintec.co.jp

**Abstract**—This paper considers an optimal life insurance for a householder subject to mortality risk. The household receives a wage income continuously, which is terminated by the householder's death. In order to protect the sudden loss, the household buys a life insurance from which they can receive some amount of insurance at the householder's death. Also, the household can invest their wealth into a financial market. The problem is to determine an optimal insurance and investment in order to maximize the expected total, discounted utility from consumption and terminal wealth. It is shown that an explicit solution is obtained for some special case.

**Keywords**—Life Insurance, Life Cycle, Investment/Consumption Model, Martingale.

## 1. INTRODUCTION

In this paper, we consider an asset allocation problem of a household. In the literature of such problems, asset classes are limited to riskless asset (bank accounts), and risky assets (stocks). In this paper, we extend the asset allocation problem in such a way to include life insurance contracts. By including them, the problem becomes not only to obtain an optimal optimal consumption and investment of the household, but also to decide how much amount of the life insurance should be invested to prepare for mortality risk of its householder.

Asset allocation problems are traced back to Merton [5] in which he derived optimal consumption and portfolio selection rules by assuming that an agent has specific utility functions in a continuous-time model. Subsequently, Merton [6] generalized his model for the case of general utility functions. In these papers, an agent decides the amount of consumption for goods or services, and the amount for investment into financial assets at each time so as to maximize their utility through their life time. In a relatively recent research, a life time model of Bodie, Merton and Samuelson [1] considered a human capital in order to add more reality to existing models. The human capital of an agent represents the present value of the total wage income which he/she will obtain in the future. By including the human capital in their model, they explicitly derived the relation between age and optimal investment strategies for the agent.

This paper considers an optimal life insurance for a household. The household lives through consumption by a wage income of the householder and a capital gain of investment into financial assets. However, if the householder dies, the wage income will terminate. The household may then want to buy a life insurance to prepare for risk of the householder's death. The household decides

an optimal insurance as well as optimal consumption and investment amount into financial assets to maximize the expected total, discounted utility from consumption and terminal wealth.

This paper is organized as follows. In the next section, we formulate our model by assuming that utility functions of the household are given, and derive an optimal insurance, consumption and investment amounts. In Section 3, we consider the special cases of exponential and power utility functions, and derive an optimal insurance, consumption and investment amounts explicitly.

## 2. THE MODEL

We consider a household which consumes a wage income of its householder to maximize the expected total, discounted utility from consumption and terminal wealth. The income of the household is only the wage of the householder, and if he/she dies, then the income terminates. Therefore, the household may be willing to buy a life insurance to protect the sudden loss from mortality risk of the householder. On the other hands, the household may want to invest its wealth into financial assets. Let the current time be zero, and assume that the householder's income terminates at time $T > 0$, i.e., his/her retirement is time $T$. The problem for the household is then to maximize the expected total, discounted utility from consumption over time 0 to $T$, and from terminal wealth at time $T$. The terminal wealth will be used for their lives after retirement or a bequest to their descendants.

We assume that our economy consists of a financial market, which is frictionless and perfect, and that every trade occurs continuously at time $t \in \mathcal{T}$, $\mathcal{T} \equiv [0, T]$. In order to make the model tractable, we also assume that the resolution of uncertainty of the economy is described by evolutions of a standard Brownian motion $Z = \{Z(t); t \in \mathcal{T}\}$ and a Poisson process $N = \{N(t); t \in \mathcal{T}\}$ with an intensity process $\lambda = \{\lambda(t); t \in \mathcal{T}\}$ defined on a given probability space $(\Omega, \mathcal{F}, P)$, where $Z$ is assumed to be independent of $N$. Here, without loss of generality, we set $Z(0) = 0$ and $N(0) = 0$. Let $\mathbb{F} \equiv \{\mathcal{F}_t; t \in \mathcal{T}\}$ be the $P$-augmentation of filtration with

$$\mathcal{F}_t = \sigma\{(Z(s), N(s)); 0 \leq s \leq t\}, \qquad \forall\, t \in \mathcal{T}.$$

The intensity process $\lambda$ is assumed to be positive, Markov and predictable with respect to $\mathbb{F}$, and satisfies

$$\int_0^T |\lambda(t)| dt < \infty \qquad \text{a.s.}$$

Hereafter, equalities and inequalities for random variables hold in the sense of a.s. (almost surely); however, we omit the notation a.s. for the sake of notational simplicity. The conditional expectation operator given $\mathcal{F}_t$ is denoted by $E_t$ with $E = E_0$.

The financial assets into which the household can invest consist of a risk-free asset and a risky asset. Let $P_0(t)$ and $P_1(t)$ be the time $t \in \mathcal{T}$ prices of the risk-free asset and the risky asset, respectively. We assume that the price processes, $P_0(t)$ and $P_1(t)$, are defined by the following stochastic differential equations (SDEs), respectively.

$$P_0(0) = p_0, \qquad \frac{dP_0(t)}{P_0(t)} = r(t)dt, \qquad t \geq 0, \tag{1}$$

and

$$P_1(0) = p_1, \qquad \frac{dP_1(t)}{P_1(t)} = \mu(t)dt + \sigma(t)dZ(t), \qquad t \geq 0, \tag{2}$$

where $p_0$ and $p_1$ are positive constants, and where $r(t)$, $\mu(t)$ and $\sigma(t)$ are progressively measurable processes with respect to $\mathbb{F}$ that satisfy

$$\int_0^T |r(t) + \mu(t) + \sigma^2(t)|dt < \infty.$$

Let $\tau$ denote the time of the householder's death, and assume that it is generated by the first passage time of the Poisson process $N$ to state 1, i.e.

$$\tau = \inf\{t \geq 0; N(t) = 1\}.$$

A life insurance that the household considers to buy is as follows. An insurance company pays the insurance amount $\theta(t)$ if the householder's death occurs at time $t$ before the terminal epoch $T$, and nothing if it occurs after $T$. It is noted that, in many situations, the insurance amount is set to be constant; however, as we shall show later, it must be a stochastic process, in general, to attain the optimal plan. Of course, in order to receive the insurance amount, the household must pay a premium $\bar{p}$ to the insurance company at time 0. We assume that the *insurance process* $\theta = \{\theta(t); t \in \mathcal{T}\}$ is bounded and adapted to $\mathbb{F}$. It is also assumed that the *income process* $y = \{y(t); t \in \mathcal{T}\}$ and the *consumption process* $c = \{c(t); t \in \mathcal{T}\}$ are bounded and adapted to $\mathbb{F}$.

Let $w(t)$ be the investment amount into the risky asset at time $t$. We refer to $w = \{w(t); t \in \mathcal{T}\}$ as a *portfolio process*. Given a portfolio process $w$, a consumption process $c$, an insurance process $\theta$, and an *income process* $y$, the *wealth process* $W = \{W(t); t \in \mathcal{T}\}$ is defined by $W(0) = W_0 - \bar{p}$ and

$$\begin{aligned} dW(t) &= \left(y(t)1_{\{N(t-)=0\}} - c(t)\right) dt + \theta(t)1_{\{N(t-)=0\}}dN(t) \\ &\quad + w(t)\left[\mu(t)dt + \sigma(t)dZ(t)\right] + \left(W(t) - w(t)\right) r(t)dt \\ &= \left(y(t)1_{\{N(t-)=0\}} - c(t)\right) dt + \theta(t)1_{\{N(t-)=0\}}dN(t) \\ &\quad + r(t)W(t)dt + w(t)\left[(\mu(t) - r(t))dt + \sigma dZ(t)\right], \qquad t \in \mathcal{T}, \end{aligned} \tag{3}$$

where $W_0$ is a given initial wealth which is assumed to be a positive constant, and where $1_{\{\cdot\}}$ denotes the indicator function.

Let $\phi(t)$ be the state price density at time $t$ which satisfies $\phi(0) = 1$, $0 < \phi(t) < \infty$, and for each $t \in \mathcal{T}$ and any $s > t$, $s \in \mathcal{T}$,

$$E_t[\phi(s)P_j(s)] = \phi(t)P_j(t), \qquad j = 0, 1. \tag{4}$$

The insurance premium is then given by

$$\bar{p} = E\left[\int_0^T \phi(t)\theta(t)1_{\{N(t-)=0\}}dN(t)\right], \tag{5}$$

since otherwise there is an arbitrage opportunity.

Consider another risky security whose price process $P_2(t)$ is defined by $P_2(0) = p_2$ and

$$dP_2(t) = P_2(t)\left(\frac{1}{p_2} - 1\right) dN(t), \qquad t \geq 0,$$

where $p_2$ is a positive constant. If we assume that the security is traded in the financial market, then it is not difficult to show that the state price density is represented as

$$\phi(t) = \exp\left\{-\int_0^t \xi(s)dZ(s) - \frac{1}{2}\int_0^t \xi^2(s)ds - \int_0^t r(s)ds + \int_0^t \ln\left(\frac{\psi(s)}{\lambda(s)}\right) dN(s) - \int_0^t (\psi(s) - \lambda(s))ds\right\}, \tag{6}$$

where

$$\xi(t) = \frac{\mu(t) - r(t)}{\sigma(t)}, \qquad \psi(t) = \frac{p_2}{1 - p_2} r(t). \tag{7}$$

The next definition is similar to the one given by Karatzas and Shreve [4].

**Definition 1.** The triplet $(c, w, \theta)$ of consumption, portfolio and insurance processes is *admissible* for the household if the corresponding wealth process satisfies

$$\phi(t)W(t) + E_t\left[\int_t^T \phi(s)y(s)1_{\{N(s-)=0\}}ds\right] + E_t\left[\int_t^T \phi(s)\theta(s)1_{\{N(s-)=0\}}dN(s)\right] \geq 0, \quad t \in \mathcal{T}. \tag{8}$$

The class of admissible processes is denoted by $\mathcal{A}$.

From (6) and (8), and by the arguments similar to Karatzas and Shreve [4], we can readily show that if $(c, w, \theta)$ is admissible, then the consumption process $c$ must satisfy the *budget constraint*

$$E\left[\int_0^T \phi(t)c(t)dt + \phi(T)W(T)\right] \leq E\left[\int_0^T \phi(t)y(t)1_{\{N(t-)=0\}}dt\right] + W_0.$$

**Lemma 1.** *For any pair of consumption process $c$ and terminal wealth $W(T)$ that satisfies*

$$E\left[\int_0^T \phi(s)c(s)ds + \phi(T)W(T)\right] = W_0 + E\left[\int_0^T \phi(s)y(s)1_{\{N(s-)=0\}}ds\right],$$

*there exists a portfolio/insurance processes $(w, \theta)$ such that $(c, w, \theta) \in \mathcal{A}$ and*

$$\phi(t)W(t) = E_t\left[\int_t^T \phi(s)\left(c(s) - (y(s) + \psi(s)\theta(s))1_{\{N(s-)=0\}}\right)ds + \phi(T)W(T)\right]$$

*for any $t \in \mathcal{T}$.*

**Proof.** Let $M(t)$ be a martingale defined by

$$M(t) = E_t\left[\int_0^T \phi(s)\left(c(s) - (y(s) + \psi(s)\theta(s))1_{\{N(s-)=0\}}\right)ds + \phi(T)W(T)\right]. \tag{9}$$

Then, by the martingale representation theorem (see, e.g., Bremaud [2] and Karatzas and Shreve [3]), there exist a progressively measurable process $\pi_1(t)$ and a predictable process $\pi_2(t)$ such that

$$\int_0^T |\pi_1^2(t) + \pi_2(t)|dt < \infty,$$

and satisfying

$$M(t) = W_0 + \int_0^t \pi_1(s)dZ(s) + \int_0^t \pi_2(s)(dN(s) - \lambda(s)ds).$$

On the other hands, from (3) and (6), we can readily show that

$$\begin{aligned} d(\phi(t)W(t)) &= \phi(t)\left[(y(t) + \theta(t)\psi(t))1_{\{N(t-)=0\}} - c(t)\right]dt + \phi(t)(w(t)\sigma(t) - \xi(t)W(t))dZ(t) \\ &\quad + \phi(t)\left[\theta(t)1_{\{N(t-)=0\}}\frac{\psi(t)}{\lambda(t)} + W(t)\frac{\psi(t) - \lambda(t)}{\lambda(t)}\right](dN(t) - \lambda(t)dt). \end{aligned}$$

Thus, if we define $w(t)$ and $\theta(t)$ so as to satisfy

$$\phi(t)(w(t)\sigma(t) - \xi(t)W(t)) = \pi_1(t) \tag{10}$$

and

$$\phi(t)\left[\theta(t)1_{\{N(t-)=0\}}\frac{\psi(t)}{\lambda(t)} + W(t)\frac{\psi(t)-\lambda(t)}{\lambda(t)}\right] = \pi_2(t), \tag{11}$$

respectively, then

$$d(\phi(t)W(t)) = dM(t) - \phi(t)\left[c(t) - (y(t) + \theta(t)\psi(t))1_{\{N(t-)=0\}}\right]dt.$$

The lemma now follows by integrating the above equation over $[0,t)$. □

Now, suppose that the household has a *time-discount factor* $e^{-\int_0^t \rho(s)ds}$, $t \in \mathcal{T}$, where $\rho(t)$ is bounded and adapted to $\mathbb{F}$, and has *utility functions* $u_i : \mathbb{R} \to (0,\infty)$, $i = 1,2$, which are strictly increasing, strictly concave and twice continuously differentiable with properties $u_i'(\infty) \equiv \lim_{x\to\infty} u_i'(x) = 0$ and $u_i'(0+) \equiv \lim_{x\downarrow 0} u_i'(x) = \infty$. The problem that the household faces is formally described as follows:

**(MP)** Given the discount factor and utility functions, find an optimal consumption/portfolio process $(\hat{c}, \hat{w})$ and an optimal insurance process $\hat{\theta}$ to maximize the expected total, discounted utility from consumption and terminal wealth

$$E\left[\int_0^T e^{-\int_0^t \rho(s)ds}u_1(c(t))dt + e^{-\int_0^T \rho(s)ds}u_2(W(T))\right],$$

over the admissible consumption/porttolio/insurance process $(c,w,\theta) \in \mathcal{A}$, that satisfy

$$E\left[\int_0^T \min\left\{0, e^{-\int_0^t \rho(s)ds}u_1(c(t))\right\}dt\right] > -\infty$$

and

$$E\left[\min\left\{0, e^{-\int_0^T \rho(s)ds}u_2(W(T))\right\}\right] > -\infty,$$

respectively.

For each utility function $u_i$, $i = 1,2$, and for each $t \in \mathcal{T}$, we shall denote by $I_i(x,t)$ the inverse function of $\frac{d}{dx}\left[u_i(x)e^{-\int_0^t \rho(s)ds}\right]$. Under the assumptions stated above, for each $t \in \mathcal{T}$, the functions $I_i(x,t)$, $i = 1,2$, exist, and are also continuous, strictly decreasing, and map $(0,\infty)$ onto itself with respect to $x$, with properties $I_i(0+,t) = \infty$ and $I_i(\infty,t) = 0$.

The householder's optimal consumption/wealth process is given by the next theorem whose proof is similar to that of Theorem 3.6.3 in Karatzas and Shreve [4], and it is omitted here.

**Theorem 1.** *Under the conditions stated above, an optimal consumption process $\hat{c}$ and the corresponding wealth process $\hat{W}$ are given, respectively, by*

$$\hat{c}(t) = I_1\left(t, \zeta\phi(t)\right), \qquad t \in \mathcal{T},$$

*where $\zeta$ is a solution of equation*

$$E\left[\int_0^T \phi(t)I_i\left(\zeta\phi(t),t\right)dt + \phi(T)I_2(\zeta\phi(T),T)\right] = W_0 + E\left[\int_0^T \phi(t)y(t)1_{\{N(t)=0\}}dt\right],$$

*and by*

$$\hat{W}(t) = \frac{1}{\phi(t)} E_t \left[ \int_t^T \phi(s)(\hat{c}(s) - (y(s) + \psi(s)\theta(s))\, 1_{\{N(s-)=0\}})ds + \phi(T)\hat{W}(T) \right], \qquad t \in \mathcal{T}, \quad (12)$$

*with* $\hat{W}(T) = I_2(\zeta\phi(T), T)$. *An optimal portfolio/insurance process* $(\hat{w}, \hat{\theta})$ *is given by (10) and (11), respectively, with* $W(t)$ *being replaced by the optimal wealth* $\hat{W}(t)$.

# 3. SOME SPECIAL CASES

In this section, we consider some cases in which the household has specific utility functions. Namely, we study the cases of exponential and power utility functions, and explicitly derive optimal consumption/portfolio/insurance processes $(\hat{c}, \hat{w}, \hat{\theta})$ given by Theorem 1 For simplicity, we assume that $r(t)$, $\mu(t)$, $\sigma(t)$, $\lambda(t)$, $y(t)$, and $\rho(t)$ are positive constants, from which $\xi(t)$ and $\psi(t)$ defined in (7) are also constant, $\xi(t) = \xi$ and $\psi(t) = \psi$ say.

## 3.1 EXPONENTIAL UTILITY FUNCTIONS

First, we consider the case in which the household has utility functions defined by

$$u_i(x) = \frac{1 - \exp(-\eta_i x)}{\eta_i}, \qquad 0 < x < \infty, \qquad i = 1, 2, \quad (13)$$

where $\eta_i$ are positive constant which represent indices of risk aversion.

It is easily seen that the functions $I_i(x, t)$ are given by

$$I_i(x, t) = -\frac{1}{\eta_i}(\ln x + \rho t), \qquad i = 1, 2.$$

From Theorem 1, the optimal consumption process and the optimal terminal wealth are given, respectively, by

$$\hat{c}(t) = -\frac{1}{\eta_1}(\ln \phi(t) + \ln \zeta + \rho t)$$

and

$$\hat{W}(T) = -\frac{1}{\eta_2}(\ln \phi(T) + \ln \zeta + \rho T).$$

Using (12) and a tedious algebra leads to

$$\hat{W}(t) = \hat{c}(t) f(t) \eta_1 - g(t) A - h(t) 1_{\{N(t-)=0\}}, \quad (14)$$

where

$$f(t) = \frac{1}{\eta_1} \frac{1 - e^{-r(T-t)}}{r} + \frac{1}{\eta_2} e^{-r(T-t)},$$

$$g(t) = \frac{1}{\eta_1} \frac{1 - e^{-r(T-t)} - r(T-t)e^{-r(T-t)}}{r^2} + \frac{1}{\eta_2}(T-t)e^{-r(T-t)},$$

$$h(t) = y \frac{1 - e^{-(\psi+r)(T-t)}}{\psi + r} + \psi \int_0^{T-t} \theta(s+t) e^{-(\psi+r)s} ds, \qquad t \in \mathcal{T},$$

and

$$A = \frac{1}{2}\xi^2 + \ln\left(\frac{\psi}{\lambda}\right)\psi - (\psi - \lambda) - r + \rho.$$

It follows, from (9) and (14), that

$$\begin{aligned} dM(t) &= d(\phi(t)\hat{W}(t)) + \phi(t)(\hat{c}(t) - (y + \psi\theta)1_{\{N(t)=0\}})dt \\ &= \phi(t)\left(f(t) - \hat{W}(t)\right)\xi dZ(t) \\ &\quad + \phi(t)\left(\hat{W}(t)\frac{\psi - \lambda}{\lambda} - \left(\ln\left(\frac{\psi}{\lambda}\right)f(t) - h(t)1_{\{N(t-)=0\}}\right)\frac{\psi}{\lambda}\right)(dN(t) - \lambda dt). \end{aligned}$$

Therefore, from (10) and (11), the optimal portfolio and insurance are given, respectively, by

$$w(t) = \frac{\mu - r}{\sigma^2}f(t)$$

and

$$\theta(t) = \frac{1 - e^{-r(T-t)}}{r}y - \ln\left(\frac{\psi}{\lambda}\right)(k_1(t) + k_2(t)),$$

where

$$k_1(t) = \frac{1}{\eta_1}\left(\frac{1 - e^{-r(T-t)} - r(T-t)e^{-r(T-t)}}{r^2}\psi + \frac{1 - e^{-r(T-t)}}{r}\right),$$

and

$$k_2(t) = \frac{1}{\eta_2}e^{-r(T-t)}(1 + \psi(T-t))$$

for all $t \in \mathcal{T}$. The insurance premium is then given by

$$\bar{p} = m_1 y - \ln\left(\frac{\psi}{\lambda}\right)\left(\frac{1}{\eta_1}m_2 + \frac{1}{\eta_2}m_3\right),$$

where

$$\begin{aligned} m_1 &= \frac{\psi}{\psi + r}(1 - e^{-(\psi+r)T}) - (1 - e^{-\psi T}), \\ m_2 &= \frac{\psi}{r^2}(1 - e^{-(\psi+r)T}) - \frac{\psi + r + \psi T}{r^2}(1 - e^{-\psi T}) + \frac{e^{-rT}}{r}(1 - e^{-\psi T} - \psi T e^{-\psi T}), \\ m_3 &= (1 + \psi T)(1 - e^{-\psi T}) - \psi e^{-rT}(1 - e^{-\psi T} - \psi T e^{-\psi T}). \end{aligned}$$

Furthermore, from (14), the optimal consumption is given by

$$\hat{c}(t) = \frac{1}{f(t)\eta_1}\left(\hat{W}(t) + g(t)A + h(t)1_{\{N(t-)=0\}}\right), \qquad t \in \mathcal{T}.$$

## 3.2 POWER UTILITY FUNCTIONS

We next consider the case in which the household have utility functions defined by

$$u_1(x) = u_2(x) = \frac{x^\alpha}{\alpha}, \qquad 0 < x < \infty, \quad \alpha \in (-\infty, 1) \setminus \{0\}, \tag{15}$$

where $\alpha$ represents the shape parameter.

In this case, it is easily seen that, from Theorem 1, the optimal consumption process and the optimal terminal wealth are given, respectively, by

$$\hat{c}(t) = \left(\zeta e^{\rho t}\phi(t)\right)^{\frac{1}{\alpha-1}}$$

and

$$\hat{W}(T) = \left(\zeta e^{\rho T}\phi(T)\right)^{\frac{1}{\alpha-1}}.$$

Again, (12) and a tedious algebra leads to

$$\hat{W}(t) = \hat{c}(t)l(t) - h(t)1_{\{N(t-)=0\}}, \tag{16}$$

where

$$l(t) = \frac{e^{B(T-t)} - 1}{B} + e^{B(T-t)}$$

with

$$B = \frac{1}{2}\frac{\alpha}{(1-\alpha)^2}\xi^2 + \left(\left(\frac{\psi}{\lambda}\right)^{\frac{1}{\alpha-1}}\psi - \lambda\right) - \frac{\alpha}{\alpha-1}\left((\psi-\lambda)+r\right) + \frac{\rho}{\alpha-1}.$$

It follows, from (9) and (16), that

$$dM(t) = -\frac{\phi(t)}{\alpha-1}\left(h(t)1_{\{N(t-)=0\}} + \alpha\hat{W}(t)\right)\xi dZ(t)$$

$$+\phi(t)\left[\hat{W}(t)\frac{\psi-\lambda}{\lambda} + \left\{\left(\left(\frac{\psi}{\lambda}\right)^{\frac{1}{\alpha-1}} - 1\right)\hat{W}(t) + \left(\frac{\psi}{\lambda}\right)^{\frac{1}{\alpha-1}}h(t)1_{\{N(t-)=0\}}\right\}\frac{\psi}{\lambda}\right](dN(t) - \lambda dt). \tag{17}$$

It is clear, from (11) and (17), that $\theta(t)$ generally depends on the value of wealth $\hat{W}(t)$. If, however, the household does not require premium for the mortality risk of the householder, i.e. $\lambda = \psi$, the optimal insurance process and portfolio process are, respectively, given by

$$\hat{\theta}(t) = \frac{1 - e^{-r(T-t)}}{r}y$$

and

$$\hat{w}(t) = \frac{\mu - r}{\sigma^2(1-\alpha)}\left[\hat{W}(t) + \frac{1 - e^{-r(T-t)}}{r}y1_{\{N(t-)=0\}}\right].$$

The premium is therefore given by

$$\bar{p} = m_1 y.$$

Furthermore, from (16), the optimal consumption is given by

$$\hat{c}(t) = \frac{1}{l(t)}\left(\hat{W}(t) + \frac{1 - e^{-r(T-t)}}{r}y1_{\{N(t-)=0\}}\right), \qquad t \in \mathcal{T}.$$

# REFERENCES


1. Z. Bodie, R.C. Merton and W. Samuelson, Labor Supply Flexibility and Portfolio Choice in a Life-Cycle Model, *Journal of Economic Dynamics and Control*, **18**, 427-449 (1992).

2. P. Bremaud, *Point Processes and Queues*, Springer, New York, (1981).

3. I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Second Edition, Springer, New York, (1991).

4. I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer, New York, (1998).

5. R. C. Merton, Life Time Portfolio Selection under Uncertainty, *Review of Economics and Statistics*, **51**, 247-257, (1969).

6. R. C. Merton, Optimum Consumption and Portfolio Rules in a Continuous-time Model, *Journal of Economic Theory*, **3**, 373-413, (1971).

# AN EFFICIENT BACKUP WARNING POLICY FOR A HARD DISK

HAJIME KAWAI[1] and HIROAKI SANDOH[2]

[1] Department of Social Systems Engineering, Tottori University
4-101, Minami, Koyama-cho, Tottori, 680-8550, Japan
[2] Department of Information and Management Science,
University of Marketing and Distribution Sciences
3-1, Gakuen-nishi-machi, Nishi, Kobe, 651-2188, Japan
kawai@sse.tottori-u.ac.jp / sandoh@umds.ac.jp

**Abstract**—In this decade, a hard disk has become an essential key component of a personal computer system. It preserves important information which is frequently updated. In case the hard disk fails, we may possibly lose such important information. This is called a hard disk failure. One of the simplest methods to cope with such a possibility of a hard disk failure is to periodically make a copy of the information to another secondary medium. This is called a backup operation.

This study discusses an efficient backup warning policy which gives us a warning to back up files at the prespecified time $T_w$ measured by the elapsed time since the previous backup operation or the recovery from a hard disk failure. For the purpose of determining the value of $T_w$, this study formulate the efficiency as a criterion, which is defined by the long-run average ratio of (i) the time spent in processing jobs effectively in the sense that their accomplishments are successfully backed up to (ii) the total time spent in processing jobs ineffectively as well as effectively, and spent in backup or recovery operations. We then clarify the conditions under which an optimal warning time exists. A numerical example is also presented.

**Keywords**—Backup, Warning, Hard disk, Efficiency, Optimal warning time

## 1. INTRODUCTION

Hard disks used for an engineering work station or a personal computer can, in recent years, be purchased at lower prices. Furthermore, a variety of application software products for a personal computer are being developed, which require a hard disk. For these reasons, the hard disk has become one of the essential components for a personal computer system as well as an engineering work station system.

A hard disk generally preserves various files, which are frequently updated. However, these files are occasionally lost because of human errors or failures of hardware devices which the computer system consists of. This is called *a hard disk failure*. One of the simplest methods for protecting us from such a serious loss is to make a backup copy of the files on magnetic tapes, removable disks, magnetic optical disks and so forth (*backup disks* for simplicity) periodically. In the case of a hard disk failure, the backup disks can partially recover the hard disk. The recovery will be partial since the data updated after the previous backup operation or the recovery from a hard disk failure cannot be recovered.

Frequent backup operations could significantly reduce the loss at a hard disk failure although they would spend much time in backup operations. On the contrary, rare backup operations could save time in backup operations while the loss time incurred by a hard disk failure would become

very large. These observations indicate the significance of determining an adequate backup timing of files on a suitable criterion.

Similar problems to the above have been discussed for the main internal memory of a main frame computer, where data stored in the main internal memory are sometimes lost because of a system failure. For such a system, many studies have been reported on rollback and recovery strategies[1–12], which provide adequate times to backup data in the main internal memory on a hard disk. These strategies were originally devised for fear of a system failure of an online banking system. In the case of a system failure of such an online banking system, all the data in the main internal memory at a failure must completely be recovered at any rate. For this reason, all the log files are also backed up on a magnetic tape. With both the data backed up on the hard disk and the log files backed up on the magnetic tape, the system data can perfectly be recovered although it spends a great deal of time and cost.

Assuming that the state of the system can perfectly be recovered up to the state at its failure, a formulation based on the renewal reward process[13] is possible. The underling idea in the formulation is quite similar to that in replacement policies for a system in the reliability context[14, 15], where the cost structure depends on the age of the failed unit at its failure[16–19].

In the problems associated with a hard disk for personal computers or workstations (*backup policy problems* for simplicity), it should be reminded that the recovery from a hard disk failure using backup disks is partial, i.e., the hard disk can only be recovered up to the state at the last backup time. This peculiarity makes the backup policy problem more complicated than that of rollback and recovery strategies for the main internal memory.

For backup policy problems, Sandoh, Kaio and Kawai[20] and Sandoh, Kawai and Ibaraki[21] have proposed a backup policy, which suggests to backup files in the hard disk at time $T$ measured by the elapsed time spent in updating or creating files after the last backup operation or the recovery from a hard disk failure, whichever occurred most recently. This policy is called *a time-managed backup policy.* For the purpose of determining the value of $T$, Sandoh, Kaio and Kawai[20] formulated the expected cost per unit time over an infinite time span as an objective function to be minimized. Sandoh, Kawai and Ibaraki[21] introduced the limiting availability as another objective function to be maximized. Sandoh and Kawai[22] have also proposed another backup policy, which insists on backing up files when $N$ jobs of creating or updating files are completed. This is called *a job-managed backup policy.* The limiting availability was introduced that was to be maximized for the purpose of determining an optimal integer $N^*$.

Under the time-managed backup policy, we may have to stop creating or updating files for a backup operation when the elapsed time since the last backup operation or the recovery reaches $T$. Such a problem can be solved by adopting the job-managed backup policy. Under the job-managed backup policy, however, some backup operations may be executed too early and others too late. This is because the processing time of each job is random.

In order to overcome both problems under the time-managed and the job-managed policies, this study proposes a warning policy for backup operations. This policy gives us a warning to back up files at the prespecified time $T_w(> 0)$ measured by the elapsed time since the previous backup operation or the recovery from a hard disk failure. The time to give us a warning is called *a warning time.* In case a job is being processed at the warning time, a backup operation is actually conducted immediately after the process of the job is completed. Such a job is called *a warned job* in the following.

For the purpose of determining the value of $T_w$, this study formulate the efficiency as a criterion, which is defined by the long-run average ratio of (i) the time spent in processing jobs effectively in the sense that their accomplishments are successfully backed up to (ii) the total time spent in processing jobs ineffectively as well as effectively, and spent in backup or recovery operations. If

a warning time $T_w = T_w^*$ maximizes the efficiency, it is optimum. We then clarify the conditions under which such an optimal warning time exists. A numerical example is also presented.

# 2. ASSUPMTIONS AND PROCESS BEHAVIOUR

## 2.1 ASSUPMTIONS

This study makes the following assumptions:

(a) The hard disk failure time $X$, follows an exponential distribution with failure rate $\lambda$, since the failures occur randomly in time. The hard disk failure can instantly be detected.

(b) We only consider the time during which a job is being processed, and thereupon we assume that a hard disk failure occurs only when a system is processing a job.

(c) The processing time $Y$ for each job of updating files is independently and identically distributed, and the cumulative distribution function(*cdf*) and the probability density function (*pdf*)of a processing time are denoted by $H(y)$ and $h(y)$, respectively.

(d) The backing up time at each backup operation consists of a setup time $\tau$ and the time proportional to the total processing time of jobs whose accomplishments are backed up. The proportional constant is denoted by $a$.

(e) The mean recovery time from a hard disk failure is given by $\mu$.

(f) No hard disk failure occurs during a recovery operation although one might occur during a backup operation.

Assumption (b) signifies that we regard a hard disk as an intermittently-used system[23]. Assumption (c) indicates that the *cdf* and the *pdf* of the total processing time for $n$ jobs are respectively given by

$$H_n(t) = \int_0^t H_{n-1}(t-y)dH(y) = H(t) * H_{n-1}(t), \quad n = 2, 3, \cdots, \tag{1}$$

$$H_1(t) = H(t),$$

and

$$h_n(t) = \frac{dH_n(t)}{dt}, \quad n = 1, 2, \cdots. \tag{2}$$

## 2.2 PROCESS BEHAVIOR

Let us here define *the excess age* as the residual processing time of a warned job at $T_w$. The processing times of jobs generate a renewal process[13], and therefore the *cdf*, $G(t)$ of the excess age $T_e$ is given (see, e.g. [13]) by

$$G(t_e) = H(T_w + t_e) - \int_0^{T_w} \overline{H}(T_w - t + t_e)m(t)dt, \tag{3}$$

where

$$m(t) = \sum_{n=1}^{\infty} h_n(t). \tag{4}$$

From assumption (a), the process of the system behavior generates a renewal reward process[13], where the renewal point is assigned to the time when one of the following two events occurs:

(i) The process of the warned job was finished and the backup operation has successfully been carried out, that is, $X > (a+1)(T_w + T_e)$.

(ii) A hard disk failure occurred during or before a backup operation and a recovery from the hard disk failure using backup disks has been completed, that is, $X \leq (a+1)(T_w + T_e)$.

The above case (i) includes the effective time, which is expressed by $T_w + T_e$, and the time over one cycle is $(a+1)(T_w + T_e)$ in this case, where one cycle refers to the time between tow successive renewal points. In the case (ii), the time over one cycle is given by $X$ and there is no effective time over one cycle.

## 3. EFFICIENCY

Let $A(T_w)$ and $B(T_w)$ respectively denote the expected time and the expected effective time over one cycle, then the efficiency $W(T_w)$ is written by

$$W(T_w) = \frac{B(T_w)}{A(T_w)}, \tag{5}$$

where

$$\begin{aligned} A(T_w) &= \int_0^\infty [(a+1)(T_w + t_e) + \tau]\, e^{-\lambda[(a+1)(T_w+t_e)+\tau]} dG(t_e) \\ &\quad + \int_0^\infty \left[ \int_0^{(a+1)(T_w+t_e)+\tau} (x+\mu)\lambda e^{-\lambda x} dx \right] dG(t_e), \end{aligned} \tag{6}$$

$$B(T_w) = \int_0^\infty (T_w + t_e) e^{-\lambda[(a+1)(T_w+t_e)+\tau]} dG(t_e). \tag{7}$$

The first and the second terms of the right-hand-side of Eq. (6) respectively express the above events (i) and (ii), while the right-hand-side of Eq. (7) shows only the event (i). In many cases, however, it is difficult to derive $G(t_e)$ in Eq. (3) in a closed form, and hence, it is also difficult to conduct the subsequent analysis using Eqs. (6) and (7).

On the other hand, $A(T_w)$ and $B(T_w)$ are also given by

$$
\begin{aligned}
A(T_w) &= \int_{T_w}^{\infty} [(a+1)v_w + \tau]\, e^{-\lambda[(a+1)v_w+\tau]} dH(v_w) \\
&+ \int_0^{T_w} \left\{ \int_0^{\infty} [(a+1)v_w + \tau]\, e^{-\lambda[(a+1)v_w+\tau]} h(v_w - t) dv_w \right\} m(t) dt \\
&+ \int_{T_w}^{\infty} \left[ \int_{v_w}^{(a+1)v_w+\tau} (x+\mu)\lambda e^{-\lambda x} dx \right] dH(v_w) \\
&+ \int_0^{T_w} \left\{ \int_{T_w}^{\infty} \left[ \int_{v_w}^{(a+1)v_w+\tau} (x+\mu)\lambda e^{-\lambda x} dx \right] h(v_w - t) dv_w \right\} m(t) dt \\
&+ \int_{T_w}^{\infty} (x+\mu)\overline{H}(x) \\
&+ \int_0^{T_w} \left[ \int_{T_w}^{\infty} (x+\mu)\overline{H}(x-t)\lambda e^{-\lambda x} dx \right] m(t) dt \\
&+ \int_0^{T_w} (x+\mu)\lambda e^{-\lambda x} dx, \qquad (8)
\end{aligned}
$$

$$
\begin{aligned}
B(T_w) &= \int_{T_w}^{\infty} v_w e^{-\lambda[(a+1)v_w+\tau]} dH(v_w) \\
&+ \int_0^{T_w} \left\{ \int_0^{\infty} v_w e^{-\lambda[(a+1)v_w+\tau]} h(v_w - t) dv_w \right\} m(t) dt. \qquad (9)
\end{aligned}
$$

where variables $v_w$, $x$, $t$ in the above equations, are used to express the accomplishment time of the warned job, the hard disk failure time, and the completion time of the job processed just prior to the warned job, respectively.

Each term in the right-hand-side of Eq. (8) respectively expresses each of the seven cases listed below:

(a) A warning had been given to the job immediately after the renewal point, and a backup operation has successfully been completed after the warned job was processed.

(b) At least one job had been processed before the warning, and a backup operation has successfully been completed after the warned job was processed.

(c) A warning had been give to the job immediately after the renewal point, and a hard disk failure occurred during a backup operation.

(d) At least one job had been processed before the warning, and a hard disk failure occurred during a backup operation for the accomplishments of processed jobs.

(e) A warning had been given to the job immediately after the renewal point, and a hard disk failure occurred after the warning and before the warned job was processed.

(f) At least one job had been processed before the warning, and a hard disk failure occurred after the warning and before the warned job was processed.

(g) A hard disk failure occurred before the warning.

The first and the second term in the right-hand-side of Eq. (9) express the above cases (a) and (b) respectively since the other cases include no efficient time in themselves. From Eqs. (8) and (9), we obtain

$$A(T_w) = \left(\frac{1}{\lambda}+\mu\right)\left[1-Q+\lambda bP\int_0^{T_w} m(t)e^{-\lambda bt}dt\right], \tag{10}$$

$$B(T_w) = R+\int_0^{T_w}(R-\lambda bPt)\,m(t)e^{-\lambda bt}dt, \tag{11}$$

where

$$P = e^{-\lambda\tau}\int_0^{\infty}\overline{H}(t)e^{-\lambda bt}dt, \tag{12}$$

$$Q = e^{-\lambda\tau}\int_0^{\infty}h(t)e^{-\lambda bt}dt, \tag{13}$$

$$R = e^{-\lambda\tau}\int_0^{\infty}th(t)e^{-\lambda bt}dt, \tag{14}$$

$$b = a+1. \tag{15}$$

The above results yield

$$W(T_w) = \frac{R+\int_0^{T_w}(R-\lambda bPt)\,m(t)e^{-\lambda bt}dt}{\left(\frac{1}{\lambda}+\mu\right)\left[1-Q+\lambda bP\int_0^{T_w}m(t)e^{-\lambda bt}dt\right]}. \tag{16}$$

We have formulated the efficiency of the proposed policy. If $T_w = T_w^*$ maximizes $W(T_w)$, it is optimum. In the succeeding section, we will examine the existence of such $T_w^*$.

## 4. EFFICIENT WARNING POLICY

By differentiating $W(T_w)$ in Eq. (5) with respect to $T_w$, we have

$$W'(T_w) = \frac{B'(T_w)A(T_w)-A'(T_w)B(T_w)}{A^2(T_w)}$$

$$= \frac{A'(T_w)}{A^2(T_w)}\left[\frac{B'(T_w)}{A'(T_w)}A(T_w)-B(T_w)\right]. \tag{17}$$

From Eq. (10), we have

$$A'(T_w) = \lambda\left(\frac{1}{\lambda}+\mu\right)bPm(T_w)e^{-\lambda bT_w} > 0. \tag{18}$$

It follows that the sign of $W'(T_w)$ agrees with that of $D(T_w)$, which is defined by

$$D(T_w) = \frac{B'(T_w)}{A'(T_w)} A(T_w) - B(T_w). \tag{19}$$

From Eqs. (10), (11) and (19), we obtain

$$\lim_{T_w \to \infty} D(T_w) = -\infty < 0, \tag{20}$$

$$\lim_{T_w \to +0} D(T_w) = \frac{R}{\lambda b P} (1 - Q - \lambda b P). \tag{21}$$

Equations (12) and (13) reveal the relationship between $P$ and $Q$, which is expressed by

$$Q = -\lambda b P + e^{-\lambda \tau}, \tag{22}$$

and thus we have

$$\lim_{T_w \to +0} D(T_w) = \frac{R}{\lambda b P} \left(1 - e^{-\lambda \tau}\right) \begin{cases} = 0, & \tau = 0 \\ > 0, & \tau > 0. \end{cases} \tag{23}$$

On the other hand, Eq. (19) yields

$$D'(T_w) = \left[\frac{B'(T_w)}{A'(T_w)}\right]' A(T_w). \tag{24}$$

Since $A(T_w) > 0$, the sign of $D'(T_w)$ coincides with that of $[B'(T_w)/A'(T_w)]'$, which satisfies

$$\left[\frac{B'(T_w)}{A'(T_w)}\right]' = -\frac{1}{\frac{1}{\lambda} + \mu} < 0. \tag{25}$$

From Eqs. (20), (23) and (25), the existence of an efficient warning time can be discussed for the following two cases:

(1) $\tau > 0$ (i.e., the setup time cannot be neglected):

In this case, the sign of $D(T_w)$ changes from positive to negative, and thus there exists a unique finite positive warning time $T_w (> 0)$.

(2) $\tau = 0$ (i.e., the setup time is negligibly small):

In this case, we have $D(T_w) < 0$ for $T_w > 0$, and therefore $T_w \to +0$. This result suggests to backup files as frequently as possible.

## 5. NUMERICAL EXAMPLES

This section illustrates the proposed warning policy assuming that the processing time of each job independently and identically follows a gamma distribution with shape parameter 2 whose *cdf* and *pdf* are respectively given by

$$H(t) = 1 - (1 + \alpha t) e^{-\alpha t}, \tag{26}$$

and

$$h(t) = \alpha^2 t e^{-\alpha t}. \tag{27}$$

Under $H(t)$ in Eq. (26), we have

$$P = \frac{2\alpha+\lambda b}{(\alpha+\lambda b)^2}e^{-\lambda\tau}, \tag{28}$$

$$Q = \frac{\alpha^2}{(\alpha+\lambda b)^2}e^{-\lambda\tau}, \tag{29}$$

$$R = \frac{2\alpha^2}{(\alpha+\lambda b)^3}e^{-\lambda\tau}, \tag{30}$$

$$m(t) = \frac{\alpha}{2}\left(1-e^{-2\alpha t}\right). \tag{31}$$

Hence, $A(T_w)$ and $B(T_w)$ in Eqs. (10) and (11) respectively become

$$A(T_w) = \left(\frac{1}{\lambda}+\mu\right)\left\{1-Q+\frac{\alpha P}{2}\left(1-e^{-\lambda b T_w}\right)-\frac{\alpha\lambda b P}{4\alpha+2\lambda b}\left[1-e^{-(2\alpha+\lambda b)T_w}\right]\right\}, \tag{32}$$

and

$$\begin{aligned} B(T_w) &= R\left\{1+\frac{\alpha}{2\lambda b}\left(1-e^{-\lambda b T_w}\right)-\frac{\alpha}{2(2\alpha+\lambda b)}\left[1-e^{-(2\alpha+\lambda b)T_w}\right]\right\} \\ &\quad +\frac{\alpha\lambda b P}{2}\left\{\frac{1}{(\lambda b)^2}\left(1-e^{-\lambda b T_w}\right)-\frac{1}{(2\alpha+\lambda b)^2}\left[1-e^{-(2\alpha+\lambda b)T_w}\right]\right. \\ &\quad \left.-T_w\left[\frac{e^{-\lambda b T_w}}{\lambda b}-\frac{e^{-(2\alpha+\lambda b)T_w}}{2\alpha+\lambda b}\right]\right\}. \end{aligned} \tag{33}$$

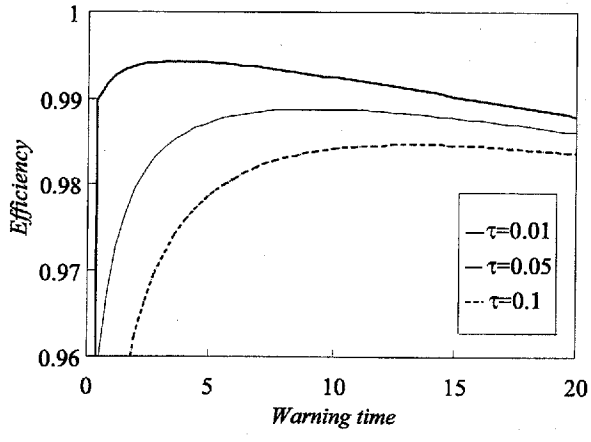


Figure 1: Efficiency

Figure 1 shows the efficiency when $\tau = 0.01$, 0.05 and 0.1 for $(\lambda, a, \mu, \alpha) = (0.001, 0.001, 0.25, 2.0)$. It is signified by $\alpha = 2$ in the gamma distribution with shape parameter 2 that the

Table 1: Efficient warning times.

$(a, \mu, \alpha) = (0.001, 0.25, 2.0)$

| $\lambda$ | $\tau = 0.01$ | | $\tau = 0.05$ | | $\tau = 0.1$ | |
|---|---|---|---|---|---|---|
| | $T_w^*$ | $W(T_w^*)$ | $T_w^*$ | $W(T_w^*)$ | $T_w^*$ | $W(T_w^*)$ |
| 0.0010 | 3.72 | 0.9943 | 9.21 | 0.9888 | 13.31 | 0.9847 |
| 0.0020 | 2.41 | 0.9922 | 6.29 | 0.9844 | 9.18 | 0.9787 |
| 0.0030 | 1.84 | 0.9905 | 4.99 | 0.9810 | 7.34 | 0.9740 |
| 0.0040 | 1.49 | 0.9890 | 4.22 | 0.9782 | 6.25 | 0.9700 |
| 0.0050 | 1.26 | 0.9877 | 3.69 | 0.9756 | 5.51 | 0.9665 |
| 0.0060 | 1.09 | 0.9865 | 3.30 | 0.9732 | 4.96 | 0.9633 |
| 0.0070 | 0.96 | 0.9854 | 3.00 | 0.9711 | 4.53 | 0.9604 |
| 0.0080 | 0.85 | 0.9843 | 2.76 | 0.9690 | 4.19 | 0.9576 |
| 0.0090 | 0.76 | 0.9832 | 2.56 | 0.9671 | 3.90 | 0.9550 |
| 0.0100 | 0.69 | 0.9822 | 2.39 | 0.9652 | 3.66 | 0.9526 |
| 0.0200 | 0.34 | 0.9726 | 1.47 | 0.9500 | 2.36 | 0.9324 |
| 0.0300 | 0.22 | 0.9632 | 1.07 | 0.9378 | 1.78 | 0.9165 |
| 0.0400 | 0.16 | 0.9538 | 0.83 | 0.9272 | 1.44 | 0.9030 |
| 0.0500 | 0.12 | 0.9446 | 0.67 | 0.9174 | 1.21 | 0.8909 |
| 0.0600 | 0.10 | 0.9354 | 0.56 | 0.9081 | 1.04 | 0.8798 |
| 0.0700 | 0.09 | 0.9263 | 0.48 | 0.8991 | 0.91 | 0.8695 |
| 0.0800 | 0.07 | 0.9173 | 0.42 | 0.8904 | 0.81 | 0.8598 |
| 0.0900 | 0.07 | 0.9085 | 0.37 | 0.8819 | 0.72 | 0.8506 |
| 0.1000 | 0.06 | 0.8997 | 0.33 | 0.8735 | 0.65 | 0.8417 |

mean processing time of each job is equal to 1.0 (hour, e.g.). It can be observed in Fig. 1 that the efficient warning time becomes larger as the setup time, $\tau$ increases. In addition, the efficiency becomes smaller on the whole when $\tau$ increases. This is because the time for a setup operation is regarded as being inefficient in this study.

Table 1 reveals the efficient warning times in the case of $(a, \mu, \alpha) = (0.001, 0.25, 2.0)$. Table 1 indicates the efficiency corresponding to the efficient warning time as well.

From Table 1, we can see that the efficient warning time decreases with increasing failure rate, $\lambda$. It is also seen that the efficiency decreases on the whole as the setup time increases and that the setup time does not affect the efficient warning time significantly when $\lambda$ takes a large value.

# 6. CONCLUSIONS

This paper proposed an efficient backup warning policy for a hard disk of an engineering workstation or a personal computer, where a warning for a backup operation is given at the elapsed time $T_w(> 0)$ since the last backup operation or the recovery from a hard disk failure. If a warning is given while we are processing a job, a backup operation is carried out immediately after we finish processing the job. The efficiency was adopted as a criterion to be maximized. It was then shown that there exists a unique efficient warning time $T_w^*$ if the setup time for a backup operation cannot be neglected. A numerical example was also presented to illustrate the theoretical underpinnings of the proposed backup warning policy formulation.

# REFERENCES


1. K.M. Chandy and C.V. Ramamoorthy, Rollback and recovery strategies for computer programs, *IEEE Trans. Computer*, **C21**, 546-556 (1972).

2. K.M. Chandy, J.C. Browne, C.W. Dissly and W.R. Uhrig, Analytical models for rollback and recovery strategies in data base system, *IEEE Trans. Software Engineering*, **SE-1**, 100-110 (1975).

3. K.M. Chandy, A survey of analytic models of rollback and recovery strategies, *Computer*, **8**, 40-47 (1975).

4. J.S.M. Verhofstadt, Recovery techniques for database systems, *ACM Computing Surveys*, **10**, 167-195 (1978).

5. F. Faccelli, Analysis of a service facility with periodic checkpointing, *Acta Informatica*, **15**, 67-81 (1981).

6. V.F. Nicola and F.J. Kylstra, A model of checkpointing and recovery with a specified number of transactions between checkpoints, *Performance 83*, (Edited by A.K. Agrawalla and S.K. Tripathi), pp. 83-100, North–Holland, Amsterdam, (1983).

7. A. Reuter, Performance analysis of recovery techniques, *ACM tods*, **9**, 526-559 (1984).

8. S. Toueg and Ö. Babaoğlu, On the optimum checkpoint selection problems, *SIAM J. Computer*, **13**, 630-649 (1984).

9. N. Kaio and S. Osaki, A note on optimum checkpointing policies, *Microelectronics & Reliability*, **25**, 451-453 (1985).

10. R. Koo and S. Toueg, Checkpointing and rollback–recovery for distributed systems, *IEEE Trans. Software Engineering*, **SE-13**, 23-31 (1987).

11. U. Sumita, N. Kaio and P.B. Goes, Analysis of effective service time with age dependent interruptions and its application to rollback policy for database management, *Queueing Systems*, **4**, 193-212 (1989).

12. S. Fukumoto, N. Kaio and S. Osaki, Evaluation for a database recovery action with periodical checkpoint generations, *Trans. Electronics, Information and Communication Engineers*, **E-74**, 2076-2082 (1991).

13. S.M. Ross, *Applied Probability Models with Optimization Applications*, Holden–Day, California, (1970).

14. R.E. Barlow and F. Proschan, *Mathematical Theory of Reliability*, John Wiley; New York, (1967).

15. R.E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, New York, (1975).

16. R. Scheaffer, Optimum age replacement policies with an increasing cost factor, *Technometrics*, **13**, 139-144 (1971).

17. R. Cléroux and M. Hanscom, Age replacement with adjustment and depreciation costs and interest charges, *Technometrics*, **16**, 235-239 (1974).

18. C. Tilquin and R. Cléroux, Block replacement policies with general cost structures, *Technometrics*, **17**, 291-298 (1975).

19. A. Ran and S.I. Rosenlund, Age replacement with discounting for a continuous maintenance cost model, *Technometrics*, **18**, 459-465 (1976).

20. H. Sandoh, N. Kaio and H. Kawai, On backup policies for a hard computer disk, *Reliability Engineering and System Safety*, **37**, 29-32 (1992).

21. H. Sandoh, H. Kawai and T. Ibaraki, An optimal backup policy for a hard computer disk depending on age under availability criterion, *Computers & Mathematics with Applications*, **24**, 57-62 (1992).

22. H. Sandoh and H. Kawai, An optimal $N$–job backup policy maximizing availability for a hard computer disk, *J. Operations Research Society of Japan*, **34**, 383-390 (1991).

23. H. Mine and H. Kawai, Preventive replacement of an intermittently-used system, *IEEE Trans. Reliability*, **R-30**, 391-392 (1981).

# RELIABILITY OF A COMMUNICATION SYSTEM WITH LIMITED NUMBER OF ROLLBACK

MITSUTAKA KIMURA[1], KAZUMI YASUI[2], TOSHIO NAKAGAWA[2] and NAOHIRO ISHII[1]

[1]Department of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan

[2]Department of Industrial Engineering, Aichi Institute of Technology, 1247 Yachigusa, Yakusa-cho, Toyota 470-0392, Japan

kimura@egg.ics.nitech.ac.jp / {yasuilab,nakagawa}@ie.aitech.ac.jp / ishii@ics.nitech.ac.jp

**Abstract**—This paper considers a communication system which consists of two processors, and studies the problem for improving its reliability by adopting the recovery techniques of checkpoint and rollback : When either processor failure or communication error occurs, the rollback operation for processors associated with such an event is carried out to the most recent checkpoint. If the rollback recovery for processors has been executed at $k$ times successively, we regard that the system has become a faulty state permanently, and interrupt it. Then, the inspection and maintenance are made and after that, the system is recovered successsfully and restarts again from the beginning of its initial state. We formulate the stochastic model with the above recovery techniques, and derive the mean time to checkpoint, the expected number of rollback operation and interruption. Further, an optimal checkpointing interval which minimizes the expected cost is analytically discussed under the assumption that the number of rollback operation is limited. Finally, some numerical examples are given and useful discussions are made.

**Keywords**—Rollback recovery, Limited rollback, Communication system, Expected cost, Checkpointing interval.

## 1. INTRODUCTION

As a computer communication technology has remarkably developed, efficient control mechanisms of a system have been actually realized by a number of processors [1]. Hence, the processing of each processor has to be carried out accurately and fast. Moreover, a system needs to restore rapidly a consistent state after transient faults, to improve the reliability of communications between processors [2]. This paper considers a communication system which consists of two processors, and studies the problem for improving its reliability, by adopting the recovery techniques of checkpoint and limited number of rollback [3], [4].

Several algorithms of rollback recovery with checkpoints have been already proposed to keep a system consistent when transient faults occur. In the previous model [6], we discussed the policy that when either processor failure or communication error occurs, the rollback operation for processors associated with such event is executed to the most recent checkpoint, and so that, the consistent state in the whole system is always maintained. That is, it was assumed that both processor failures and communication errors are transient, i.e., these are unlikely to recur after rollback operation. In this paper, we assume that the system becomes like failure if the number of rollback operation for processors is greater than a threshold level. That is, if the rollback recovery for processors has been

executed at $k$ times successively, we regard that the system has become a faulty state permanently, and interrupt it. Then, we make the inspection and maintenance of the system, and after that, it is recovered successsfully and restarts again from the beginning of its initial state.

We formulate a stochastic model of a communication system with the above recovery policy: The mean time between checkpoints, the expected number of rollback operations due to processor failures or communication errors, and the expected number of interruptions are obtained, using the theory of Markov renewal processes [5]. Further, we derive the expected cost and discuss analytically an optimal checkpointing interval which minimizes it, under the assumption that the number of rollback operation is limited. Finally, some numerical examples are given.

## 2. MODEL AND ANALYSIS

The system consists of two processors, which is called A and B, and the control mechanisms are realized by communications between processors. We observe only about communicaton behavior of processor A.

(1) The system begins to operate at time 0, and takes checkpoints for all processes that are relevant to the operation of A at scheduled time $T$. Any transmissions which have not finished until time $T$ deal with no transmission with each other.

(2) The demand for transmissions between A and B has a general distribution A(t) with mean $1/\alpha$.

(3) A message is divided into n pieces of segments because it is necessary to ensure the reliability of transmissions, and each segment is sent from a sender to a receiver with acknowledgment by handshake as follows:

  (i) Each corresponding answer of ACK( positive acknowledgement) or NAK(negative acknowledgement) from a receiver to a sender judges whether the transmission of a segment succeeds or does not. The communication of a message terminates when $n$ times of ACK have been accepted from a receiver.

  (ii) When NAK has been received or no answer has been recieved until a limited time, we retransmit a message of the same segment. If the retransmission does not succeed again, it is judged that communication errors have occurred.

  (iii) The time needed for the transmission of a segment has a distribution $a(t)$, and the probability that it succeeds is $p(0 < p \leq 1)$.

(4) Failures of processor A and processors B occur independently according to distributions $F_A(t)$ and $F_B(t)$, respectively. Then, we define the probability distribution $\overline{F}(t) \equiv \overline{F_A}(t)\overline{F_B}(t)$ with mean $1/\lambda$, where $\overline{\Phi}(t) \equiv 1 - \Phi(t)$ represents a survival function of any function $\Phi(t)$.

(5) When either processor failures or communication errors have occurred, the rollback operation for processors associated with such events is executed from that time to its most recent checkpoint.

  (i) Any transmissions which have not finished until that time deal with no transmission with each other.

  (ii) The system is regenerated by the rollback operation.

  (iii) The time required for rollback recovery has a general distribution $G(t)$ with mean $1/\mu$.

(6) If the rollback recovery for processors has been executed at $k$ times successively, the system is inspected and maintained. After that, the system restarts again from the beginning of its initial state.

(i) The system is regenerated by the inspection and maintenance.

(ii) The total time required for the inspection and maintenance has a general distribution $V(t)$ with mean $v$.

Under the above assumptions, we define the following states of the system:

State $S_0$: The system begins to operate or restart.

State $S_F$: Either processor failure or communication error occur and the rollback recovery starts.

State $S_K$: Rollback operation has executed at $k$ times, and the inspection and maintenance starts.

State $S_T$: Checkpoint of the system is made at time $T$.

The system states defined above form a Markov renewal process [6], where $S_T$ is an absorbing state and $S_0$ is a regeneration point.

We can derive the mean time $\ell_{S_0,S_T}$ from the beginning of operation to the next checkpoint, from Appendix 1:

$$\ell_{S_0,S_T} = \frac{1}{\overline{F}(T)\overline{X}(T)} \left[ \int_0^T \overline{F}(t)\overline{X}(t)dt + \frac{1}{\mu}[1 - \overline{F}(T)\overline{X}(T)] \right] + \frac{v[1 - \overline{F}(T)\overline{X}(T)]^k}{1 - [1 - \overline{F}(T)\overline{X}(T)]^k}. \quad (1)$$

Note that $X(t)$ is a probability distribution that communication errors occur. The expected number of rollback operations caused by processor failures or communication errors and the expected number of interruptions are, respectively, from Appendix 2,

$$M_F = \frac{1}{\overline{F}(T)\overline{X}(T)} - 1, \quad (2)$$

$$M_K = \frac{1}{1 - [1 - \overline{F}(T)\overline{X}(T)]^k} - 1. \quad (3)$$

## 3. OPTIMAL CHECKPOINTING INTERVAL

Let $c_1$ be the cost for the operation of the system, $c_2$ be the cost for a rollback recovery of communication errors or processor failures, and $c_3$ be the cost for inspection and maintenance. We define that the expected cost per unit of time until the next checkpoint is

$$C(T) \equiv \frac{c_1 + c_2 M_F + c_3 M_K}{\ell_{S_0,S_T}}. \quad (4)$$

We seek an optimal checkpointing interval which minimizes $C(T)$ in Eq.(4) for $c_3 \geq c_2 > c_1$, and discuss analytically it. From Eq.(1),Eq.(2) and Eq.(3), we can rewrite Eq.(4) as follows:

$$C(T) = \frac{c_1\overline{Y}(T)[1-Y(T)^k] + c_2Y(T)[1-Y(T)^k] + c_3\overline{Y}(T)Y(T)^k}{\left[\int_0^T \overline{Y}(t)dt + \frac{1}{\mu}Y(T)\right][1-Y(T)^k] + v\overline{Y}(T)Y(T)^k}, \tag{5}$$

where

$$Y(T) \equiv 1 - \overline{F}(T)\overline{X}(T).$$

Let $\gamma(t) \equiv y(t)/\overline{Y}(t)$ where $y(t)$ is a density of $Y(t)$. Differentiating $C(T)$ in Eq.(5) with respect to $T$ and setting it equal to zero, we have

$$\frac{\left[\begin{array}{c} \gamma(T)\left[\int_0^T \overline{Y}(t)dt + \frac{1}{\mu}Y(T)\right]\left(\frac{c_3}{c_2-c_1}\right)Y(T)^{k-1}\left[\frac{k\overline{Y}(T)}{1-Y(T)^k} - Y(T)\right] \\ + \left[\gamma(T)\int_0^T \overline{Y}(t)dt - Y(T)\right]\left[1-Y(T)^k\right] \\ + \gamma(T)v\left\{\frac{kY(T)^{k-1}(\overline{Y}(T))^2}{1-Y(T)^k} - \left(\frac{c_2}{c_2-c_1}\right)Y(T)^{k-1}\left[\frac{k\overline{Y}(T)}{1-Y(T)^k} - Y(T)\right]\right\} \\ - \left[1+\frac{1}{\mu}\gamma(T)\right]\left(\frac{c_3}{c_2-c_1}\right)\overline{Y}(T)Y(T)^k \end{array}\right]}{\left[1+\frac{1}{\mu}\gamma(T)\right]\left[1-Y(T)^k\right]}$$

$$= \frac{c_1}{c_2-c_1}. \tag{6}$$

Denoting the left-hand side of Eq.(6) by $L_k(T)$, we have the following policy, from Appendix 3:

(i) If $\gamma(t)$ is strictly increasing in t, $\int_0^\infty \overline{Y}(t)dt + 1/\mu > c_1v/(c_3+c_2-c_1)$ and $L_1(\infty) > c_1/(c_2-c_1)$, there exists a finite and unique $T_1^*$ which satisfies $L_1(T) = c_1/(c_2-c_1)$.

(ii) If $\gamma(t)$ is strictly increasing in t, $\int_0^\infty \overline{Y}(t)dt - 1/\gamma(\infty) > c_1v/c_3$ and $L_\infty(\infty) > c_1/(c_2-c_1)$, there exists a finite and unique $T_\infty^*$ which satisfies $L_\infty(T) = c_1/(c_2-c_1)$, and is $T_1^* \le T_\infty^*$.

(iii) If $L_k(T)$ is strictly decreasing in $k$, there exists a finite and unique $T_k^*$ which satisfies $L_k(T) = c_1/(c_2-c_1)$, and $T_1^* \le T_k^* \le T_\infty^* (k = 2, 3, \cdots)$.

## 4. NUMERICAL EXAMPLE

We consider the paticular case that $A(t)$ is exponential and the transmission time of a segment can be neglected because it is much smaller than the other times, i.e., $A(t) \equiv 1 - e^{-\alpha t}$ and $a(t) \equiv 1$ for $t \ge 0$. Let $\lambda(t) \equiv f(t)/\overline{F}(t)$ where $f(t)$ is a density of $F(t)$. Then, we can rewrite (6) as

$$\frac{\left[\begin{array}{c} [\lambda(T)+\alpha(1-x)]\left[\int_0^T \overline{Y}(t)dt + \frac{1}{\mu}Y(T)\right]\left(\frac{c_3}{c_2-c_1}\right)Y(T)^{k-1}\left[\frac{k\overline{Y}(T)}{1-Y(T)^k} - Y(T)\right] \\ + [\lambda(T)+\alpha(1-x)]\left[\int_0^T \overline{Y}(t)dt - Y(T)\right]\left[1-Y(T)^k\right] \\ + [\lambda(T)+\alpha(1-x)]v\left\{\frac{kY(T)^{k-1}(\overline{Y}(T))^2}{1-Y(T)^k} - \left(\frac{c_2}{c_2-c_1}\right)Y(T)^{k-1}\left[\frac{k\overline{Y}(T)}{1-Y(T)^k} - Y(T)\right]\right\} \\ - \left[1+\frac{1}{\mu}[\lambda(T)+\alpha(1-x)]\right]\left(\frac{c_3}{c_2-c_1}\right)\overline{Y}(T)Y(T)^k \end{array}\right]}{\left[1+\frac{1}{\mu}[\lambda(T)+\alpha(1-x)]\right]\left[1-Y(T)^k\right]}$$

$$= \frac{c_1}{c_2 - c_1}. \tag{7}$$

where $x \equiv [p(2-p)]^n (0 < x \leq 1)$ and $Y(t) = 1 - \overline{F}(t)e^{-\alpha(1-x)t}$.

We compute numerically an optimal checkpointing interval $T^*$ which satisfies Eq.(7). It is assumed that failures of processor A or processor B are caused by random factors of processors. Thus, failures occur according to a Gamma distribution with order 2, i.e., $F(t) \equiv 1-(1+2\lambda t)e^{-2\lambda t}$.

Suppose that the mean time $1/\mu$ of rollback is a unit of time, the mean time of failures is $\mu/\lambda$ = 1800 or 3600, the mean time of inspection and maintenance is $v$ = 60 or 360, the mean time of demand for communcations between A and B is $\mu/\alpha$ = 30. For example, when $1/\mu$ = 1 second, $1/\lambda$ = 30, 60 minutes. The number of segments is $n$ = 1, 4, 8, and the transmission of undivided message fails with probability $q$, and hence, the probability of accepting ACK for one segment when a message is divided into $n$ is $p \equiv 1 - q/n$.

Introduce the following costs : A cost of checkpoint is $c_1$ = 1, the loss costs of rollback recovery for communication errors and processor failures are $c_2/c_1$ = 10, and the loss costs of inspection and maintenance are $c_3/c_2$ = 2, 4.

Table 1 gives optimal checkpointing intervals $\mu T^*/60$ and expected cost $C(T^*) \times 10^{-4}$ when $q$ = 0.1, $\mu/\alpha$ = 30 and $c_2/c_1$ = 10. These values are scaled to a unit of minute in time. This shows that $T^*$ decrease with $c_3/c_2$ and $n$, increase with $\mu/\lambda$ and $\mu v$ for the same value $c_3/c_2$. Similarly, $T^*$ also increase and the expected cost $C(T^*)$ decrease with $k$ for the same value $n$. Hence, it is better to make the checkpoint at a maximum $T^*$ when $k$ goes to infinity. However, when $k$ is large, $T^*$ little depend on $c_3/c_2$, and become constant.

Table 1 Numerical values of optimal time $\mu T^*/60$ and expected cost $C(T^*) \times 10^{-4}$ to minimize $C(T)$ when $q$ = 0.1, $\mu/\alpha$ = 30 and $c_2/c_1$ = 10.

| $c_3/c_2$ | $\mu v$ | $\mu/\lambda$ | – | $n=1$ | | | $n=4$ | | | $n=8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $k$ | | | $k$ | | | $k$ | | |
| | | | | 1 | 4 | $\infty$ | 1 | 4 | $\infty$ | 1 | 4 | $\infty$ |
| 2 | 60 | 1800 | $C(T^*) \times 10^{-4}$ | 170.3 | 73.3 | 71.5 | 100.6 | 48.5 | 48.2 | 88.8 | 44.5 | 44.3 |
| | | | $\mu T^*/60$ | 5.0 | 8.7 | 10.7 | 4.8 | 9.7 | 10.4 | 4.8 | 9.8 | 10.3 |
| | | 3600 | $C(T^*) \times 10^{-4}$ | 13.3 | 54.1 | 51.3 | 62.6 | 28.3 | 27.9 | 50.6 | 24.2 | 24.1 |
| | | | $\mu T^*/60$ | 10.1 | 14.5 | 22.4 | 9.7 | 18.8 | 20.9 | 9.6 | 19.4 | 20.7 |
| | 360 | 1800 | $C(T^*) \times 10^{-4}$ | 149.2 | 73.1 | 71.5 | 94.1 | 48.5 | 48.2 | 84.5 | 44.5 | 44.3 |
| | | | $\mu T^*/60$ | 5.5 | 8.7 | 10.7 | 5.2 | 9.7 | 10.3 | 5.1 | 9.8 | 10.3 |
| | | 3600 | $C(T^*) \times 10^{-4}$ | 119.0 | 53.9 | 51.3 | 59.8 | 28.3 | 27.9 | 48.9 | 24.2 | 24.1 |
| | | | $\mu T^*/60$ | 10.9 | 14.8 | 22.4 | 10.1 | 18.9 | 20.9 | 9.9 | 19.4 | 20.6 |
| 4 | 60 | 1800 | $C(T^*) \times 10^{-4}$ | 260.2 | 74.5 | 71.5 | 142.2 | 48.8 | 48.2 | 122.2 | 44.7 | 44.3 |
| | | | $\mu T^*/60$ | 3.6 | 7.9 | 10.7 | 3.5 | 9.3 | 10.3 | 3.5 | 9.5 | 10.3 |
| | | 3600 | $C(T^*) \times 10^{-4}$ | 210.8 | 55.6 | 51.3 | 91.7 | 28.6 | 27.9 | 71.5 | 24.4 | 24.1 |
| | | | $\mu T^*/60$ | 7.3 | 12.6 | 22.4 | 7.1 | 17.7 | 20.9 | 7.1 | 18.6 | 20.6 |
| | 360 | 1800 | $C(T^*) \times 10^{-4}$ | 229.9 | 74.4 | 71.5 | 134.3 | 48.8 | 48.2 | 116.8 | 44.7 | 44.3 |
| | | | $\mu T^*/60$ | 4.0 | 8.0 | 10.7 | 3.7 | 9.3 | 10.3 | 3.7 | 9.5 | 10.3 |
| | | 3600 | $C(T^*) \times 10^{-4}$ | 189.1 | 55.5 | 51.3 | 88.0 | 28.5 | 27.9 | 69.5 | 24.4 | 24.1 |
| | | | $\mu T^*/60$ | 7.8 | 12.7 | 22.4 | 7.3 | 17.8 | 20.9 | 7.2 | 18.7 | 20.6 |

## 5. CONCLUSIONS

We have considered the reliability of a communication system by applying the recovery tequniques of checkpoint and rollback, under the assumption that the number of rollback operation is limited: We

have formulated the stochastic model where the consistent state is restored by rollback when either processor failures or communication errors have occurred, and if the rollback recovery for processors has been executed at $k$ times successively, we interrupt the system operation and make the inspection and maintenance. We have derived the mean time to checkpoint, the expected number of rollback recovery by processor failures or communication errors, and the expected number of interruption. Further, we have discussed analytically the optimal checkpointing interval which minimizes the expected cost.

From the numerical example, we have shown that the optimal checkpointing interval decreases with the rate of costs for rollback operations and interruption, and increases with limited number of rollback operations. Moreover, we have understood that optimal checkpointing interval reaches mostly a fixed value which is given by the parameters of $\mu/\lambda$ and $k$.

After this, it would be important to improve and evaluate the reliability of a system with multi-communicaton mechanisms from various practical viewpoints.

## REFERENCES


1. K. Yoneda, T. Matsubara and Y. Koga, Investigation of multi-processor system with rollback function, *Technical Report of IEICE*, **FTS97-20**, pp.27-33, (1997).

2. K.M. Chandy and Lamport, Distributed snapshots : Determining global states of distributed systems, *ACM Trans. Comput. Syst.*, **3, No.1**, pp.63-75, (1985).

3. R. Koo and S. Toueg, Checkpointing and rollback-recovery for distributed systems, *IEEE Trans. Software Eng.*, **SE-13, No.1**, pp.23-31, (1987).

4. R.E. Strom and S.A. Yemini, Optimistic recovery in distributed systems, *ACM Trans. Comput. Syst.*, **3, No.3**, pp.204-226, (1985).

5. S. Osaki, *Applied Stochastic System Modeling*, Springer-Verlag, (1992).

6. M. Kimura, K. Yasui, T. Nakagawa and N. Ishii, Optimal Checkpointing Interval of a Communication System with Rollback Recovery, *Proceedings of The First Western Pacific and Third Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp.295-304,(1999).


## 6. APPENDIX

### 1. Derivation of mean time $\ell_{S_0,S_T}$

Using the mass functions of Markov renewal processes [5], Laplace-Stieltjes (LS) transforms $q_{ij}(s)$ of the transition probabilities $Q_{ij}(t)$ from state $i(i = S_0)$ to state $j(j = S_F, S_T, S_K)$ are given by the following equations :

$$q_{S_0,S_F}(s) \equiv \int_0^T e^{-st}\overline{F}(t)dX(t) + \int_0^T e^{-st}\overline{X}(t)dF(t), \tag{A.1}$$

$$q_{S_0,S_K}(s) \equiv [q_{S_0,S_F}(s)g(s)]^k, \tag{A.2}$$

$$q_{S_0,S_T}(s) \equiv \sum_{i=1}^{k}[q_{S_0,S_F}(s)g(s)]^{i-1}e^{-sT}\overline{X}(T)\overline{F}(T), \tag{A.3}$$

where

$$W_i(t) \equiv A(t) * [pa(t) + (1-p)pa^{(2)}(t)]^{(i)} \quad (i = 1, 2, \cdots, n),$$

$$X(t) \equiv \sum_{j=0}^{\infty} W_n^{(j)}(t) \sum_{i=1}^{n} W_{i-1}(t) * [(1-p)a(t)]^{(2)},$$

and $\Phi^{(j)}(t)$ is the $j$-fold convolution of $\Phi(t)$ with itself and $\Phi^{(0)}(t) \equiv 1$ for $t \geq 0$. Then, LS transforms $h_{S_0,S_T}(s)$ which is the mean time from the beginning of the operation to the next checkpoint is

$$h_{S_0,S_T}(s) \equiv \frac{q_{S_0,S_T}(s)}{1 - q_{S_0,S_K}(s)v(s)}. \tag{A.4}$$

Therefore, the mean time $\ell_{S_0,S_T}$ is

$$\begin{aligned} \ell_{S_0,S_T} &\equiv \lim_{s\to 0} \frac{-dh_{S_0,S_T}(s)}{ds} \\ &= \frac{1}{\overline{F}(T)\overline{X}(T)} \left[ \int_0^T \overline{F}(t)\overline{X}(t)dt + \frac{1}{\mu}[1 - \overline{F}(T)\overline{X}(T)] \right] + \frac{v[1-\overline{F}(T)\overline{X}(T)]^k}{1-[1-\overline{F}(T)\overline{X}(T)]^k}. \end{aligned} \tag{A.5}$$

## 2. Analysis of $M_F$ and $M_K$

LS transforms $\tilde{m}_F(s)$ and $\tilde{m}_K(s)$ of the expected number of rollbacks caused by processor failures and communication errors or the expected number of interruptions are, respectively

$$\tilde{m}_F(s) = \frac{\sum_{j=1}^{k}(j-1)[q_{S_0,S_F}(s)]^{j-1}e^{-sT}\overline{F}(T)\overline{X}(T) + kq_{S_0,S_K}(s)}{1 - q_{S_0,S_K}(s)v(s)}, \tag{A.6}$$

$$\tilde{m}_K(s) = \frac{q_{S_0,S_K}(s)}{1 - q_{S_0,S_K}(s)v(s)}. \tag{A.7}$$

Therefore, the expected number of rollback recovery $M_F$ and interruption $M_K$ per unit of time are, respectively

$$M_F \equiv \lim_{s\to 0} \tilde{m}_F(s) = \frac{1}{\overline{F}(T)\overline{X}(T)} - 1, \tag{A.8}$$

$$M_K \equiv \lim_{s\to 0} \tilde{m}_K(s) = \frac{1}{1-[1-\overline{F}(T)\overline{X}(T)]^k} - 1. \tag{A.9}$$

## 3. Analysis of $T_k^*$ which satisfy Eq.(6)

Let $L_k(T)$ be the left-hand side of Eq.(6). First, when $k = 1$, we have

$$L_1(0) = 0, \tag{A.10}$$

$$L_1(\infty) = \frac{\left[ [\gamma(\infty)\int_0^\infty \overline{Y}(t)dt - 1]\left(1 + \frac{c_3}{c_2-c_1}\right) - \gamma(\infty)v\left(\frac{c_1}{c_2-c_1}\right) \right]}{1 + \frac{1}{\mu}\lambda(\infty)}, \tag{A.11}$$

$$L_1'(T) = \frac{\gamma'(T)\left\{\left(1+\frac{c_3}{c_2-c_1}\right)\left[\int_0^T \overline{Y}(t)dt + \frac{1}{\mu}Y(T)\right] - \left(\frac{c_1}{c_2-c_1}\right)v\right\}}{\left[1+\frac{1}{\mu}[\lambda(T)+\alpha(1-x)]\right]^2}. \tag{A.12}$$

where $\Phi'(t)$ is a density of $\Phi(t)$. Then, taking $L_1'(T) = 0$, we have,

$$\int_0^T \overline{Y}(t)dt + \frac{1}{\mu}Y(T) = \frac{c_1 v}{c_3 + c_2 - c_1}, \tag{A.13}$$

which is strictly increasing in $T$. Thus, if $\int_0^\infty \overline{Y}(t)dt + 1/\mu > c_1 v/(c_3 + c_2 - c_1)$, there exists a finite and unique $T_1 (0 < T_1 < \infty)$ which satisfies Eq.(A.13). Hence, when $T > T_1$, $L_1(T)$ is increasing in $T$.

Therefore, we have the following policy:

(i) If $\gamma(t)$ is strictly increasing in t, $\int_0^\infty \overline{Y}(t)dt + 1/\mu > c_1 v/(c_3+c_2-c_1)$ and $L_1(\infty) > c_1/(c_2-c_1)$, there exists a finite and unique $T_1^* (T_1 < T_1^* < \infty)$ which satisfies $L_1(T) = c_1/(c_2 - c_1)$.

Next, when $k = \infty$, we have, from [6], if $\gamma(t)$ is strictly increasing in t and $L_\infty(\infty) > c_1/(c_2-c_1)$, there exists a finite and unique $T_\infty^*$ which satisfy $L_\infty(T) = c_1/(c_2 - c_1)$.

Further, we consider the case that $L_1(T) > L_\infty(T)$, i.e.,

$$\int_0^T \overline{Y}(t)dt - \frac{Y(T)}{\gamma(T)} > \frac{c_1}{c_3}v.$$

Letting $Q(T) \equiv \int_0^T \overline{Y}(t)dt - \frac{Y(T)}{\gamma(T)}$, we have

$$Q(0) = 0, \tag{A.14}$$

$$Q'(T) = \frac{Y(T)\gamma'(T)}{[\gamma(T)]^2}. \tag{A.15}$$

Then, if $\gamma(t)$ is strictly increasing in t, $Q(T)$ is also strictly increasing in $T$. We easily have, from $T_1$ which satisfies Eq.(A.13),

$$\begin{aligned} Q(T_1) &= \int_0^{T_1} \overline{Y}(t)dt - \frac{Y(T_1)}{\gamma(T_1)} \\ &= \frac{c_1 v}{c_3 + c_2 - c_1} - \frac{1}{\mu}Y(T_1) - \frac{Y(T_1)}{\gamma(T_1)}, \end{aligned} \tag{A.16}$$

and

$$Q(T_1) - \frac{c_1 v}{c_3} = -\frac{c_1 v(c_2 - c_1)}{c_3(c_3 + c_2 - c_1)} - \frac{1}{\mu}Y(T_1) - \frac{Y(T_1)}{\gamma(T_1)} < 0. \tag{A.17}$$

Therefore, we have the following result:

(ii) If $\int_0^\infty \overline{Y}(t)dt - 1/\gamma(\infty) > c_1 v/c_3$, then there exists $T_2 (T_1 < T_2 < \infty)$ which satisfies $\int_0^T \overline{Y}(t)dt - Y(T)/\gamma(T) = c_1 v/c_3$, and we have that $L_1(T) > L_\infty(T)$ for $T > T_2$.

It would be very difficult to show that $L_k(t)$ is strictly decreasing in $k$. We could show that if $L_k(t)$ is strictly decreasing in $k$, there exists a finite and unique $T_k^*$ which satisfies $L_k(T) = c_1/(c_2 - c_1)$, and $T_1^* \le T_k^* \le T_\infty^* (k = 2, 3, \cdots)$.

# AN OPTIMAL JOIN POLICY TO THE QUEUE IN PROCESSING TWO KINDS OF JOBS

JUNJI KOYANAGI and HAJIME KAWAI
Department of Social Systems Engineering, Faculty of Engineering, Tottori University
Koyama Minami 4-101, Tottori, 680-8552, Japan
junji@sse.tottori-u.ac.jp / kawai@sse.tottori-u.ac.jp

**Abstract**—We deal with a situation where a worker processes two kinds of jobs. Job A (JA) can be processed only at a queueing system. He must join the queue when he decides to process JA and stays there until he completes JA. Job B (JB) is completed after processing several steps and each step needs a constant time. At the end of each step, he can know whether JB is completed or not and decides whether he joins the queue. If JB is completed, he joins the queue to process JA. If he decides to join the queue, he processes the rest of JB after JA. The objective is to minimize the expected time until two jobs are completed. We prove a monotone property of the optimal policy by a dynamic programming formulation.

**Keywords**—Optimal join, Dynamic programming, Monotone policy.

## 1. INTRODUCTION

In queueing theory, a customer is usually assumed to arrive at the system without his own policy. However, the customers sometimes decide whether to join the queue or not. Typically, the decision depends on the waiting cost and service merit. Naor[1] proposed the system in which the customer decides whether to join the queue. The decision is made on the basis of waiting cost, service merit and toll to enter the system. He showed that the admission control by the toll yields the better performance of the system. Bell and Stidham[2] dealt with a static control in a multifacility model, which assigns the arriving customers to the multiple servers with determined probabilities. It is shown that the socially optimal control uses more servers than the individually optimal control. In the shortest queue problem (Winston[3]), the behavior of joining the shortest queue, which is the individually optimal control, is also the socially optimal control.

In this paper we deal with the model, in which one customer can decide whether to join the queue and discuss his (individually) optimal policy. His objective is the minimization of the expected time for processing two jobs. One job (job A) is processed in a queueing system and the other job (job B) can be processed if he is not in the queueing system. To complete job B, several steps must be processed. At each end of step he can decide whether to join the queue. If he decides to join the queue, he resumes job B after finishing job A. To minimize the total processing time, it is desirable to minimize the time in the queueing system.

This problem may be considered as a model which explains the behavior of a man, for example, in an amusement park. In an amusement park, there are many facilities that enjoy people. Some of them are very popular and they usually have long queues. In this case, we sometimes enjoy the other minor facilities and wait for the queue to be shorter. When we think the queue becomes short enough, we join the queue. Though the objective and the situation in an amusement park are very complicated, our model can be considered as a primitive model of this situation.

Our problem is formulated as a dynamic programming problem (Ross[4]). It is shown that the optimal policy has a monotone property. The monotone property is similar to the switch curve structure introduced in Warland[5].

In the next section, we describe our model. In Section 3, the formulation and the analysis are shown. In the last section, we supply the numerical examples to confirm our results.

# 2. MODEL DESCRIPTION

We consider a worker who processes the two types of jobs. Type A job (JA) can be processed only in a queueing system. Type B job (JB) are processed while he is not in the queueing system. The queueing system where JA is processed has Poisson arrivals and an exponential server. The arrival rate may depend on the queue length $i$ and it is denoted by $\lambda_i$. The service rate is denoted by $\mu$. There is only one JA and one JB. To complete JB, we need to process several steps whose number is distributed with distribution $R_k$. The distribution $R_k$ denotes the probability that more than or equal to $k$ steps are needed to complete JB. Each step needs the constant time $T$ and after each step he can know whether there are more steps to complete JB. If JB is completed, he joins the queue and waits until JA is finished. If the steps are still left, he decides whether to join the queue or to continue JB, with the information of the queue length and the number of steps he has processed. If he chooses to join the queue, he waits for JA to be finished and resume JB. If he chooses to process JB, he will decide again after the step. Our objective is to minimize the total expected processing time of two jobs.

## 2.1 FORMULATION BY DYNAMIC PROGRAMMING

Let us define the following notation for optimality equation.

$(i,k)$ : State $(i,k)$ indicates that the queue length is $i$, and JB has not been completed after $k$ steps are finished.

$V(i,k)$ : $V(i,k)$ is the optimal expected time for state $(i,k)$.

$W(i,k)$ : $W(i,k)$ is the expected time for choosing to continue JB at state $(i,k)$ and optimal behavior thereafter.

$D(i,k)$ : $D(i,k)$ is the optimal action for state $(i,k)$.

$$D(i,k) = \begin{cases} 1 & \text{if it is optimal to join the queue,} \\ 2 & \text{if it is optimal to continue JB.} \end{cases}$$

$S_k$ : The conditional probability that the total number of the steps is $k$, given the total number of the steps is more than $k$. ($\overline{S}_k \equiv 1 - S_k$)

$$S_k = (R_k - R_{k+1})/R_k$$

$M_k$ : Expected residual time to complete JB for state $(i,k)$,

$$M_k = T \sum_{m=k+1}^{\infty} R_m/R_{k+1}.$$

$P_{ij}$ : The probability that the queue length changes to $j$ after time $T$, given the initial length $i$.

$Q_i$ : The expected queue length after time $T$, given the initial length $i$.

$$Q_i = \sum_{j=0}^{\infty} jP_{ij}$$

With these notation we obtain the following optimality equation.

$$W(i,k) = T + \sum_{j=0}^{\infty} P_{ij}\Big(S_{k+1}(j+1)/\mu + \overline{S}_{k+1}V(j,k+1)\Big) \tag{1}$$

$$V(i,k) = \min\{M_k + (i+1)/\mu,\ W(i,k)\} \tag{2}$$

Optimal action $D(i,k)$ is determined by

$$D(i,k) = \begin{cases} 1 & \text{if } M_k + (i+1)/\mu \le W(i,k) \\ 2 & \text{if } M_k + (i+1)/\mu > W(i,k) \end{cases}$$

We assume the following conditions for our model.

**Condition 1.** *The arrival rate $\lambda_i$ and the probability $S_k$ satisfy the following conditions.*

1. *$\lambda_i$ is a decreasing function of $i$.*
2. *$S_k$ is a monotone function of $k$ and there exists $N$ and $\varepsilon > 0$ such that $S_k > \varepsilon$ for all $k > N$.*

**Lemma 1.** *The transition probability of the queue length has the following properties.*

1. *For all $m$, $\sum_{j=m}^{\infty} P_{ij}$ is increasing in $i$.*
2. *The inequality $Q_{i+1} - Q_i \le 1$ holds.*

This lemma is easily derived by Condition 1.1.

The value of $V(i,k)$ is obtained by the following iteration (successive approximation, Wessels[6]).

$$V^0(i,k) \equiv 0 \text{ for all } i,k \tag{3}$$

$$W^{n+1}(i,k) = T + \sum_{j=0}^{\infty} P_{ij}\Big(S_{k+1}(j+1)/\mu + \overline{S}_{k+1}V^n(j,k+1)\Big) \tag{4}$$

$$V^{n+1}(i,k) = \min\{M_k + (i+1)/\mu, W^n(i,k)\} \tag{5}$$

We prove some properties of $V(i,k)$ and $W(i,k)$ by mathematical induction with respect to $n$.

**Lemma 2.** *The functions $V(i,k)$ and $W(i,k)$ have the following properties.*

1. *If $S_k$ is increasing in $k$,*

$$V(i,k+1) - V(i,k) \ge M_{k+1} - M_k \text{ and} \tag{6}$$

$$W(i,k+1) - W(i,k) \ge M_{k+1} - M_k. \tag{7}$$

2. *If $S_k$ is decreasing in $k$,*

$$V(i,k+1) - V(i,k) \leq M_{k+1} - M_k \text{ and} \tag{8}$$

$$W(i,k+1) - W(i,k) \leq M_{k+1} - M_k. \tag{9}$$

**Proof.**
We prove the case that $S_k$ is increasing. When $S_k$ is decreasing, the proof is similar, therefore it is omitted.

As the first step of the induction, the inequality obviously holds for $V^0(i,k)$. Then, we show that if $V^n(i,k+1) - V^n(i,k) \geq M_{k+1} - M_k$, then $W^{n+1}(i,k+1) - W^{n+1}i,k) \geq M_{k+1} - M_k$.

$$\begin{aligned}
& W^{n+1}(i,k+1) - W^{n+1}(i,k) \\
&= \sum_{j=0}^{\infty} P_{ij}\Big[(S_{k+2} - S_{k+1})(j+1)/\mu \\
&\quad + \overline{S}_{k+2}V^n(j,k+2) - \overline{S}_{k+1}V^n(j,k+1)\Big] \\
&\geq \sum_{j=0}^{\infty} P_{ij}\Big[(S_{k+2} - S_{k+1})(j+1)/\mu \\
&\quad + \overline{S}_{k+2}\{V^n(j,k+1) + M_{k+2} - M_{k+1}\} - \overline{S}_{k+1}V^n(j,k+1)\Big] \\
&= \sum_{j=0}^{\infty} P_{ij}\Big[(S_{k+2} - S_{k+1})\{(j+1)/\mu - V^n(j,k+1)\} \\
&\quad + \overline{S}_{k+2}(M_{k+2} - M_{k+1})\Big] \\
&\geq \sum_{j=0}^{\infty} P_{ij}\Big[(S_{k+2} - S_{k+1})(-M_{k+1}) + \overline{S}_{k+2}(M_{k+2} - M_{k+1})\Big] \\
&= \sum_{j=0}^{\infty} P_{ij}\Big(\overline{S}_{k+2}M_{k+2} - \overline{S}_{k+1}M_{k+1}\Big) \\
&= M_{k+1} - M_k
\end{aligned}$$

$V^{n+1}(i,k+1) - V^{n+1}(i,k) \geq M_{k+1} - M_k$ is obvious by $\min\{x,y\} - \min\{a,b\} \geq \min\{x-a, y-b\}$.
Since $V^n(i,k)$ and $W^n(i,k)$ converges to $V(i,k)$ and $W(i,k)$ respectively, Lemma 2 holds. □

By Lemma 2, the following theorem holds.

**Theorem 1.** *The optimal policy has the following properties.*

1. *If $S_k$ is increasing and $D(i,k) = 1$ for some state $(i,k)$, then $D(i,l) = 1$ for $l \geq k$.*
2. *If $S_k$ is decreasing and $D(i,k) = 2$ for some state $(i,k)$, then $D(i,l) = 2$ for $l \geq k$.*

**Proof.**
We prove the theorem when $S_k$ is increasing.
$D(i,k) = 1$ indicates $M_k + (i+1)/\mu \leq W(i,k)$, then

$$M_k + (i+1)/\mu \leq W(i,k) \leq W(i,k+1) + M_k - M_{k+1}.$$

Therefore, $M_{k+1} + (i+1)/\mu \leq W(i,k+1)$ holds and it implies $D(i,k+1) = 1$. Repeating this argument, we obtain Theorem 1.

The next lemma is also concerning to the optimal policy.

**Lemma 3.** *The functions $V(i,k)$ and $W(i,k)$ satisfy the following inequalities.*

$$V(i+1,k) - V(i,k) \leq 1/\mu \tag{10}$$

$$W(i+1,k) - W(i,k) \leq 1/\mu \tag{11}$$

**Proof.**
For $n = 0$, the result obviously holds. We show that $V^n(i+1,k) - V^n(i,k) \leq 1/\mu$ implies $W^{n+1}(i+1,k) - W^{n+1}(i,k) \leq 1/\mu$. Here, let us define $\delta(j,k)$ by

$$\delta(j,k) = \begin{cases} S_{k+1}/\mu + \overline{S}_{k+1} V^n(0,k) & (j=0) \\ S_{k+1}/\mu + \overline{S}_{k+1}(V^n(j,k) - V^n(j-1,k)) & (j \geq 1) \end{cases} \tag{12}$$

Note that $\delta(j,k) \leq 1/\mu$ for $j \geq 1$. Then

$$\begin{aligned}
& W^{n+1}(i+1,k) - W^{n+1}(i,k) \\
&= \sum_{j=0}^{\infty} P_{i+1\,j}\Big(S_{k+1}(j+1)/\mu + \overline{S}_{k+1} V^n(j,k)\Big) \\
&\quad - \sum_{j=0}^{\infty} P_{ij}\Big\{S_{k+1}(j+1)/\mu + \overline{S}_{k+1} V^n(j,k)\Big\} \\
&= \sum_{j=0}^{\infty} P_{i+1\,j} \sum_{m=0}^{j} \delta(m,k) - \sum_{j=0}^{\infty} P_{ij} \sum_{m=0}^{j} \delta(m,k) \\
&= \sum_{m=0}^{\infty} \delta(m,k) \sum_{j=m}^{\infty} P_{i+1\,j} - \sum_{m=0}^{\infty} \delta(m,k) \sum_{j=m}^{\infty} P_{ij} \\
&= \sum_{m=1}^{\infty} \delta(m,k)\Big(\sum_{j=m}^{\infty} P_{i+1\,j} - \sum_{j=m}^{\infty} P_{ij}\Big) \\
&\leq \sum_{m=1}^{\infty} 1/\mu\Big(\sum_{j=m}^{\infty} P_{i+1\,j} - \sum_{j=m}^{\infty} P_{ij}\Big) \\
&= 1/\mu(Q_{i+1} - Q_i) \\
&\leq 1/\mu
\end{aligned}$$

The first inequality holds by $\delta(m,k) \leq 1/\mu$ $(m \geq 1)$ (by inductive assumption) and Lemma 1.1 and the last inequality holds by Lemma 1.2. $V^{n+1}(i+1,k) - V^{n+1}(i,k) \leq 1/\mu$ is obvious by $\min\{x,y\} - \min\{a,b\} \leq \max\{x-a, y-b\}$.

By Lemma 3, we have the next theorem.

**Theorem 2.** *If* $D(i,k) = 2$, *then* $D(j,k) = 2$ $(i \leq j)$.

**Proof.**
$D(i,k) = 2$ indicates $M_k + (i+1)/\mu \geq W(i,k)$, then

$$M_k + (i+1)/\mu \geq W(i,k) \geq W(i+1,k) - 1/\mu.$$

Therefore, $M_k + (i+2)/\mu \geq W(i+1,k)$ holds, which implies $D(i+1,k) = 2$. Repeating this argument, we obtain Theorem 2.

By Theorem 1 and Theorem 2, the changes of optimal action happen at most once, as $i$ or $k$ increases. Thus optimal policy has the following monotone structure.

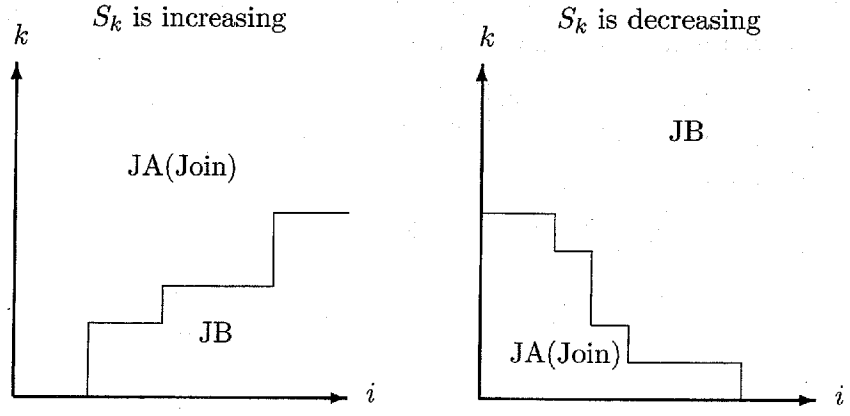


Figure 1: Monotone Property of Optimal Policy

## 3. NUMERICAL EXAMPLE

In this section we show numerical examples.

First, we show the case that $S_k$ is increasing.

1. The service and arrival rates of the queueing systems are $\mu = 1.8$,
   $\lambda_0 = 1.9$, $\lambda_1 = 1.9$, $\lambda_2 = 1.9$, $\lambda_3 = 1.9$, $\lambda_4 = 1.8$,
   $\lambda_5 = 1.5$, $\lambda_6 = 1.5$, $\lambda_7 = 1.0$, $\lambda_8 = 1.0$, $\lambda_9 = 1.0$, $\lambda_i = 0$ $(i \geq 10)$.

2. The processing time and distribution of step of JB are $T = 1$,
   $S_0 = 0$, $S_1 = 0$, $S_2 = 0.1$, $S_3 = 0.1$, $S_4 = 0.2$,
   $S_5 = 0.5$, $S_6 = 0.5$, $S_7 = 0.5$, $S_8 = 0.5$, $S_9 = 0.6$,
   $S_k = 0.9$ $(k \geq 10)$

With these values, the optimal policy is shown in Table 1.

Next, we show the case that $S_k$ is decreasing.

1. The service and arrival rates of the queueing systems are $\mu = 1.8$, $T = 1$,
   $\lambda_0 = 1.9$, $\lambda_1 = 1.9$, $\lambda_2 = 1.9$, $\lambda_3 = 1.9$, $\lambda_4 = 1.8$,
   $\lambda_5 = 1.5$, $\lambda_6 = 1.5$, $\lambda_7 = 1.0$, $\lambda_8 = 1.0$, $\lambda_9 = 1.0$, $\lambda_i = 0$ $(i \geq 10)$.

Table 1: Optimal policy for increasing $S_k$

| $k$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 10 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 5 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 0 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0 | | | | | 5 | | | | | $i$ |

2. The processing time and distribution of step of JB are $T = 1$,
   $S_0 = 0.7$, $S_1 = 0.7$, $S_2 = 0.7$, $S_3 = 0.7$, $S_4 = 0.5$,
   $S_5 = 0.5$, $S_6 = 0.5$, $S_7 = 0.5$, $S_8 = 0.5$, $S_9 = 0.2$,
   $S_{10} = 0.2$, $S_{11} = 0.2$, $S_{12} = 0.2$, $S_k = 0.1$ $(k \geq 13)$

With these values, the optimal policy is shown in Table 2.

Table 2: Optimal policy for decreasing $S_k$

| $k$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 10 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 5 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0 | | | | | 5 | | | | | $i$ |

# REFERENCES


1. P. Naor, On the regulation of queue size by levying tolls, *Econometrica*, **37**, 15–24 (1969).

2. C.E. Bell and S. Stidham, Individual and social optimization in the allocation of customers to alternative servers, *Management Science*, **29**, 831–839 (1983).

3. W. Winston, Optimality of the shortest line discipline, *Journal of Applied Probability*, **14**, 181–189 (1977).

4. S. M. Ross, *Applied Stochastic Models with Optimization Applications*, Holden-Day, San Francisco (1970).

5. J. Walrand, *An Introduction to Queueing Networks*, Prentice Hall, New Jersey (1988).

6. J. Wessels, Markov Programming by Successive Approximations with Respect to Weighted Supremum Norms, *Journal of Mathematical Analysis and Applications*, **58**, 326–335 (1977).

# AN OPTIMAL MAINTENANCE TIME OF AUTOMATIC MONITORING SYSTEM OF ATM WITH TWO KINDS OF BREAKDOWNS

SYOUJI NAKAMURA[1], CUNHUA QIAN[2], ITUKI HAYASHI[3] and TOSHIO NAKAGAWA[4]

[1]Systems Division, The Bank of Nagoya, Ltd.
1-chome-501 Kounosu, Tennpaku-ku, Nagoya 468-0003, Japan
[2]Department of Industrial Engineering, Aichi Institute of Technology
1247 Yachigusa, Yakusa-cho, Toyota 470-0392, Japan
[3] Hardware Application Engineering Division, Hitachi Chubu Software, Ltd.
[4]Department of Industrial Engineering, Aichi Institute of Technology
1247 Yachigusa, Yakusa-cho, Toyota 470-0392, Japan

pfa00744@nifty.ne.jp / qch64317@ie.aitech.ac.jp / hayashi@ts.hitachi.co.jp / nakagawa@ie.aitech.ac.jp

**Abstract**—All automatic tellers machines (ATM) in a bank make an unmanned driving on a weekend and holidays, and an automatic monitoring system continuously watches the operation of ATM through the telecommunication network. There are two kinds of troubles according to the installed places of ATM. : One is the trouble which occurs inside the branch of a bank where ATM make a manned driving except a weekend and holidays, and the other is the one which occurs outside the branch where ATM always make an unmanned driving. Two kinds of breakdowns are introduced, and the expected cost for an unmanned driving period is obtained. A maintenance policy which minimizes the expected cost is analytically derived. Finally, a numerical example is given and some useful discussions are made.

**Keywords**—ATM of bank, Two breakdowns, Expected cost, Maintenance policy.

## 1. INTRODUCTION

Most automatic tellers machines (ATMs) are connected with the online system of a bank and improve the efficiency of business about since 1975. The operational times of ATMs are greatly increased with the driving on a weekend and holidays in recent years. Further, ATMs have various functions such as the transfer of cash, the contract and cancellation of deposit and account, the reception of loan, and so on. Moreover, ATMs are now planning to connect with other organizations, and so, their networks are expanded on every place and become an indispensable infrastructure in a daily life. In such situations, it is very important to consider an automatic monitoring system of ATMs, because adequate maintenance for troubles and breakdowns have to be promptly done from the viewpoints of trust and customer's service.

A bank consigns the replenishment of cash, and the check and maintenance of ATMs to a guard company[1]. There are roughly two kinds of ATMs according to their installed places: One is an ATM which is set up in the branch of a bank, and the other is in department stores, stations, supermarkets or other public facilities, which is called the outside branch ATM.

An automatic monitoring system continuously watches the operation of outside branch ATMs because they always make an unmanned driving. However, the inside branch ATM is watched by a bank employee in the branch on weekdays, and is done at the control center on holidays. Further, a

bank employee checks an ATM at the beginning time of the next day after holidays. Even if some troubles have occurred in an ATM on holidays, they are removed and a bank employee restores it to a normal condition on the next day. At the control center, a monitoring system displays the state of troubles in the terminal unit and outputs them. Moreover, there might be phone calls to report the situation of troubles by users in an ATM. When troubles are displayed in the terminal unit, a watch member at the control center can remove some of them, by operating the terminal unit remotely according to their state. If a watch member cannot remove them, he reports this fact to a guard company. A guard member can remove promptly troubles or breakdowns of ATM.

It is assumed in this paper that there are two kinds of troubles, which might break down indirectly, and breakdown directly. This paper proposes a stochastic model with two kinds of breakdowns: An ATM is checked at time $t_0$ after trouble occurrence. When the distributions of two breakdowns, the checking cost and the loss cost due to breakdowns are introduced, the expected cost of the inside branch ATM for an unmanned driving period is obtained. An optimal maintenance policy, which minimizes the expected cost, is analytically derived. Finally, a numerical example is given and some useful discussions are made.

## 2. MODEL

An automatic monitoring system watches ATM by the polling selecting method through a telephone line, and displays the state of ATM. The state can be classified in the following four states:

**state 0:** ATM is normal. There is no trouble in ATM.

**state 1:** Some troubles occur in ATM. There is a possibility that it will break down soon. For example, it is warning that the cash and the receipt are running out soon, or ATM is choked up with the cash and the card. If a watch member at the control center removes troubles, they are not included in state 1.

**state 2:** ATM is checked at time $t_0$ after trouble occurrences in state 1. A guard member goes to the ATM place and removes troubles before it breaks down. This is an easy work, which changes the cashbox or replenishes the receipt and the journal form.

**state 3:** ATM breaks down until time $t_0$ after trouble occurrences(breakdown 1), i.e., it breaks down before a guard member arrives at the ATM place. He recovers the breakdown by changing the cashbox or replenishing the receipt and the journal form.

**state 4:** ATM breaks down by mechanical factors(breakdown 2). For example, the power supply stops or ATM is choked up with the cash and the card. A guard member goes to the ATM place and recovers the breakdown. Therefore, ATM cannot be used from the breakdown to the arrival time of a guard member. The maintenance time of breakdown 2 is usually longer than that of breakdown 1 in state 3.

Figure 1 shows the transition relation between above states.

In the operation of ATM, troubles associated with the cash, the receipt form and the journal form would occur at most one time for a short time span such as a weekend and holidays. It is supposed that an ATM has to operate during the interval $[0, T]$ and the trouble occurs only at most one time in this interval.

It is assumed that troubles occur according to a general distribution $F_0(t)$, and after trouble occurrences, the time to breakdown 1 has a general distribution $F_1(t)$. Further, the time to breakdown 2 is independent of the occurrences of troubles and breakdown 1, and has a general distribution

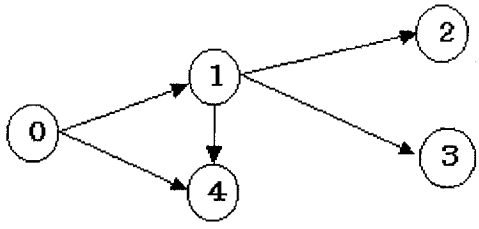


Figure 1: Figure of state of transition.

$F_2(t)$. If there are two or more ATMs in the same booth, four states are defined as the state of the last operating ATM.

We give the following probabilities that events such as troubles and breakdowns occur during $[0, T]$, where $\bar{F}_i \equiv 1 - F_i \quad (i = 0, 1, 2)$.

**(i)** The probability that troubles and breakdown 2 do not occur during (0,T ] is

$$\bar{F}_0(T)\bar{F}_2(T). \tag{1}$$

**(ii)** The probability that breakdown 2 occurs before trouble occurrence during (0,T ] is

$$\int_0^T \bar{F}_0(x)dF_2(x). \tag{2}$$

**(iii)** The probability that ATM is checked at T without breakdowns after trouble occurrence is

$$\bar{F}_2(T)\int_{T-t_0}^{T} \bar{F}_1(T-x)dF_0(x). \tag{3}$$

**(iv)** The probability that breakdown 1 occurs after trouble occurrence (see Figure 2) is

$$\int_{T-t_0}^{T} dF_0(x)\int_0^{T-x} \bar{F}_2(x+y)dF_1(y). \tag{4}$$

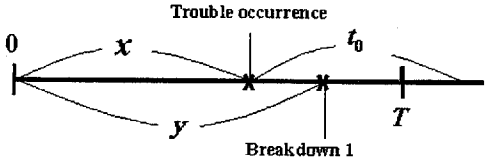


Figure 2: Breakdown 1 occurrence

**(v)** The probability that breakdown 2 occurs after trouble occurrence is

$$\int_{t-t_0}^{T} dF_0(x) \int_x^T \bar{F}_1(y-x) dF_2(y). \tag{5}$$

**(vi)** The probability that ATM is checked at $t_0$ after trouble occurrence is

$$\bar{F}_1(t_0) \int_0^{T-t_0} \bar{F}_2(t_0+x) dF_0(x). \tag{6}$$

**(vii)** The probability that breakdown 1 occurs until time $t_0$ after trouble occurrence is

$$\int_0^{T-t_0} dF_0(x) \int_0^{t_0} \bar{F}_2(x+y) dF_1(y). \tag{7}$$

**(viii)** The probability that breakdown 2 occurs until time $t_0$ after trouble occurrence (see Figure 3) is

$$\int_0^{T-t_0} dF_0(x) \int_x^{x+t_0} \bar{F}_1(y-x) dF_2(y). \tag{8}$$

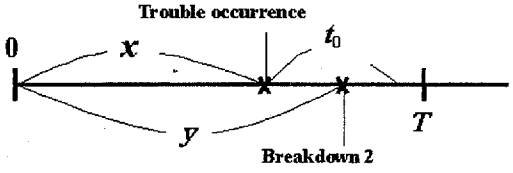


Figure 3: Breakdown 2 occurrence

Evidently, we have

(3)+(4)+(5)

$$= \int_{T-t_0}^{T} dF_0(x) \left[ \bar{F}_2(T) \bar{F}_1(T-x) + \int_0^{T-x} \bar{F}_2(x+y) dF_1(y) + \int_x^T \bar{F}_1(y-x) dF_2(y) \right]$$

$$= \int_{T-t_0}^{T} \bar{F}_2(x) dF_0(x), \tag{9}$$

(6)+(7)+(8)

$$= \int_0^{T-t_0} dF_0(x) \left[ \bar{F}_2(t_0+x) \bar{F}_1(t_0) + \int_0^{t_0} \bar{F}_2(x+y) dF_1(y) + \int_x^{x+t_0} \bar{F}_1(y-x) dF_2(y) \right]$$

$$= \int_0^{T-t_0} \bar{F}_2(x) dF_0(x). \tag{10}$$

Hence, it is proved that

$$(1)+(2)+(9)+(10) = \bar{F}_0(T)\bar{F}_2(T) + \int_0^T \bar{F}_0(x) dF_2(x) + \int_0^T \bar{F}_2(x) dF_0(x) = 1.$$

## 3. EXPECTED COST

We introduce the following costs:

$c_0$ = cost at $T$. An ATM stops at time $T$. A bank employee checks an ATM before it begins to operate on the next day, and replenishes the cash, the journal and receipt forms.

$c_1$ = checking cost at time $t_0$. A guard member refills up the cash cassette, and if necessary, replenishes the journal and receipt forms. A cost $c_1$ is higher than $c_0$ because a guard member specially has $t_0$ go to the ATM place.

$c_2$ = cost for breakdown 1. An ATM has stopped until a guard member arrives at time $t_0$ after breakdown 1 occurrence. Any customers cannot use it and have to use ATMs of other banks. In this case, not only customers pay the commission to other banks, but also a bank pays the commission for customers' usage. A cost $c_2$ includes the whole cost which is the sum of cost $c_1$ and the loss cost for breakdown 1.

$c_3$ = cost of breakdown 2. An ATM breaks down directly, and has stopped until a guard member arrives at the ATM place. The maintenance time and cost for breakdown 2 would be usually longer and higher than those of breakdown 1, respectively. It can be seen in general that $c_3 > c_2 > c_1 > c_0$.

The total expected cost of ATM during $[0, T]$ is given by

$$\begin{aligned} C(t_0) &= c_0\bar{F}_2(T)\left[\bar{F}_0(T) + \int_{T-t_0}^{T} \bar{F}_1(T-x)dF_0(x)\right] + c_1\bar{F}_1(t_0)\int_0^{T-t_0} \bar{F}_2(t_0+x)dF_0(x) \\ &\quad + c_2\left[\int_{T-t_0}^{T} dF_0(x)\int_0^{T-x} \bar{F}_2(x+y)dF_1(y) + \int_0^{T-t_0} dF_0(x)\int_0^{t_0} \bar{F}_2(x+y)dF_1(y)\right] \\ &\quad + c_3\left[\int_0^{T} \bar{F}_0(x)dF_2(x) + \int_{T-t_0}^{T} dF_0(x)\int_x^{T} \bar{F}_1(y-x)dF_2(y)\right. \\ &\quad \left. + \int_0^{T-t_0} dF_0(x)\int_x^{x+t_0} \bar{F}_1(y-x)dF_2(y)\right] \qquad (0 \le t_0 \le T). \end{aligned} \tag{11}$$

## 4. OPTIMAL POLICY

It is a problem to determine when a guard member goes to the ATM place after trouble occurrence. For example, if troubles occur near at time $T$, it would be unnecessary to send a guard member. We find an optimal time $t_0^*$ $(0 \le t_0^* \le T)$ which minimizes the expected cost $C(t_0)$ in (11). In particular case of $t_0 = 0$, i.e., when an ATM is maintained immediately after trouble occurrences, the expected cost is

$$C(0) = c_0\bar{F}_2(T)\bar{F}_0(T) + c_1\int_0^{T} \bar{F}_2(x)dF_0(x) + c_3\int_0^{T} \bar{F}_0(x)dF_2(x). \tag{12}$$

In particular case of $t_0 = T$, i.e., when an ATM is not maintained until time T even if troubles

occur, the expected cost is

$$C(T) = c_0\bar{F}_2(T)\left[\bar{F}_0(T) + \int_0^T \bar{F}_1(T-x)dF_0(x)\right]$$

$$+c_2\int_0^T dF_0(x)\int_0^{T-x}\bar{F}_2(x+y)dF_1(y)$$

$$+c_3\left[\int_0^T \bar{F}_0(x)dF_2(x) + \int_0^T dF_0(x)\int_x^T \bar{F}_1(y-x)dF_2(y)\right]. \tag{13}$$

Next, suppose that distributions $F_0(t)$ and $F_2(t)$ are exponential, i.e., $F_0(t) = 1 - e^{-\lambda_0 t}$ and $F_2(t) = 1-e^{-\lambda_2 t}$. Further, assume that $F_1(t)$ has a density $f_1(t)$, and define that $\gamma_1(t) \equiv f_1(t)/\bar{F}_1(t)$ with $\gamma_1(0) \equiv 0$ which represents the failure rate of breakdown 1. Then, differentiating $C(t_0)$ with respect to $t_0$ and setting it equal to zero, we have

$$\left[(c_2 - c_1)\gamma_1(t_0) + (c_3 - c_1)\lambda_2\right]\frac{e^{(\lambda_0+\lambda_2)(T-t_0)} - 1}{\lambda_0 + \lambda_2} = c_1 - c_0. \tag{14}$$

In general, it would be very difficult to derive an optimal time $t_0^*$ analytically.

## 5. NUMERICAL EXAMPLE

Suppose that the distribution $F_1(t)$ of time to breakdown 1 has the IFR property [2],i.e., $F_1(t) = 1-e^{-\lambda_1 t^m}$ $(m > 1)$. Figure 4 draws the expected cost $C(t_0)$ for $t_0$ when $T = 16$(hours), $\lambda_0 = 5/1000$ (1/hours), $\lambda_1 = 7/200$ (1/hours), $\lambda_2 = 5/200$ (1/hours), $c_0 = 4.5$, $c_1 = 6.0$, $c_2 = 7.0$, $c_3 = 8.5$. It is shown from this figure that $t_0^* = 1.00$ (hours) and $C(t_0^*) = 4.745$. We have to dispatch a guard member after 60 minutes from trouble occurrence, and he make the maintenance of an ATM. In actual operations, a guard member usually goes to the branch of ATMs from about 20 minutes to 60 minutes even if one of them in the booth breaks down, and sequentially makes the maintenance of ATMs with troubles. The above model, where a guard member arrives there at 60 minutes after trouble occurrence, would be suitable for the above real situations.

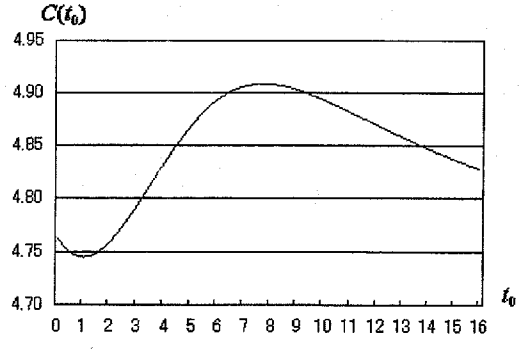


Figure 4: Graph of total expected cost.

## 6. CONCLUSIONS

We form a stochastic model of an automatic monitoring system for an ATM in a bank: Assuming the occurrences of two breakdowns where one occurs after some troubles and the other occurs directly, we obtain the expected cost during an unmanned period. Further, we discuss numerically an optimal time $t_0^*$ which is the checking time of an ATM after trouble occurrences. This maintenance policy woul be applied to an actual monitoring system by suitable modifications.

## REFERENCES


1. S. Nakamura, H. Sandoh and T. Nakagawa, Optimal number of cashboxes for an unmanned ATM, *J. of Operational Research Soc. of Japan,* **42**, 663-666 (1998).

2. R. E. Barlow and F. Proschan, *Mathematical Theory of Reliability,* John Wiley & Sons, New York (1965).

# A STRUCTURAL APPROXIMATION METHOD TO GENERATE THE OPTIMAL AUTO-SLEEP SCHEDULE FOR A COMPUTER SYSTEM

HIROYUKI OKAMURA, TADASHI DOHI and SHUNJI OSAKI

Department of Industrial and Systems Engineering, Hiroshima University,
4-1 Kagamiyama 1 Chome, Higashi-Hiroshima 739-8527, Japan.

{okamu, dohi, osaki}@gal.sys.hiroshima-u.ac.jp

**Abstract**—This paper addresses a problem of how to determine the optimal auto-sleep schedule when the computer user should turn the hard disk or the display to a sleep mode in order to save the electrical power after the computer has not been accessed. We propose a stochastic model to obtain the optimal sleep timing strategy which minimizes the expected electrical power consumed per unit time in the steady-state, where access requirements arrive at the system according to a renewal process and are processed by a general service time. Then the phase-type approximations are proposed to generate the optimal auto-sleep schedule approximately. We investigate the performance of the phase-type approximation through a simulation study.

**Keywords**—auto-sleep scheduling, power saving, renewal process, phase-type distribution, EM algorithm, approximation.

## 1. INTRODUCTION

Recently, the auto-sleep function of the hard disk or the display in a computer system is rapidly recognized to be important in terms of power management. In fact, the auto-sleep function is equipped in almost computer systems as a standard function. Then the optimal design for the auto-sleep function is the most important problem, in particular, for notebook computers with limited capacity of battery. For example, on the hard disk of a computer, the electrical power consumed to warm up from sleep mode is larger than that consumed in the normal operation. Thus, it is not always effective to design the system such that moves its state to the sleep mode whenever there is no access requirement.

First, the optimal design problem for the auto-sleep function was considered by Sandoh, Hirakoshi and Kawai [1]. Dohi, Kaio and Osaki [2] proposed a statistical non-parametric method to estimate the optimal sleep timing for the same problem. However, it is noted that the seminal works above simplified the underlying problem extremely and was incomplete for representation of stochastic behavior of the auto-sleep system. More valid formulations were made by Okamura, Dohi and Osaki [3, 4]. They considered two kinds of models (Type I model and Type II model) with and without cancellation of access requirements arrived at the system, respectively. More specifically, Type I model with cancellation assumes that other access requirements arrived at the system while one job has been processed are canceled, and focuses on the multi-use circumstance for a desktop computer unit. On the other hand, Type II model corresponds to a buffer system in which other access requirements are accumulated while one job has been processed, and deals with the multi-job system such as network printers. Okamura, Dohi and Osaki [3, 4] proved that the optimal sleep timing strategies for both models are *the switching strategies*, *i.e.*, turn always the system to a sleep

mode after the process for a job is completed, or do not at all, if the access requirements arrive according to the homogeneous Poisson process.

However, if the arrival of access requirements follows more general stochastic processes such as the renewal process, it is difficult to obtain the optimal sleep timing explicitly. Okamura, Dohi and Osaki [3, 4] applied the simple parametric approximation methods by Miyazawa [5] and the usual diffusion approximation to generate the optimal auto-sleep schedule, but could not obtain the satisfactory approximation performance. The main reason is that the arrival process may belong to a more wide class of stochastic processes. In this paper, we apply the phase-type approximations to generate the optimal auto-sleep timing which minimizes the expected power consumed per unit time in the steady-state for Type I model. Altiok [6] and Heijden [7] showed that the phase-type approximations are useful to represent the general probability distributions. Asmussen and Koole [8] also proved that the phase-type renewal process is weakly dense in the class of stationary simple point processes.

The paper is planed as follows. Section 2 describes the auto-sleep model under consideration and gives an implicit form of the expected power consumed per unit time in the steady-state under the assumption that the arrival of access requirements follows the renewal process. Section 3 concerns the approximation problem for the expected power consumed per unit time in the steady-state. Then, the phase-type approximation is introduced to represent the access requirements process. Furthermore, two statistical estimation methods with the phase-type approximation are developed. Section 4 is devoted to investigate the approximation performance for the proposed methods through a simulation study. Finally, the paper is concluded with some remarks.

# 2. MODEL DESCRIPTION

## 2.1 NOTATION AND ASSUMPTIONS

Suppose that the access requirements arrive at the system according to an ordinary renewal process $\{N(t);\ t > 0\}$. Denote a sequence of inter-arrival times between $(k-1)$-th and $k$-th arrivals by $\{X_k;\ k = 1, 2, \cdots\}$. Then, $X_k$ are the non-negative i.i.d. random variables, having the probability distribution $F(t)$ with mean $1/\lambda$ $(> 0)$ and variance $\sigma_a$ $(> 0)$. The tasks required by the $k$-th access are processed with the times $S_k$, which are the non-negative i.i.d. random variables having the probability distribution $H(t)$ with finite mean $1/\mu$ $(> 0)$ and variance $\sigma_s$ $(> 0)$. It is assumed that the system under consideration can take the following states;

**Busy:** The system processes some tasks required by accesses, where the set-up time $\tau$ $(> 0)$ is needed before processing each task. After the present task is completed, the state of system moves to the idle state. During the busy state, the electrical power consumed per unit time is $P_1$ $(> 0)$.

**Idle:** No access requirement occurs, after one task is completed. If a new access requirement occurs until the total spent time in the idle period becomes $t_0$, the system begins to process it after elapsing $\tau$ time units. Otherwise, the state of system moves to the sleep state at the moment when the total spent time in the idle period becomes $t_0$. Throughout this paper, we call $t_0$ *the auto-sleep time.* The electrical power consumed per unit time during the idle period is also $P_1$ $(> 0)$.

**Sleep:** The sleep state is the lower-power state, so that the electrical power consumed per unit time is less than that in the other states. To simplify the discussion, we assume that the electrical power consumed per unit time in the sleep state is zero. When an access requirement occurs, the sleep mode terminates immediately and the state of system moves to the warm-up state.

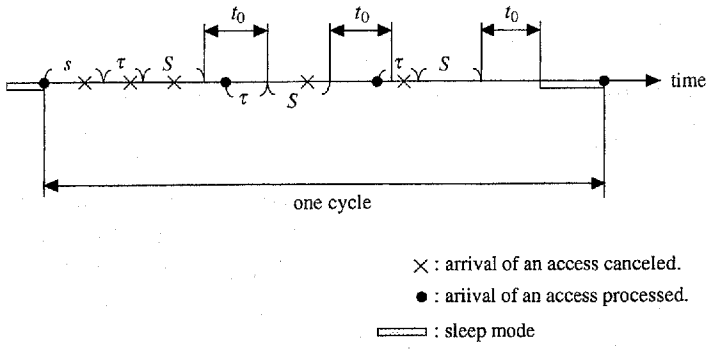


Figure 1: Possible realization of the stochastic system.

**Warm-up:** In order to begin processing a task from the sleep mode, $s$ ($> 0$) time units are needed for warming-up. Hence, after $s + \tau$ time units are elapsed, the process for the task is started. In the warm-up state, the electrical power $P_2$ ($> 0$) is consumed per unit time, where $P_2 > P_1$.

In this paper, we assume that the other access requirements are canceled while the system is busy. Hence, the state of system moves to an idle when each task is completed. Figure 1 is depicted the possible realization of the auto-sleep system.

## 2.2 FORMULATION OF THE EXPECTED POWER CUNSUMED PER UNIT TIME IN THE STEADY-STATE

Let us consider the expected power consumed per unit time in the steady-state as a criterion to evaluate the system performance. In this model, since the other access requirements arrive during the processing of the previous task, the time length of an idle period can be represented as the residual life of the arrival process. Define the residual life of the arrival process by $\gamma_t$ having the distribution function $I(x|t)$, where the subscript $t$ is the elapsed time. Define $M(t)$ as the renewal function of the arrival process. Then the residual life distribution is given by

$$I(x|t) = F(t+x) - \int_0^t \overline{F}(t+x-y)dM(y), \tag{1}$$

where, in general, $\overline{\psi}(\cdot) = 1 - \psi(\cdot)$.

Now we define the time period from the beginning of warm-up state to the next beginning of that as one cycle. Using the residual life $\gamma_t$, we can derive the mean time length of one cycle;

$$T(t_0) = s + \tau + 1/\mu + \int_0^\infty \mathrm{E}[\gamma_{s+\tau+x}]dH(x) + \mathrm{E}[N]\left(\tau + 1/\mu + \int_0^\infty \mathrm{E}[\gamma_{\tau+x}]dH(x)\right), \tag{2}$$

where $\mathrm{E}[N]$ is the expected number of transitions from idle to busy during one cycle, and the probability mass function is

$$\Pr\{N = n\} = \int_0^\infty I(t_0|s+\tau+x)\overline{I}(t_0|\tau+x)I(t_0|\tau+x)^{n-1}dH(x), \qquad \text{for } n = 1, 2, \cdots. \tag{3}$$

Hence, it is found that the expected number of transitions from idle to busy during one cycle is

$$\mathrm{E}[N] = \frac{\Pr\{\int_0^\infty \gamma_{s+\tau+x} \le t_0\}dH(x)}{\Pr\{\int_0^\infty \gamma_{\tau+x} > t_0\}dH(x)} = \frac{\int_0^\infty I(t_0|s+\tau+x)dH(x)}{\int_0^\infty \overline{I}(t_0|\tau+x)dH(x)}. \tag{4}$$

In a fashion similar to the mean time length, the expected power consumed during one cycle is

$$C(t_0) = \Big\{\frac{\rho}{1-\rho}P_1 + P_2\Big\}s + \frac{P_1\tau}{1-\rho} + P_1\Big\{\frac{\rho}{1-\rho}\mathrm{E}[\eta_{s+\tau}] + \mathrm{E}[\eta_{s+\tau} \wedge t_0]\Big\}$$

$$+\mathrm{E}[N]\Big\{\frac{P_1\tau}{1-\rho} + P_1\Big(\frac{\rho}{1-\rho}\mathrm{E}[\eta_\tau] + \mathrm{E}[\eta_\tau \wedge t_0]\Big)\Big\}, \tag{5}$$

where $\mathrm{E}[\eta_t \wedge t_0] = \mathrm{E}[\min(\eta_t, t_0)] = \int_0^{t_0} u dI(u \mid t) + t_0\overline{I}(t_0 \mid t)$. Therefore, from the usual renewal reward theorem, we can obtain the expected power consumed per unit time in the steady-state, $V(t_0) = C(t_0)/T(t_0)$. Then, the problem is to find the optimal auto-sleep time $t_0^*$ which minimizes the expected power consumed per unit time in the steady-state, *i.e.*, $\min_{0\le t_0<\infty} V(t_0)$.

# 3. THE PHASE-TYPE APPROXIMATION

## 3.1 FORMULATION OF THE EXPECTED POWER VIA THE PHASE-TYPE APPROXIMATION

In general, it is difficult to obtain the explicit form of the expected power consumed per unit time in the steady-state for the renewal arrival case. This is due to an analytical difficulty to represent the renewal function. In this section, we propose a structural approximation scheme to generate the optimal auto-sleep schedule effectively, applying the phase-type approximation method. These are based on the fact that an ordinary renewal arrival process can be approximated well by a phase-type renewal process, so that we give an approximation form of the residual distribution $I(t|x)$ in Eqs. (2) and (5).

Before developing the phase-type approximation, we describe the phase-type renewal process. Consider a Markov process on the state space $\{1, 2, \cdots, m+1\}$, where $\{1, 2, \cdots, m\}$ denote the transient states called the *phases*, and $\{m+1\}$ means the absorbing one. The initial probability vector for the Markov process is given by $(\boldsymbol{\alpha}, 0)$, where $\alpha$ is the $1 \times m$ probability vector. Until the absorption in the state $m+1$, the process behaves similar to the Markov process with an infinitesimal generator $T$, where $T$ is a matrix with components $\lambda_{ij}$ $(> 0)$, $1 \le i, j \le m$, $j \ne i$ and $-\lambda_{ii}$ $(< 0)$. In our model, the absorption implies the occurrence of events, *i.e.* the arrival of access requirements. After the absorption, the process is restarted at the phase having the initial probability vector. Then, the time interval of successive arrivals can be represented by the phase-type distribution with parameter $(\boldsymbol{\alpha}, T)$, where the inter-arrival time distribution becomes $F_{PH}(t) = 1 - \boldsymbol{\alpha}\exp(Tt)\boldsymbol{e}$ with a column vector $\boldsymbol{e}$ of 1s.

Let us now return our argument to the phase-type approximation. Denote $N_t$ and $J_t$ be the number of arrivals in $(0, t]$ and the internal state of arrival at time $t$, respectively, where the internal states can be interpreted as the states of various factors which cause the arrival of access requirements. We define the transition probability; $P_{ij}(n, t) = \Pr\{N_t = n, J_t = j \mid N_0 = 0, J_0 = i\}$ and the matrix $P(n, t)$ with components $P_{ij}(n, t)$. Then, the Kolmogorov's forward equation is given by

$$\frac{d}{dt}P(0,t) = P(0,t)T,$$

$$\frac{d}{dt}P(n+1,t) = P(n+1,t)T + P(n,t)\boldsymbol{T}^{\boldsymbol{0}}\boldsymbol{\alpha}, \qquad \text{for } n = 1, 2, \cdots,$$

$$P(0,0) = I, \qquad P(n,0) = O, \qquad \text{for } n = 1, 2, \cdots, \tag{6}$$

where $\boldsymbol{T}^{\mathbf{0}}= -T\boldsymbol{e}$ is the column vector and where $I$ and $O$ are an identity matrix and a zero matrix, respectively. Letting $P^*(z,t) = \sum_{n=0}^{\infty} P(n,t)z^n$ be the matrix generating function, from Eq. (6), we obtain

$$P^*(z,t) = \sum_{n=0}^{\infty} P(n,t)z^n = \exp\left\{(T + z\boldsymbol{T}^{\mathbf{0}}\boldsymbol{\alpha})t\right\}. \tag{7}$$

Hence, we can derive the probability vector $\boldsymbol{g}(t)$ with component $g_j(t)$ which means the probability that the state of process at time $t$ is $j$, that is,

$$\boldsymbol{g}(t) = \boldsymbol{\alpha}\exp\left\{(T + \boldsymbol{T}^{\mathbf{0}}\boldsymbol{\alpha})t\right\}. \tag{8}$$

Therefore, from the Markov property for the phase-type renewal process, it is found that the residual life distribution can be written as

$$I_{PH}(x|t) = 1 - \boldsymbol{g}(t)\exp(Tx)\boldsymbol{e}. \tag{9}$$

Finally, the residual life distribution in Eqs. (2) and (5) can be approximated by $I(x|t) \approx I_{PH}(x|t)$, which leads to an approximation form of the expected power consumed per unit time in the steady-state.

## 3.2 STATISTICAL ESTIMATION PROCEDURE

Since the phase-type renewal process is composed of two stochastic processes which are observable and unobservable, usual statistical estimation methods such as the method of maximum likelihood cannot be used for model parameters. Thus, we introduce the following two estimation methods for the phase-type approximation method.

*(i) The moment matching*

Heijden [7] proposed the following moment matching conditions. If there are $n$ unknown-parameters, they are determined by fitting the first $n$ moments to the sample moments estimated from real data. If the inter-arrival time distribution of the phase-type renewal process obeys the following Coxian-2 distribution;

$$T = \begin{bmatrix} -\lambda_1 & 0 \\ \lambda_2 & -\lambda_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\alpha} = (1-a, a), \tag{10}$$

then the estimators for the parameters are given by

$$= \frac{\lambda_2}{\lambda_1}(m_1\lambda - 1), \tag{11}$$

$$\lambda_1 = \frac{3m_1m_2 - m_3 - \sqrt{m_3^3 + 18m_2^3 + 24m_1^3m_3 - 9m_1m_2(3m_1m_2 + 2m_3)}}{3m_2^2 - 2m_1m_3} \tag{12}$$

and

$$\lambda_2 = \frac{2(m_1\lambda_1 - 1)}{m_2\lambda_1 - 2m_1}, \tag{13}$$

where $m_1$, $m_2$ and $m_3$ are the first three moments of inter-arrival time.

*(ii) The EM-algorithm for phase-type distribution*

The EM (expectation-maximization) algorithm is an iterative method for the maximum likelihood estimation [9, 10]. It is useful to parameterize statistical models including the incomplete data. Suppose that $Y = u(X)$ is observed and that $X$ is unobserved, where $Y$ and $X$ have the probability density functions $g_\gamma$ and $f_\gamma$, respectively. Then, the $(n+1)$-th step in the EM algorithm is to find the value $\gamma_{n+1}$ which maximizes

$$\gamma \to \mathrm{E}[\log f_\gamma(X) \mid u(X) = y;\, \gamma_n], \tag{14}$$

where $y$ is the observed data and $\gamma_n$ is the current estimate after the $n$ steps in the EM-algorithm (see *e.g.* [11] for detail). In particular, when the inter-arrival time distribution has the phase-type distribution, the EM-algorithm is given as follows:

Let $(y_1, y_2, \ldots, y_n)$ be the observed sample data. Then, the $(k+1)$-th iteration of the EM-algorithm is given by

E-Step: Calculate;

$$\pi_i^{(k+1)} = \sum_{l=1}^{n} \mathrm{E}[\pi_i^{(k)} | y_l;\, \hat{\boldsymbol{\alpha}}^{(k)}, \hat{T}^{(k)}], \quad \text{for } i = 1, \cdots, m, \tag{15}$$

$$\xi_i^{(k+1)} = \sum_{l=1}^{n} \mathrm{E}[\xi_i^{(k)} | y_l;\, \hat{\boldsymbol{\alpha}}^{(k)}, \hat{T}^{(k)}], \quad \text{for } i = 1, \cdots, m, \tag{16}$$

$$\Lambda_{ij}^{(k+1)} = \sum_{l=1}^{n} \mathrm{E}[\Lambda_{ij}^{(k)} | y_l;\, \hat{\boldsymbol{\alpha}}^{(k)}, \hat{T}^{(k)}], \quad \text{for } i \neq j, i = 1, \cdots, m \text{ and } j = 1, \cdots, m. \tag{17}$$

M-Step: Generate the new estimates;

$$\hat{\alpha}_i^{(k+1)} = \frac{\pi_i^{(k+1)}}{n}, \quad \hat{t}_{ij}^{(k+1)} = \frac{\Lambda_{ij}^{(k+1)}}{\xi_i^{(k+1)}}, \quad \hat{t}_{ii}^{(k+1)} = -\left( \frac{\Lambda_{i0}^{(k+1)}}{\xi_i^{(k+1)}} + \sum_{j=1, j\neq 1}^{m} \hat{t}_{ij}^{(k+1)} \right), \tag{18}$$

where $\hat{\alpha}_i$ and $\hat{t}_{ij}$ are the elements of $\hat{\boldsymbol{\alpha}}$ and $\hat{T}$, respectively. In the above expressions, $\pi_i$ is the number of Markov processes starting from the state $i$, $\xi_i$ is the total time spent in the state $i$ and $\Lambda_{ij}$ is the total number of jumps from the state $i$ to $j$.

# 4. NUMERICAL EXAMPLES

In this section, we investigate the approximation performance of the phase-type methods proposed in Section 3. Suppose that the arrival of access requirements follows the renewal process with the Weibull inter-arrival time distribution; $F(t) = 1 - \exp\{-(t/\beta_a)^{m_a}\}$, where $m_a = 0.5$ and $\beta_a = \rho/\Gamma(1+1/m_a)$ denote the shape and scale parameters of the Weibull distribution, respectively, and where $\Gamma(\cdot)$ is the standard gamma function. We also suppose that the processing time distribution is the exponential distribution; $H(t) = 1 - \exp(-t)$. The other model parameters are fixed as $P_1 = 1.0$, $P_2 = 4.0$, $\tau = 0.1$ and $s = 1.0$. In our approximation scheme, the inter-arrival time distribution of the phase-type renewal process is equivalent to the Coxian-2 distribution. In addition to the phase-type

Table 1: The optimal auto-sleep time based on the equilibrium approximation.

| $\rho$ | $\hat{t}_0^*$ | $\hat{V}(t_0^*)$ | $V(\hat{t}_0^*)$ |
|---|---|---|---|
| 0.1 | 0.000 | 0.159 | 0.299 (0.278, 0.320) |
| 0.2 | 0.000 | 0.298 | 0.484 (0.454, 0.515) |
| 0.3 | 0.006 | 0.421 | 0.617 (0.579, 0.655) |
| 0.4 | 0.189 | 0.529 | 0.762 (0.728, 0.797) |
| 0.5 | 0.476 | 0.620 | 0.804 (0.761, 0.846) |
| 0.6 | 0.806 | 0.695 | 0.871 (0.825, 0.918) |
| 0.7 | 1.154 | 0.756 | 0.896 (0.855, 0.937) |
| 0.8 | 1.514 | 0.806 | 0.952 (0.911, 0.994) |
| 0.9 | 1.874 | 0.847 | 0.918 (0.873, 0.962) |

Table 2: The optimal auto-sleep time based on the phase-type approximations.

| | moment matching | | | EM-algorithm | | |
|---|---|---|---|---|---|---|
| $\rho$ | $\hat{t}_0^*$ | $\hat{V}(t_0^*)$ | $V(\hat{t}_0^*)$ | $\hat{t}_0^*$ | $\hat{V}(t_0^*)$ | $V(\hat{t}_0^*)$ |
| 0.1 | 0.000 | 0.379 | 0.299 (0.278, 0.320) | 0.000 | 0.334 | 0.299 (0.278, 0.320) |
| 0.2 | 0.000 | 0.608 | 0.484 (0.454, 0.515) | 0.000 | 0.622 | 0.484 (0.454, 0.515) |
| 0.3 | 3.128 | 0.705 | 0.665 (0.626, 0.704) | 0.000 | 0.644 | 0.617 (0.579, 0.655) |
| 0.4 | 3.265 | 0.747 | 0.790 (0.762, 0.819) | 0.037 | 0.852 | 0.773 (0.736, 0.809) |
| 0.5 | 3.058 | 0.778 | 0.815 (0.780, 0.851) | 0.421 | 0.922 | 0.808 (0.764, 0.851) |
| 0.6 | 2.826 | 0.805 | 0.861 (0.824, 0.899) | $\infty$ | 1.000 | 0.997 (0.991, 1.003) |
| 0.7 | 2.625 | 0.828 | 0.881 (0.848, 0.914) | $\infty$ | 1.000 | 1.000 (1.000, 1.000) |
| 0.8 | 2.454 | 0.850 | 0.942 (0.906, 0.977) | $\infty$ | 1.000 | 1.000 (1.000, 1.001) |
| 0.9 | 2.310 | 0.870 | 0.918 (0.880, 0.956) | 0.797 | 0.710 | 0.964 (0.908, 1.019) |

approximation, we calculate the optimal auto-sleep time based on the equilibrium approximation [4] and compare their precision, where the equilibrium approximation is to represent the residual life distribution with the equilibrium distribution of inter-arrival time, that is,

$$I(t|x) \approx F_e(t) = \lambda \int_0^t \overline{F}(u)du. \tag{19}$$

Tables 1 and 2 present the optimal auto-sleep times and their associated minimum expected powers consumed per unit time in the steady-state, based on the equilibrium approximation and the phase-type approximations. In the phase-type approximations, we use the moment matching and the EM-algorithm to estimate the model parameters. Furthermore, we estimate numerically the expected power by the Monte Carlo simulation, provided that the auto-sleep time is given by the estimated optimal solution. On each table, the values in brackets indicate the lower and upper bounds on the confidence interval with significant level 95%, and are calculated by the simulation. From Tables 1 and 2, it is observed that the expected powers estimated by the equilibrium approximation and the moment matching tend not to belong to the corresponding confidence intervals. On the other hand, the phase-type approximation with EM-algorithm can estimate the expected power consumed per unit time within the confidence intervals with significant level 95%. These results show that the the phase-type approximation with EM-algorithm is efficient to calculate the expected power consumed per unit time approximately. However, in estimating the optimal auto-sleep time, the phase-type approximation with EM-algorithm does not always give the best solutions. From Tables 1 and 2, it can be observed that the estimation results for the optimal

auto-sleep time by both the equilibrium approximation and the moment matching are better than that by EM-algorithm. Thus, we can conclude that the equilibrium approximation and the moment matching are useful methods to estimate the optimal auto-sleep time. Also, if one wants to obtain more reliable estimate of the expected power consumed per unit time, the EM-algorithm may function better than the others.

## 5. CONCLUDING REMARKS

In this paper, we have considered the stochastic auto-sleep model under the renewal arrival process, and have proposed two kinds of phase-type approximation methods to represent the expected power consumed per unit time in the steady-state. Based on these approximations, we have calculated the optimal auto-sleep schedule which minimizes the expected power consumed per unit time in the steady-state. In numerical examples, we have investigated the approximation performance for the proposed methods. As a result, we have shown that the phase-type approximations could be useful for finding the optimal auto-sleep time approximately in the heavy traffic circumstance.

## REFERENCES


1. Sandoh, H., Hirakoshi, H. and Kawai, H. (1996), An optimal time to sleep for an auto-sleep system, *Computers & Operations Research*, **23**, 221–227.
2. Dohi, T., Kaio, N. and Osaki, S. (1997), Nonparametric approach to power-saving strategies for a portable personal computer, *Electronics and Communications in Japan, Part 3*, **80**, 80–90.
3. Okamura, H., Dohi, T. and Osaki, S. (1999), On the effect of power saving by auto-sleep functions on a computer system I – modeling by a renewal process – (in Japanese), *Transactions on IPSJ*, **39**, 1858-1869.
4. Okamura, H., Dohi, T. and Osaki, S. (2000), On the effects of power saving by auto-sleep functions on a computer system II – queueing models – (in Japanese), *Transactions on IPSJ*, **40**, 1027–1040
5. Miyazawa, M. (1989), A generalized Pollaczek-Khintchine formula for the $GI/GI/1/K$ queue and its application to approximation, *Stochastic Models*, **3**, 53–65.
6. Altiok, T. (1985), On the phase-type approximations of general distributions , *IIE Transactions*, **17**, 110–116.
7. Heijden, M. C. (1988), On the three-moment approximation of a general distribution by a Coxian distribution, *Probability in the Engineering and Informational Sciences*, **2**, 257–261.
8. Asmussen, S. and Koole, G. (1993), Marked point processes as limits of Markovian arrival streams, *Journal of Applied Probability*, **30**, 356–372.
9. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society, Series B*, **39**, 1–38.
10. Wu, C. F. J. (1983), On the convergence properties of the EM algorithm, *Annals of Statistics*, **11**, 95–103.
11. Asmussen, S., Nerman, O. and Olsson, M. (1996), Fitting phase-type distributions via the EM algorithm, *Scandinavian Journal of Statistics*, **23**, 419–441.

# REPLACEMENT POLICIES FOR A SHOCK MODEL WITH MAINTENANCE AND MINIMAL REPAIR

CUNHUA QIAN[1], SYOUJI NAKAMURA[2] and TOSHIO NAKAGAWA[3]

[1]Department of Industrial Engineering, Aichi Institute of Technology
1247 Yachigusa, Yakusa-cho, Toyota 470-0392, Japan
[2]Systems Division, The Bank of Nagoya, Ltd.
1-chome-501 Kounosu, Tennpaku-ku, Nagoya 468-0003, Japan
[3]Department of Industrial Engineering, Aichi Institute of Technology
1247 Yachigusa, Yakusa-cho, Toyota 470-0392, Japan

qch64317@ie.aitech.ac.jp / pfa00744@nifty.ne.jp / nakagawa@ie.aitech.ac.jp

**Abstract**—This paper considers replacement polices for an extended cumulative damage model with maintenance at each shock and minimal repair at failure: Shocks occur at a non-homogeneous Poisson process. A system undergoes maintenance at each shock when the total damage does not exceed a failure level $K$, undergoes minimal repair at each shock when the total damage exceeds a failure level $K$, and is replaced at time $T$ or at failure $N$, whichever occurs first. The expected cost rate is obtained and optimal $T^*$ and $N^*$ to minimize the expected cost are analytically discussed. It is shown that this model would be applied to the backup of secondary storage files in a database system as an example.

**Keywords**—Shock model, Minimal repair, Replacement, Maintenance, Backup policy.

## 1. INTRODUCTION

In recent years, the database in computer systems has become very important in the highly information-oriented society. In particular, the reliable database is the most indispensable instrument in on-line transaction processing systems such as real-time systems used for account of bank. The data in a computer system are frequently updated by adding or deleting them, and are stored in floppy disks or other secondary media. However, data files in secondary media are sometimes broken by several errors due to noises, human errors and hardware faults. In this case, we have to reconstruct the same files from the beginning.

The most simple and dependable method to ensure the safety of data would be always to make the backup copies of all files in other places as *total backup*, and to take out them if files in the original secondary media are broken. But, this method would take hours and costs when files become large. To make the backup copies efficiently, we make the backup copies of only updated files which have changed or are new since the last full backup when the total update files do not exceed a threshold level $K$. We call it *incremental backup*. This would reduce significantly both duration time and size of backup [1]. Conversely, we perform *full backup* at periodic time $T$, or at $N$-th update since the total updated files have exceeded a threshold level $K$, whichever occurs first. It is assumed that the database system returns to an initial state by the full backup.

Cumulative damage models, where a system suffers damage due to shocks and fails when the total amount of damage exceeds a failure level $K$, generate a cumulative process [2]. Some aspects of damage models from reliability viewpoints were discussed by Esary, Marshall and Proschan [3].

It is of great interest that a system is replaced before failure as preventive maintenance. The replacement policies where a system is replaced before failure at time $T$ [4], at shock $N$ [5], or at

damage $Z$ [6, 7] were considered. Nakagawa and Kijima [8] applied the periodic replacement with minimal repair [9] at failure to a cumulative damage model and obtained optimal values $T^*$, $N^*$ and $Z^*$ which minimize the expected cost.

This paper considers an extended cumulative damage model with maintenance at each shock and minimal repair at each failure. Reliability measures of this model are derived, using the theory of cumulative processes. Further, this is applied to the backup of files in a database system.

## 2. PROBLEM FORMULATION

Suppose that shocks occur at a nonhomogeneous Poisson process with an intensity function $\lambda(t)$ and a mean-value function $R(t)$, *i.e.*, $R(t) \equiv \int_0^t \lambda(u)du$. Then, the probability that shocks occur exactly $j$ times during $[0,t]$ is [10]

$$H_j(t) \equiv \frac{[R(t)]^j}{j!}e^{-R(t)} \quad (j = 0, 1, 2, \cdots). \tag{1}$$

Further, an amount $Y_j$ of damage due to the $j$-th shock has a probability distribution $G_j(x) \equiv P_r\{Y_j \leq x\}$ $(j = 1, 2, \cdots)$ with finite mean. Then, the total damage $Z_j \equiv \sum_{i=1}^{j} Y_i$ to the $j$-th damage shock where $Z_0 \equiv 0$ has a distribution

$$G^{(j)}(x) \equiv \Pr\{Z_j \leq x\} = G_1 * G_2 * \cdots * G_j(x) \quad (j = 0, 1, 2, \cdots), \tag{2}$$

where $G^{(0)}(x) \equiv 1$ for $x \geq 0$, 0 for $x < 0$, and the asterisk mark represents the Stieltjes convolution, *i.e.*, $a * b(t) \equiv \int_0^t b(t-u)da(u)$ for any functions $a(t)$ and $b(t)$. Then, the probability that the total damage exceeds exactly a failure level $K$ at $j$-th shock is $G^{(j-1)}(K) - G^{(j)}(K)$. Let $Z(t)$ be the total amount of damage at time $t$. Then, the distribution of $Z(t)$ is [3]

$$\Pr\{Z(t) \leq x\} = \sum_{j=0}^{\infty} H_j(t)G^{(j)}(x). \tag{3}$$

Consider the system which should operate for an infinite time span and assume: When the total damage does not exceed a failure level $K$, the system undergoes maintenance at each shock, and the maintenance cost is $c_2 + c_0(x)$ when the total damage is $x$ $(0 \leq x < K)$. It is assumed that the function $c_0(x)$ is continuous and strictly increasing and $c_0(0) \equiv 0$. When the total damage exceeds a failure level $K$, the system undergoes minimal repair at each failure, and the repair cost is $c_3$, where $c_3 = c_2 + c_0(K)$. The system is replaced at periodic time $T$, or at failure $N$, whichever occurs first, and the replacement cost is $c_1$, where $c_3 < c_1$. The maintenance time, the repair time and the replacement time are negligible, *i.e.*, the time considered here is measured only by the total operating time of the system. Then, the probability $P_T$ that the system is replaced at time $T$ is

$$P_T = \sum_{j=0}^{\infty} H_j(T)G^{(j)}(K) + \sum_{j=0}^{\infty}[G^{(j)}(K) - G^{(j+1)}(K)] \sum_{i=j+1}^{j+N-1} H_i(T) \tag{4}$$

$$= \sum_{j=0}^{\infty}[G^{(j)}(K) - G^{(j+1)}(K)] \sum_{i=0}^{j+N-1} H_i(T), \tag{5}$$

and the probability $P_N$ that the system is replaced at failure $N$ is

$$P_N = \sum_{j=0}^{\infty}[G^{(j)}(K) - G^{(j+1)}(K)] \sum_{i=j+N}^{\infty} H_i(T) \tag{6}$$

$$= \sum_{j=0}^{\infty}[G^{(j)}(K) - G^{(j+1)}(K)] \int_0^T H_{j+N-1}(t)\lambda(t)dt, \tag{7}$$

where $G^{(i)}(K) \equiv 1$ for $i < 0$. It is evident that

$$P_T + P_N = \sum_{j=0}^{\infty} H_j(T)G^{(j-N+1)}(K) + \sum_{j=0}^{\infty} H_j(T)[1 - G^{(j-N+1)}(K)] = 1.$$

Let $M_1(T)$ and $M_2(T, N)$ denote the expected numbers of maintenances and minimal repairs until replacement, respectively. Then, from (4) and (6), we have

$$M_1(T) = \sum_{j=0}^{\infty} jH_j(T)G^{(j)}(K) + \sum_{j=0}^{\infty} j[G^{(j)}(K) - G^{(j+1)}(K)] \sum_{i=j+1}^{\infty} H_i(T)$$

$$= \sum_{j=1}^{\infty} H_j(T) \sum_{i=1}^{j} G^{(i)}(K), \tag{8}$$

$$M_2(T, N) = \sum_{j=0}^{\infty}[G^{(j)}(K) - G^{(j+1)}(K)]\{\sum_{i=j+1}^{j+N-1}(i-j)H_i(T) + \sum_{i=j+N}^{\infty}(N-1)H_i(T)\}$$

$$= \sum_{j=1}^{\infty} H_j(T) \sum_{i=j-N+2}^{j} [1 - G^{(i)}(K)]. \tag{9}$$

Thus, the total expected cost $E[C]$ to replacement is

$$E[C] = c_1 + \sum_{j=1}^{\infty} H_j(T) \sum_{i=1}^{j} \int_0^K [c_2 + c_0(x)]dG^{(i)}(x) + c_3 M_2(T, N). \tag{10}$$

Let $E[U]$ denote the mean time to replacement. Then, from (5) and (7), we have

$$E[U] = \sum_{j=0}^{\infty}[G^{(j)}(K) - G^{(j+1)}(K)] \int_0^T tH_{j+N-1}(t)\lambda(t)dt + TP_T$$

$$= \sum_{j=0}^{\infty} G^{(j-N+1)}(K) \int_0^T H_j(t)dt. \tag{11}$$

Therefore, from (10) and (11), by using the theory of renewal process [11], the expected cost per unit time is $C(T, N) = E[C]/E[U]$.

# 3. OPTIMAL POLICY

Suppose that shocks occur at a Poisson process with rate $\lambda$, *i.e.*, $\lambda(t) = \lambda$, $R(t) = \lambda t$ and $H_j(t) = [(\lambda t)^j/j!]e^{-\lambda t}$ $(j = 0, 1, 2, \cdots)$. Further, assume that the cost of maintenances is proportional to the total damage, *i.e.*, $c_2 + c_0(x) = c_2 + c_0 x$ $(0 \le x < K)$. Then, the expected cost per unit of time is

$$\frac{C(T,N)}{\lambda} = c_3 + \frac{c_1 - A(T,N)}{\sum_{j=1}^{\infty} H_j(T) \sum_{i=1}^{j} G^{(i-N)}(K)}, \tag{12}$$

where

$$A(T,N) \equiv c_3 \sum_{j=1}^{\infty} H_j(T) \sum_{i=0}^{j-N} [G^{(i)}(K) - G^{(i+1)}(K)] + c_0 \sum_{j=1}^{\infty} H_j(T) \sum_{i=1}^{j} \int_0^K G^{(i)}(x)dx. \tag{13}$$

If $M(K) \equiv \sum_{j=1}^{\infty} G^{(j)}(K) < \infty$, then $C(0,N) \equiv \lim_{T\to 0} C(T,N) = \infty$ for all $N$ and $C(\infty,\infty) \equiv \lim_{T\to\infty, N\to\infty} C(T,N) = c_3\lambda$. Thus, there exists a positive pair $(T^*, N^*)$ $(0 < T^*, N^* \le \infty)$ which minimizes $C(T,N)$.

**Remark 1** *The expected cost per unit of time when the system is replaced only at failure $N$ is*

$$\frac{C(N)}{\lambda} \equiv \lim_{T\to\infty} \frac{C(T,N)}{\lambda} = c_3 + \frac{c_1 - c_3 - c_0 \int_0^K M(x)dx}{M(K) + N} \quad (N = 1, 2, \cdots). \tag{14}$$

*If $\int_0^K M(x)dx > (c_1 - c_3)/c_0$ then $N^* = 1$, and the system should be replaced at the first shock after the total damage has exceeded a failure level $K$. Conversely, if $\int_0^K M(x)dx \le (c_1 - c_3)/c_0$ then $N^* = \infty$, and the resulting cost is $C(\infty)/\lambda = c_3$.*

In general, let an optimal pair $(T^*, N^*)$ denote a positive solution which minimizes $C(T,N)$. It is evident that

$$\sum_{j=1}^{\infty} H_j(T) \sum_{i=1}^{j} G^{(i-N-1)}(K) - \sum_{j=1}^{\infty} H_j(T) \sum_{i=1}^{j} G^{(i-N)}(K) = \sum_{j=N+1}^{\infty} H_j(T)[1 - G^{(j-N)}(K)] > 0,$$

and

$$A(T,N) - A(T,N+1) = c_3 \sum_{j=N}^{\infty} H_j(T)[G^{(j-N)}(K) - G^{(j-N+1)}(K)] > 0.$$

Thus, we have the following property for $(T^*, N^*)$:

**Remark 2** *If $c_1 \le A(T^*,N)$ for some $N$ then $N^* = 1$, and if $c_1 > A(T^*,N)$ for all $N$, then $N^* = \infty$.*

## 3.1 MINIMAL REPAIR MODEL

First, consider an optimal policy for the minimal repair model, *i.e.*, the system undergoes minimal repair at each shock when the total damage exceeds a failure level $K$, and the system is replaced at time $T$. Since we put $N = \infty$ in (12), the expected cost per unit of time is

$$\frac{C_1(T)}{\lambda} \equiv \lim_{N\to\infty} \frac{C(T,N)}{\lambda} = c_3 + \frac{c_1/\lambda - c_0 \sum_{j=0}^{\infty} \int_0^K G^{(j+1)}(x)dx \int_0^T H_j(t)dt}{T}. \tag{15}$$

Since $C_1(0) \equiv \lim_{T\to 0} C_1(T) = \infty$ and $C_1(\infty) \equiv \lim_{T\to\infty} C_1(T) = c_3\lambda$, then there exists a positive $T_1^*$ $(0 < T_1^* \le \infty)$ which minimizes (15). A necessary condition that a finite $T_1^*$ minimizes $C_1(T)$ is

given by differentiating $C_1(T)$ with respect to $T$ and setting it equal to zero. Hence, from (15), we have

$$\sum_{j=1}^{\infty} H_j(T) \sum_{i=1}^{j} \int_0^K [G^{(i)}(x) - G^{(j)}(x)]dx = \frac{c_1}{c_0}. \tag{16}$$

Letting $Q(T)$ be the left-hand side of (16), we have

$$Q(0) \equiv \lim_{T\to 0} Q(T) = 0,$$

$$Q(\infty) \equiv \lim_{T\to\infty} Q(T) = \int_0^K M(x)dx,$$

$$Q'(T) = \lambda \sum_{j=1}^{\infty} H_j(T) j \int_0^K [G^{(j)}(x) - G^{(j+1)}(x)]dx > 0.$$

Thus, $Q(T)$ is a strictly increasing function from 0 to $\int_0^K M(x)dx$.

Therefore, we have:

**Theorem 1** *If $\int_0^K M(x)dx > c_1/c_0$ then there exists a finite and unique $T_1^*$ $(0 < T_1^* < \infty)$ which minimizes $C_1(T)$, and it satisfies (16). The resulting cost is*

$$\frac{C_1(T_1^*)}{\lambda} = c_3 - c_0 \sum_{j=0}^{\infty} H_j(T_1^*) \int_0^K G^{(j+1)}(x)dx. \tag{17}$$

*If $\int_0^K M(x)dx \le c_1/c_0$ then $T_1^* = \infty$ and $C_1(\infty)/\lambda = c_3$.*

**Example 1** Suppose that a database is updated according to a Poisson process with rate $\lambda$. Further, an amount of only files, which changed or are new since the last full backup, arises from the $j$-th update, is $Y_j$. It is assumed that each $Y_j$ has a probability distribution $G_j(x) = 1 - e^{-\mu x}$, *i.e.*, $G^{(j)}(x) = 1 - \sum_{i=0}^{j-1}[(\mu x)^i/i!]e^{-\mu x}$ $(j = 1, 2, \cdots)$ and $M(K) = \mu K$. We replace *shock* by *update*, *damage* by *dumped files*, *maintenance* by *incremental backup*, *minimal repair* by *total backup* and *replacement* by *full backup*. Then, equation (16) is simplifird as

$$\sum_{j=2}^{\infty} H_j(T) \sum_{i=1}^{j-1} i G^{(i+1)}(K) = \frac{c_1}{c_0/\mu}. \tag{18}$$

Letting $Q_1(T)$ be the left-hand side of (18), we have that $Q_1(0) = 0$, and $Q_1(\infty) = (\mu K)^2/2$. Thus, $Q_1(T)$ is a strictly increasing function of $T$ from 0 to $(\mu K)^2/2$. If $\mu K^2/2 > c_1/c_0$ then there exists a finite and unique $T_1^*$ $(0 < T_1^* < \infty)$ which satisfies (18), and the resulting cost is

$$\frac{C_1(T_1^*)}{\lambda} = c_2 + \frac{c_0}{\mu} \sum_{j=0}^{\infty} H_j(T_1^*) \sum_{i=1}^{j+1} G^{(i)}(K). \tag{19}$$

If $\mu K^2/2 \le c_1/c_0$ then $T_1^* = \infty$, and the resulting cost is $c_3\lambda$.

It is supposed that the total volume of files is $5 \times 10^5$ trucks and a threshold level $K$ is $3 \times 10^5$ trucks which correspond to 60% of the total volume. Table 1 gives the optimal full backup times $\lambda T_1^*$, the resulting costs $C_1(T_1^*)/\lambda$ for $c_1 = 70, 90, 110, 140, 200, 260, 320, 440$, and $\mu K = 12, 24$ when $c_2 = 10$ and $c_0 = 2 \times 10^{-4}$. It is found from the optimal policy that if $30\mu K > c_1$ then

$T_1^* < \infty$, and conversely, if $30\mu K \le c_1$ then $T_1^* = \infty$ and $C_1(\infty)/\lambda = 70$. This shows that both optimal $T_1^*$ and costs $C_1(T_1^*)$ are increasing with $c_1$, and $C_1(T_1^*)$ are decreasing with $\mu K$. However, $T_1^*$ are smaller for small $c_1$, and conversely, are greater for large $c_1$, when $\mu K$ is smaller. This reason would be explained that if the cost $c_1$ is small then it is better to perform the full backup early, but if $c_1$ is large then it is better to do it lately, especially when its mean updated file is large.

Table 1. Optimal full backup times $\lambda T_1^*$ and resulting costs $C_1(T_1^*)/\lambda$ for minimal repair model

| $c_1$ | | 70 | 90 | 110 | 140 | 200 | 260 | 320 | 440 |
|---|---|---|---|---|---|---|---|---|---|
| $\mu K = 12$ | $\lambda T_1^*$ | 5.418 | 6.211 | 6.953 | 8.020 | 10.163 | 12.652 | 16.675 | $\infty$ |
| | $C_1(T_1^*)/\lambda$ | 41.272 | 44.726 | 47.771 | 51.787 | 58.427 | 63.731 | 67.935 | 70.000 |
| $\mu K = 24$ | $\lambda T_1^*$ | 7.486 | 8.492 | 9.393 | 10.611 | 12.740 | 14.636 | 16.422 | 19.981 |
| | $C_1(T_1^*)/\lambda$ | 31.206 | 33.710 | 35.947 | 38.947 | 44.089 | 48.475 | 52.341 | 58.961 |

For example, when the mean time of update is $1/\lambda = 1$ day, $c_1 = 320$ and $\mu K = 12$, the optimal full backup time $T_1^*$ is about 17 days. In this case, $K/(\lambda/\mu) = 12$ days, and note that it represents the mean time until the total updated files exceed a threshold level $K$.

## 3.2 PREVENTIVE REPLACEMENT MODEL

In this section, consider an optimal policy for the preventive replacement model, *i.e.*, the system is replaced at periodic time $T$, or at failure, whichever occurs first. Putting that $N = 1$ in (12), the expected cost per unit of time is

$$\frac{C_2(T)}{\lambda} = \frac{c_1/\lambda + \sum_{j=0}^{\infty} \int_0^K (c_2 + c_0 x) dG^{(j+1)}(x) \int_0^T H_j(t)dt}{\sum_{j=0}^{\infty} G^{(j)}(K) \int_0^T H_j(t)dt}. \tag{20}$$

Since $C_2(0) \equiv \lim_{T\to 0} C_2(T) = \infty$ and from (14),

$$\frac{C_2(\infty)}{\lambda} \equiv \lim_{T\to\infty} \frac{C(T,1)}{\lambda} = c_3 + \frac{c_1 - c_3 - c_0 \int_0^K M(x)dx}{M(K)+1}, \tag{21}$$

there exists a positive $T_2^*$ $(0 < T_2^* \le \infty)$ which minimizes (20). A necessary condition that a finite $T_2^*$ minimizes $C_2(T)$ is given by differentiating $C_2(T)$ with respect to $T$ and setting it equal to zero. Hence, from (20), we have

$$\sum_{j=0}^{\infty} [V(T)G^{(j)}(K) - \int_0^K (c_2 + c_0 x) dG^{(j+1)}(x)] \int_0^T H_j(t)\lambda dt = c_1, \tag{22}$$

where

$$V(T) \equiv \frac{\sum_{j=0}^{\infty} \int_0^K (c_2 + c_0 x) dG^{(j+1)}(x) H_j(T)}{\sum_{j=0}^{\infty} G^{(j)}(K) H_j(T)}. \tag{23}$$

Letting $U(T)$ be the left-hand side of (22), we have

$$U(0) \equiv \lim_{T\to 0} U(T) = 0,$$

$$U(\infty) \equiv \lim_{T\to\infty} U(T) = V(\infty)[1+M(K)] - \int_0^K (c_2 + c_0 x)dM(x),$$

$$U'(T) = V'(T)\sum_{j=0}^{\infty} G^{(j)}(K)\int_0^T H_j(t)\lambda dt,$$

where $V(\infty) \equiv \lim_{T\to\infty} V(T)$. If $V(T)$ is a strictly increasing function, $U(T)$ is also a strictly increasing function from 0 to $U(\infty)$.

Therefore, we have:

**Theorem 2** *If $V'(T) > 0$ and $U(\infty) > c_1$ then there exists a finite and unique $T_2^*$ which minimizes $C_2(T)$, and it satisfies (22). The resulting cost is $C_2(T_2^*)/\lambda = V(T_2^*)$. If $V'(T) \le 0$ or $U(\infty) \le c_1$ then $T_2^* = \infty$ and $C_2(\infty)/\lambda$ is given in (21). This corresponds to the case of $N^* = 1$ in Remark 2.*

**Example 2** In example 1, we perform full backup at periodic time $T$, or when the total update files have exceeded a threshold lever $K$, whichever occurs first. When $G_j(x) = 1 - e^{-\mu x}$, *i.e.*, $G^{(j)}(x) = 1 - \sum_{i=0}^{j-1}[(\mu x)^i/i!]e^{-\mu x}$ $(j = 1, 2, \cdots)$ and $M(K) = \mu K$, equation (22) is simplifird as

$$\sum_{j=0}^{\infty}[V(T)G^{(j)}(K) - c_2 G^{(j+1)}(K) - \frac{c_0}{\mu}(j+1)G^{(j+2)}(K)]\sum_{i=j+1}^{\infty} H_i(T) = c_1, \tag{24}$$

and equation (23) is

$$V(T) = c_2\frac{\sum_{j=0}^{\infty} G^{(j+1)}(K)H_j(T)}{\sum_{j=0}^{\infty} G^{(j)}(K)H_j(T)} + \frac{c_0}{\mu}\frac{\sum_{j=0}^{\infty}(j+1)G^{(j+2)}(K)H_j(T)}{\sum_{j=0}^{\infty} G^{(j)}(K)H_j(T)}. \tag{25}$$

Table 2 gives the optimal full backup times $\lambda T_2^*$, the resulting costs $C_2(T_2^*)/\lambda$ for $c_1 = 70$, 90, 110, 140, 200, 260, 320, 440, and $\mu K = 12$, 24 when $c_2 = 10$ and $c_0 = 2 \times 10^{-4}$. This shows that both optimal $T_2^*$ and costs $C_2(T_2^*)$ are increasing with $c_1$, and $\lambda T_2^* < \mu K$.

Table 2. Optimal full backup times $\lambda T_2^*$ and resulting costs $C_2(T_2^*)/\lambda$ for preventive replacement model

| $c_1$ | | 70 | 90 | 110 | 140 | 200 | 260 | 320 | 440 |
|---|---|---|---|---|---|---|---|---|---|
| $\mu K = 12$ | $\lambda T_2^*$ | 6.311 | 7.922 | 11.094 | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| | $C_2(T_2^*)/\lambda$ | 40.318 | 43.225 | 45.555 | 47.692 | 52.308 | 56.923 | 61.538 | 70.769 |
| $\mu K = 24$ | $\lambda T_2^*$ | 7.515 | 8.553 | 9.505 | 10.843 | 13.443 | 16.396 | $\infty$ | $\infty$ |
| | $C_2(T_2^*)/\lambda$ | 31.196 | 33.686 | 35.903 | 38.856 | 43.831 | 47.919 | 51.200 | 56.000 |

It is found from Table 1 and Table 2 that $\lambda T_1^* < \lambda T_2^*$ and $C_1(T_1^*)/\lambda > C_2(T_2^*)/\lambda$, that is, the preventive replacement model is better than the minimal repair one. But, if $T_2^* = \infty$ and $C_2(\infty)/\lambda > C_1(\infty)/\lambda = c_3$, *i.e.*, $c_1 > c_3 + \int_0^K M(x)dx$, then the system should undergoes minimal repair at each shock foreever.

In general, note that $dA(T,N)/dT > 0$ in (13) and $A(T,N)$ is strictly decreasing in $N$. From Remark 2, if $c_1 > A(\infty,1) = c_3 + \int_0^K M(x)dx$, then $(T^*,N^*) = (\infty,\infty)$, and if $c_1 \leq c_3 + \int_0^K M(x)dx$, then $N^* = 1$ and $T^* = T_2^*$.

**Remark 3** *If* $c_1 > c_3 + \int_0^K M(x)dx$ *then* $(T^*,N^*) = (\infty,\infty)$; *If* $c_1 \leq c_3 + \int_0^K M(x)dx$ *then* $(T^*,N^*) = (T_2^*,1)$.

# 4. CONCLUSIONS

We have proposed the extended cumulative damage model with maintenance at each shock and minimal repair at failure, and is replaced at scheduled time $T$ or at failure $N$, whichever occurs first. Using the theory of cumulative processes, we derive the expected cost and discuss the optimal replacement policy which minimizes it.

Further, we have shown that this would be applied to the backup of secondary storage files in the database system. Thus, by estimating the costs of backups and the amount of dumped files from actual data and by modifying some suppositions, we could practically determine a scheduled time of full backup. These formulations and results would be applied to other management policies for computer systems [12].

**Acknowledgment** This form a part of research results by the Hori Information Science Promotion Foundation.

# REFERENCES

1. K. Suzuki and K. Nakajima, Storage management software, *Fujitsu,* **46**, 389-397 (1995).
2. D. R. Cox, *Renewal Theory,* Methuen, London (1962).
3. J. D. Esary, A. W. Marshall and F. Proschan, Shock models and wear processes, *Annals of Probability,* **1**, 627-649 (1973).
4. H. M. Taylor, Optimal replacement under additive damage and other failure models., *Naval Res. Logist. Quart,* **22**, 1-18 (1975).
5. T. Nakagawa, A summary of discrete replacement policies, *European J. of Operational Research,* **17**, 382-392(1984).
6. R. M. Feldman, Optimal replacement with semi-Markov shock models, *J.Appl. Prob.,* **13**, 108-117 (1976).
7. T. Nakagawa, On a replacement problem of a cumulative damage model, *Operational Research Quarterly,* **27** 895-900 (1976).
8. T. Nakagawa and M. Kijima, Replacement policies for a cumulative damage model with minimal repair at failure, *IEEE Trans. Reliability,* **13**, 581-584 (1989).
9. R. E. Barlow and F. Proschan, *Mathematical Theory of Reliability,* John Wiley & Sons, New York (1965).
10. S. Osaki, *Applied Stochastic Systems Modeling,* Springer Verlag, Berlin (1992).
11. S. M. Ross, *Applied Probability Models with Optimization Applications,* Holden-Day, San Francisco (1970).
12. T. Satow, K. Yasui and T. Nakagawa, Optimal garbage collection policies for a database in a computer system, *RAIRO-Operations Research,* **30** 359-372 (1996).

# OPTIMAL INSPECTION POLICIES FOR A SCALE

HIROAKI SANDOH[1] and NOBUKO IGAKI[2]

[1] Department of Information and Management Science,
University of Marketing and Distribution Sciences
3-1, Gakuen-nishi-machi, Nishi, Kobe, 651-2188, Japan
[2] Department of Business Administration, Tezukayama University
1-1, Tezukayama, Nara, 631-8501, Japan
sandoh@umds.ac.jp / igaki@tezukayama-u.ac.jp

**Abstract**—In the final stage of manufacturing some specific products such as chemical ones, they weigh each product using a scale and mark the weighing result on each product. However, the scale will occasionally undergo malfunction or a failure during the weighing process. The products weighed by such a scale will be shipped with different marks from their actual weights. In the case of chemical products, those products with wrong marks can be regarded as defective products.

This study proposes two types of periodical inspection policies for a scale to adjust it or to detect its malfunction or a failure followed by repair. The inspection in this study involves adjustment operations by which the malfunction or failures of the scale can be detected and the scale will recover from its malfunction or failure. Under Policy I, the scale is inspected at time $iT(i = 1, 2, \cdots)$. Under Policy II, we consider a situation where the scale is inspected every morning before we start daily work of weighing products, which can be observed in the actual circumstances. For such a case, we can carry out an inspection at $i\tau/N(i = 0, 1, 2, \cdots, N-1)$, where $\tau$ signifies the working hours per day and an integer $N$ denotes the inspection frequency to be conducted per day.

Two types of objective functions are considered; (1) the fraction defective and (2) the long-run average cost under each policy. Under Policy I (Policy II), we examine the existence of an inspection interval $T_\alpha$ (inspection frequency $N_\alpha$) which guarantees that the fraction defective does not exceed a prespecified value $\alpha(0 < \alpha < 1)$. An economical inspection interval $T^*$ (inspection frequency $N^*$) minimizing the long-run average cost is also discussed. Numerical examples are presented to illustrate the proposed inspection policy formulations.

**Keywords**—Inspection policy, Scale, Fraction defective, Long-run average cost.

## 1. INTRODUCTION

In the final stage of manufacturing some specific products such as chemical ones, there is a process in which each product is weighed using a scale to mark each weight on the product. This process is not emphasized generally and its associated cost is reduced as much as possible since it does not affect the product quality. However, the scale occasionally undergoes its malfunction or failures, and this malfunction or a failure can be detected only by an inspection. When the scale becomes out of order, it will indicate different weight for each product from the actual one, and hence each product will be shipped with a different mark from each actual weight. In the case of chemical products particularly, this will be a significant problem if their consumers believe the wrong weights indicated on them and use them for chemical reactions.

The present study concentrates on the products that are marked with wrong weights and calls them *defectives*. In addition, it is postulated that we cannot devote a large expense to this weighing

process as observed in the actual circumstances. We then discuss two types of inspection policies for a scale.

Under Policy I, we conduct an inspection to a scale at $iT(i = 1, 2, \cdots)$ to adjust the scale or to detect malfunction or a failure followed by repair. Under Policy II, we consider a situation where the scale is inspected every morning before we start daily work of weighing products. For such a case, we can inspect the scale at $i\tau/N(i = 0, 1, 2, \cdots, N-1)$, where $\tau$ signifies the working hours per day and an integer $N$ denotes an inspection frequency to be carried out per day.

Two types of objective functions are considered; (1) the fraction defective and (2) the long-run average cost under each policy. Under Policy I (Policy II), we examine the existence of an inspection interval $T_\alpha$ (inspection frequency $N_\alpha$) which guarantees that the fraction defective does not exceed a prespecified value $\alpha(0 < \alpha < 1)$. An economical inspection interval $T^*$ (inspection frequency $N^*$) minimizing the long-run average cost is also discussed. Numerical examples are presented to illustrate the proposed inspection policy formulations.

On the other hand, inspection policies have a long validated history. Most of studies associated with inspection policies have considered to carry out an inspection with a view to detecting a system failure which cannot be detected instantly[1–30]. Among these studies, Barlow and Proschan[3], Munford and Shahani[4,5], Tadikamalla[14], Wattanapanom and Shaw[15], Nakagawa and Yasui[16,17], Kaio and Osaki[20,21] have proposed methods for obtaining inspection points in time $\{x_1, x_2, \cdots\}$. Weiss[2] and Kaio and Osaki[23] have considered models under imperfect inspections, and Luss and Kander[9] have dealt with a model when time required for an inspection is not negligible. Zacks and Fenske[6], Lussand and Kander[8], and Kander[13] have discussed inspection policies for a $n$-unit system. Approximately optimal policies have been studied by Munford and Shahani[4], and Anbar[10]. Yum and MacDowell[25] and Gassandras and Han[29] have applied inspection policies to a manufacturing system.

The above studies assume that since we cannot detect a system failure instantly, we incur cost depending on the period over which we leave the failed system as it is. The cost in is, however, based on not a concrete but an abstract concept. In addition, most of the above studies focus on the period from the time when we start to use a system to the time when the system failure is detected. In addition, the above studies focus on the time interval from when we start to use the system to when we detect the system failure. In this study, however, the period over which the scale is left to be out of order corresponds to the volume of defectives. In addition, the scale is used again after it is adjusted by an inspection.

## 2. ASSUMPTIONS

In this study, we make the following assumptions:

(1) The malfunction or a failure of the scale can be detected only by an inspection. Furthermore, an inspection involves adjustment operations by which the scale can recover from its malfunction or a failure. Hence the scale enters its normal state immediately after an inspection.

(2) The number of products to be weighed is very large and thus we can regard it as being continuous by corresponding their volume to the time to be spent in weighing them.

(3) The malfunction or failure time distribution of the scale is expressed by $F(t)$ with mean $\mu$, i.e.,

$$\mu = \int_0^\infty t dF(t) = \int_0^\infty \overline{F}(t) dt \tag{1}$$

# 3. POLICY I

This section discusses Policy I under which we conduct an inspection at time $iT(i = 1, 2, \cdots)$ to adjust the scale or to detect its malfunction or a failure. From assumption (1), the process behavior generates a renewal reward process[31, 32] where a renewal point corresponds to the time when an inspection is completed.

## 3.1 FRACTION DEFECTIVE

Since we regard products with different marks from their actual weights as defectives, the fraction defective in this study can be defined by the ratio of the volume of shipped defective products to that of all the shipped products. From the renewal reward theory[31, 32], the fraction defective $Q_1(T)$ under Policy I is given by

$$\begin{aligned} Q_1(T) &= \lim_{t\to+\infty} \frac{E[\text{time during which the scale is out of order over } (0,t]]}{t} \\ &= \frac{B_1(T)}{A_1(T)}, \end{aligned} \tag{2}$$

where $A_1(T)$ and $B_1(T)$ respectively denote the expected cycle length and the expected time during which the scale is out of order over one cycle.

Since we have

$$A_1(T) = T, \tag{3}$$

$$B_1(T) = 0 \times \overline{F}(T) + \int_0^T (T-t)dF(t) = \int_0^T F(t)dt, \tag{4}$$

the fraction defective in Eq. (2) becomes

$$Q_1(T) = \frac{\int_0^T F(t)dt}{T} = 1 - \frac{\int_0^T \overline{F}(t)dt}{T}. \tag{5}$$

We here consider an upper limit for the inspection interval $T_\alpha$ which makes the fraction defective equal to $100\alpha\%$ or less for a prespecified value of $\alpha$ $(0 < \alpha < 1)$. From Eq. (5), we have

$$\lim_{T\to+0} Q_1(T) = 0, \tag{6}$$

$$\lim_{T\to+\infty} Q_1(T) = 1, \tag{7}$$

$$Q_1'(T) = \frac{-[T\overline{F}(T) - \int_0^T \overline{F}(t)dt]}{T^2}. \tag{8}$$

Let $R(T)$ denote the numerator of the right-hand-side of Eq. (8), i.e., let

$$R(T) = -T\overline{F}(T) + \int_0^T \overline{F}(t)dt, \tag{9}$$

then we have

$$R'(T) = Tf(T) > 0, \tag{10}$$

$$\lim_{T \to +0} R(T) = 0. \tag{11}$$

This indicates that $R(T) > 0$ for $T > 0$, and thus we have $Q_1'(T) > 0$. From Eqs. (6) and (7), there exists a finite upper limit $T_\alpha (> 0)$ for an inspection interval which satisfies $Q_1(T_\alpha) \leq \alpha (0 < \alpha < 1)$.

### 3.2 ECONOMICAL INSPECTION INTERVAL

In this subsection, we discuss an economical inspection interval $T^*$ which minimizes the long-run average cost. From the renewal reward theory, the long-run average cost of the proposed inspection policy is given by

$$C_1(T) = \lim_{t \to +\infty} \frac{E[\text{total cost over } (0,t]]}{t} = \frac{c_1 \int_0^T (T-t) dF(t) + c_2}{T}$$

$$= \frac{c_1 \int_0^T F(t) dt + c_2}{T}. \tag{12}$$

It should be noted that the above formulation coincides with that of Moldel II for block replacement policy proposed by Osaki[33].

By differentiating $C_1(T)$ with respect to $T$, we can show that $C_1'(T) \geq 0$ agrees with

$$R(T) \geq \frac{c_2}{c_1}, \tag{13}$$

where $R(T)$ is given by Eq. (9). From Eqs. (10) and (11), if

$$\lim_{T \to +\infty} R(T) = \mu > \frac{c_2}{c_1}, \tag{14}$$

there exists a unique finite economical inspection interval $T^* (> 0)$ which minimizes $C_1(T)$. If the inequality in (14) does not hold, we have $C_1'(T) \leq 0$ and thus $T^* = +\infty$ which suggests to conduct no inspections.

## 4. POLICY II

This section considers a situation where we perform an inspection to make adjustment to the scale every morning just before we start our daily work of weighing products. In such a situation, we can divide our daily work hours $\tau$ into $N$ divisions. At $i\tau/N (i = 0, 1, 2, N-1,\ N = 1, 2, \cdots)$, we perform an inspection to adjust the scale or to detect its malfunction or failure followed by repair. It should be noted that $N = 1$ corresponds to the policy which conducts only one inspection every morning before we start our daily work. From assumption (1), the process behavior generates a renewal reward process where a renewal point corresponds to the time immediately after the inspection is finished.

## 4.1 FRACTION DEFECTIVE

The definition of the fraction defective is identical to that in **3.1**. The fraction defective $Q_2(N)$ under Policy II is given by

$$Q_2(N) = \lim_{t\to+\infty} \frac{E[\text{time during which the scale is out order over } (0,t]]}{t}$$

$$= \frac{B_2(N)}{A_2(N)}, \quad N = 0, 1, 2, \cdots, N-1, \tag{15}$$

where $A_2(N)$ and $B_2(N)$ are the expected cycle length and the expected time representing the volume of defective products per one cycle.

We here have

$$A_2(N) = \frac{\tau}{N}, \tag{16}$$

$$B_2(N) = 0 \times \overline{F}\left(\frac{\tau}{N}\right) + \int_0^{\frac{\tau}{N}} \left(\frac{\tau}{N} - t\right) dF(t) = \int_0^{\frac{\tau}{N}} F(t)dt, \tag{17}$$

and therefore $Q_2(N)$ in Eq. (15) becomes

$$Q_2(N) = \frac{\int_0^{\frac{\tau}{N}} F(t)dt}{\frac{\tau}{N}} = 1 - \frac{\int_0^{\frac{\tau}{N}} \overline{F}(t)dt}{\frac{\tau}{N}}. \tag{18}$$

We here consider a lower limit for the inspection frequency $N_\alpha$ that makes the fraction defective of products equal to $100\alpha\%$ or less for a prespecified value of $\alpha$ $(0 < \alpha < 1)$. It is convenient to introduce $u$ defined by

$$u = \frac{\tau}{N}, \quad N = 1, 2, \cdots, \tag{19}$$

and then we have

$$Q_2(u) = Q_1(u), \quad u \in (0, \tau]. \tag{20}$$

Hence, $Q_2(u)$ is strictly increasing in $u$ from 0 to $Q_2(\tau)$. Consequently, if

$$Q_2(\tau) = Q_1(\tau) = 1 - \frac{\int_0^{\tau} \overline{F}(t)dt}{\tau} > \alpha, \tag{21}$$

then there exists an upper limit $u_\alpha(> 0)$ satisfying $Q_2(u) \leq \alpha$ for a prespecified value of $\alpha$. This indicates that there exists a lower limit $N_\alpha(\geq 1)$ that satisfies $Q_2(N_\alpha) \leq \alpha$.

## 4.2 ECONOMICAL INSPECTION FREQUENCY

The long-run average cost of the proposed policy is, from the renewal reward theory, given by

$$C_2(N) = \frac{c_1 \int_0^{\frac{\tau}{N}} \left(\frac{\tau}{N} - t\right) dF(t) + c_2}{\frac{\tau}{N}}$$

$$= \frac{c_1 \int_0^{\frac{\tau}{N}} F(t)dt + c_2}{\frac{\tau}{N}}, \quad N = 0, 1, 2, \cdots, N-1. \tag{22}$$

Let us again introduce $u$ in Eq. (19), and we have

$$C_2(u) = C_1(u), \quad u \in (0, \tau]. \tag{23}$$

Hence if

$$R(\tau) = -\tau\overline{F}(\tau) + \int_0^\tau \overline{F}(t)dt > \frac{c_2}{c_1}, \tag{24}$$

there exists a unique $u^*$ minimizing $C_2(u)$ in relation to $u$, and therefore there exists a finite economical integer $N^*(\geq 1)$ that minimizes $C_2(N)$ with respect to $N$. If the inequality in (24) does not hold, we have $C'(u) \leq 0$ which signifies $u^* = \tau$, i.e., $N^* = 1$. This indicates that it is the optimum to conduct an inspection only just before we start our daily work of weighing products.

# 5. NUMERICAL EXAMPLES

This section assumes an exponential failure(malfunction) time distribution with failure rate $\lambda = 1/\mu$.

## 5.1 POLICY I

Under the exponential distribution, the fraction defective $Q_1(T)$ in Eq. (2) becomes

$$Q_1(T) = 1 - \frac{1 - e^{-\lambda T}}{\lambda T}, \quad T > 0, \tag{25}$$

and the long-run average cost $C_1(T)$ in Eq. (12) yields

$$C_1(T) = \frac{c_1\left(e^{-\lambda T} + \lambda T - 1\right) + c_2\lambda}{\lambda T}, \quad T > 0. \tag{26}$$

Table 1 shows values of inspection interval $T_\alpha$ for $\alpha = 0.01$, 0.05 and 0.1 in the case of $\lambda = 0.2(\mu = 5)$. Figure 1 indicates $C_1(T)$ for $c_2 = 1$ with $c_1 = 10$, 20, 30, 40 and 50, while Table 2 shows $T^*$ with $C_1(T^*)$. It is observed in Figure 1 and Table 2 that the economical inspection interval $T^*$ decreases with increasing $c_1$, which can intuitively explained.

Table 1: Inspection interval.

| $\alpha$ | 0.01 | 0.05 | 0.1 |
|---|---|---|---|
| $T_\alpha$ | 0.100 | 0.517 | 1.072 |

Table 2: Economical inspection interval.

| $c_1$ | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| $T^*$ | 1.07 | 0.74 | 0.6 | 0.52 | 0.46 |
| $C_1(T^*)$ | 1.93 | 2.76 | 3.40 | 3.93 | 4.40 |

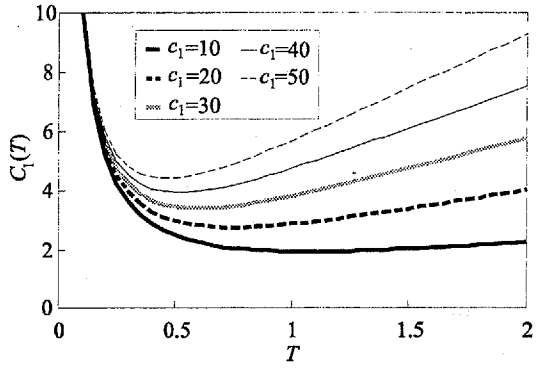


Figure 1: Long-run average cost.

## 5.2 POLICY II

Under the exponential distribution with failure rate $\lambda = 1/\mu$, the fraction defective $Q_2(T)$ in Eq. (15) becomes

$$Q_2(N) = 1 - \frac{1 - e^{-\lambda\tau/N}}{\lambda\tau/N}, \quad N = 1, 2, \cdots, \tag{27}$$

while the long-run average cost $C_2(N)$ in Eq. (22) becomes

$$C_2(N) = \frac{c_1\left(e^{-\lambda\tau/N} + \lambda\tau/N - 1\right) + c_2\lambda}{\lambda\tau/N}, \quad N = 1, 2, \cdots. \tag{28}$$

Table 3 shows values of inspection interval $N_\alpha$ for $\alpha = 0.01$, 0.05 and 0.1 in the case of $\lambda = 0.2(\mu = 5)$. Figure 2 depicts $C_2(N)$ for $c_2 = 1$ with $c_1 = 10$, 20, 30, 40 and 50, while Table 4 reveals $N^*$ along with $C_2(N^*)$. It is observed that the economical inspection frequency $N^*$ increases as $c_1$ becomes large, which can also be explained intuitively.

Table 3: Inspection frequency.

| $\alpha$ | 0.01 | 0.05 | 0.1 |
|---|---|---|---|
| $N_\alpha$ | 10 | 2 | 1 |

Table 4: Economical inspection frequency.

| $c_1$ | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| $N^*$ | 1 | 1 | 2 | 2 | 2 |
| $C_2(N^*)$ | 1.94 | 2.87 | 3.45 | 3.93 | 4.42 |

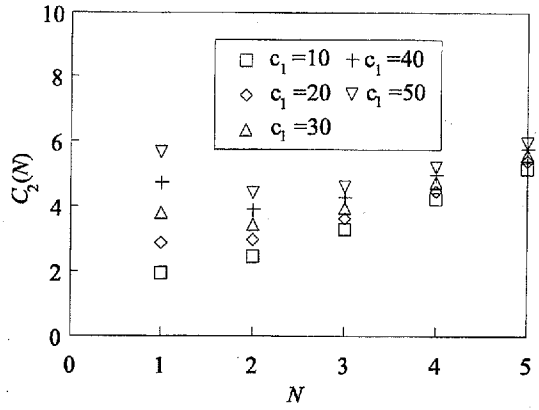


Figure 2: Long-run average cost.

# 6. CONCLUSIONS

In the final stage of manufacturing some specific products such as chemical ones, there is a process in which each product is weighed using a scale to mark each weight on the product. However, the scale occasionally undergoes its malfunction or failures. When the scale becomes out of order, it will show different weight for each product from the actual one, and hence each product will be shipped with a different mark from each actual weight.

This study focused on the products that are marked with wrong weight and regarded them as defectives. We then discussed two types of inspection policies for a scale. Under Policy I, we conduct an inspection to a scale at $iT(i = 1, 2, \cdots)$. Under Policy II, we considered a situation where the scale was inspected every morning before we started daily work of weighing products. For such a case, we considered to inspect the scale at $i\tau/N(i = 0, 1, 2, \cdots, N-1)$, where $\tau$ signified the working hours per day.

Two types of objective functions were considered; (1) the fraction defective and (2) the long-run average cost under each policy. Under Policy I (Policy II), we clarified the condition under which an finite inspection interval $T_\alpha$ (finite inspection frequency $N_\alpha$) exists, which guarantees that the fraction defective does not exceed a prespecified value $\alpha(0 < \alpha < 1)$. An economical inspection interval $T^*$ (inspection frequency $N^*$) minimizing the long-run average cost was then discussed. Numerical examples were presented to illustrate the proposed inspection policy formulations.

This study dealt with a case where an inspection involves adjustment operations, but there is a case where an inspection does not include adjustment activities. A model considering such a case will be discussed taking another opportunity.

# REFERENCES

1. R.E. Barlow and L.C. Hunter, Optimum preventive maintenance policies, *Operations Research,* **8**, 90-100 (1960).

2. G.H. Weiss, A problem in equipment maintenance, *Management Science,* **8**, 266-277 (1962).

3. R.E. Barlow and F. Proschan, *Mathematical Theory of Reliability,* Wiley, New York, (1965).

4. A.G. Munford and A.K. Shahani, A nearly optimal inspection policy, *Operational Research Quarterly,* **23**, 373-379 (1972).

5. A.G. Munford and A.K. Shahani, An inspection policy for the Weibull case, *Operational Research Quarterly,* **24**, 453-458 (1973).

6. S. Zacks and W.J. Fenske, Sequential determination of inspection epochs for reliability system with general lifetime distributions, *Naval Research Logistics Quarterly,* **20**, 377-386 (1973).

7. J.B. Keller, Optimum checking schedules for systems subject to random failure, *Management Science,* **21**, 256-260 (1974).

8. H. Luss and Z. Kander, A preparedness model dealing with $N$ systems operating simultaneously, *Operations Research,* **24**, 117-128 (1974).

9. H. Luss and Z. Kander, Inspection policies when duration of checking is non-negligible, *Operational Research Quarterly,* **25**, 299-309 (1974).

10. D. Anbar, An asymptotically optimal inspection policy, *Naval Research Logistics Quarterly,* **23**, 211-218 (1976).

11. W.G. Schneeweiss, On the mean duration of hidden faults in periodically checked systems, *IEEE Trans. on Reliability,* **R-25**, 346-348 (1976).

12. W.G. Schneeweiss, Duration of hidden faults in randomly checked systems, *IEEE Trans. on Reliability,* **R-26**, 328-330 (1977).

13. Z. Kander, Inspection policies for deteriorating equipment characterized by $N$ quality levels, *Naval Research Logistics Quarterly,* **25**, 243-255, (1978).

14. P.R. Tadikamalla, An inspection policy for the gamma failure distributions, *Naval Research Logistics Quarterly,* **25**, 243-255 (1979).

15. N. Wattanapanom and L. Shaw, Optimal inspection schedules for failure detection in a model where tests hasten failures, *Operations Research,* **27**, 303-317 (1979).

16. T. Nakagawa and K. Yasui, Approximate calculation of inspection policy with Weibull failure times, *IEEE Trans. on Reliability,* **R-28**, 403-404 (1979).

17. T. Nakagawa and K. Yasui, Approximate calculation of optimal inspection times, *Journal of Operational Research Society,* **31**, 851-853 (1980).

18. A. Gupta and H. Gupta, Optimal inspection policy for multistage production process with alternate inspection plans, *IEEE Trans. on Reliability,* **R-30**, 161-162 (1981).

19. T. Nakagawa, Periodic inspection policy with preventive maintenance, *Naval Research Logistics Quarterly,* **31**, 33-40 (1984).

20. N. Kaio and S. Osaki, Some remarks on optimum inspection policies, *IEEE Trans. on Reliability,* **R-33**, 277-279 (1984).

21. N. Kaio and S. Osaki, Analytical considerations on inspection policies, in Stochastic Models in Reliability Theory, (Edited by S. Osaki and Y. Hatoyama), pp. 53-71, Springer-Verlag, Heidelberg, (1984).

22. N. Kaio and S. Osaki, Optimal inspection policies: A review and comparison, *Journal of Mathematical Analysis and Applications,* **119**, 3-20 (1986).

23. N. Kaio and S. Osaki, Optimal inspection policy with two types of imperfect inspection Probabilities, *Microelectronics and Reliability,* **26**, 935-942 (1986).

24. N. Kaio and S. Osaki, Inspection policies: Comparisons, in *Reliability Theory and Applications,* (Edited by S. Osaki and J. Cao), pp. 140-147, World Scientific, Singapore, (1987).

25. B.J. Yum and E.D. McDowell, Optimal inspection policies in a serial production system including scarp rework and repair: A MILD approach, *International Journal of Production Research,* **25**, 1451-1464 (1987).

26. N. Kaio and S. Osaki, Inspection policies: Comparisons and modifications, *Recherche Opérationnelle / Operations Research,* **22**, 387-400 (1988).

27. N. Kaio and S. Osaki, Comparison of inspection policies, *Journal of the Operational Research Society,* **40**, 499-503 (1989).

28. D.J.D. Wijnmalen and J.A.M. Hontelez, Review of Markov decision algorithm for optimal inspections and revisions in a maintenance system with partial information, *European Journal of Operational Research,* **62**, 96-104 (1992).

29. C.G. Gassandras and Y. Han, Optimal inspection policies for a manufacturing station, *European Journal of Operational Research,* **63**, 35-53 (1992).

30. N. Kaio, T. Dohi and S. Osaki, Inspection policy with failure due to inspection, *Microelectronics and Reliability,* **34**, 599-602 (1994).

31. S. M. Ross, *Applied Probability Models with Optimization Applications,* Holden-Day, San Francisco, (1970).

32. S. M. Ross, *Introduction to Probability Models, 5th edition,* Academic Press, New York, (1993).

33. S. Osaki, *Applied Stochastic System Modeling,* Springer-Verlag, Berlin, (1992).

# OPTIMAL REPLACEMENT POLICIES FOR A TWO-UNIT SYSTEM WITH SHOCK DAMAGE INTERACTION

TAKASHI SATOW and SHUNJI OSAKI

Department of Industrial and Systems Engineering, Hiroshima University,
Higashi-Hiroshima 739-8527, Japan

{satow, osaki}@gal.sys.hiroshima-u.ac.jp

**Abstract**—In this paper we consider a two component system where component 1 failures occur according to a Poisson process. Each component 1 failure cause a random amount of damage to component 2 leading to its failure when the total damage exceeds a specified level. We study a two parameter maintenance policy which minimize the expected cost per unit of time for infinite time operation.

**Keywords**—Shock damage, Cumulative damage, Optimal policy, Control limit

## 1. INTRODUCTION

There is a vast literature on the maintenance of unreliable systems. Bulk of them deal with single component system. Valdez-Flores and Feldman [1] present a comprehensive review where references to earlier review papers can be found. In contrast, the maintenance of multi-component systems has received less attention and is an area of considerable research activity. Most of the models deal with the case where the component failures are independent. For a review of maintenance models for multi-component system, see Thomas [2], Cho and Parlar [3] and Dekker [4].

In a multi-component system, the failure times are often stochastically dependent [5]. Özekici [6] deals with optimal periodic replacement policy with statistically dependent failure times. Murthy and Nguyen [7] deal with a formulation where failure of a component has an effect on one or more of remaining components. They call this "failure interaction" and suggest two different types (Types 1 and 2) of interactions. In Type 1 failure interaction, a natural failure of a component can induce the failure (call "induced" failure) of one or more of remaining components. In Type 2 failure interaction, the failure affects the performance (e.g., the failure rate) of one or more of the remaining components. Murthy and Wilson [8] discusses the estimation problem for Type 1 failure interaction model with different data structures. Nakagawa and Murthy [9] deals with a two component (labelled components 1 and 2) system. Whenever component 1 fails, it causes a random amount of damage to component 2. The damage accumulates and leads to component 2 failure when it exceeds a specified level $K$. Component 1 failures occur according to a non-homogeneous Poisson process and are rectified through minimal repair. They considered two maintenance policies (one and two parameter policies) and derived conditions for the optimal parameters for the policies.

This paper deals with a two component system with component failures as in Nakagawa and Murthy [9]. We formulate a maintenance policy involving two parameters ("2-parameter policy") and derive an expression for the expected cost per unit time for infinite time operation. We give an optimal policy to minimize the expected cost per unit of time. We examine two special cases of this policy by letting one of the parameters assume their upper limits. The 1-parameter poli-

cies correspond to some well known policies studied by earlier researchers. These include the age replacement [10] and the control limit policies with additive damage [11] and [12].

The outline of this paper is as follows. In section 2, we give the details of the model formulation. Section 3 deals with the analysis of the 2-parameter maintenance policy. The special cases of this policy are considered in the next section. Section 4 deals with three 1-parameter maintenance policies. Section 5 deals with some numerical examples and we conclude with some comments in Section 6.

## 2. MODEL DESCRIPTION

We consider a system composed of two components (denoted as units 1 & 2). Unit 1 is repairable and it undergoes minimal repair at failure. The time to repair is small so that it can be ignored. As a result, unit 1 failures occur according to a nonhomogeneous Poisson process with intensity function $r(t)$ and a mean-value function $R(t)$, i.e., $R(t) \equiv \int_0^t r(u)du$, and $r(t)$ is increasing in $t$. Let $S_j$ $(j = 0, 1, 2, \cdots)$ be the random variable denoting the occurrence time of $j^{th}$ unit 1 failure with $S_0 = 0$. Then, the probability that $j$ or more unit 1 failures occurring during $(0, t]$ is given by

$$H_j(t) \equiv \mathrm{P_r}\{S_j \leq t\} = \sum_{i=j}^{\infty} \frac{[R(t)]^i}{i!} e^{-R(t)} \quad (j = 0, 1, 2, \cdots). \tag{1}$$

Let $N(t)$ be the total number of unit 1 failures by time $t$. The probability that exactly $j$ failures occur until time $t$ is given by

$$F_j(t) \equiv \mathrm{P_r}\{N(t) = j\} = H_j(t) - H_{j+1}(t) \quad (j = 0, 1, 2, \cdots). \tag{2}$$

Whenever unit 1 fails, it causes a random amount of damage $\{Y_j\}$ $(j = 0, 1, 2, \cdots)$ to unit 2. $Y_j$ is a sequence of identical and independent r.v with distribution $G(x)$, *i.e.*, $\mathrm{P_r}\{Y_j \leq x\} = G(x)$. The damage is additive and let $Z_j$ be the damage after the $j^{th}$ failure of unit 1 with $Z_0 = 0$ . Then $Z_j$ is a cumulative process (see [13]) with $Z_j \equiv \sum_{i=1}^{j} Y_i$ $(j = 1, 2, \cdots)$ and

$$\mathrm{P_r}\{Z_j \leq x\} \equiv G^{(j)}(x) \quad (j = 0, 1, 2, \cdots), \tag{3}$$

where $G^{(j)}(x)$ is the $j$-fold Stieltjes convolution of $G(x)$ with itself, and $G^{(0)}(x) \equiv 1$ for $x \geq 0$, and 0 for $x < 0$ .

Unit 2 fails whenever the total damage exceeds a failure level $K$. A system failure occurs whenever unit 2 fails because both units fail simultaneously. We assume that unit 2 is not repairable and as a result, a failed system needs to be replaced by a new one. Note that such replacements are unplanned replacements.

A system failure, in general, results in a high cost. One way of reducing this cost is to replace the system preventively, based on some policy, which reduces the likelihood of system failure. From a cost point of view, a preventive replacement is cheaper than failure replacement. However, a preventive replacement implies discarding some useful life of the system. Hence, preventive replacement needs to be done in a manner which achieves a suitable trade off between this loss versus the risk of a failure. We consider the following 2-parameter policy:

The system is replaced through a failure replacement when unit 2 fails (which corresponds to the damage for unit 2 exceeding $K$) or earlier through a preventive replacement when one of the following conditions occur:

**(i)** system reaches an age $T$,

**(ii)** the total damage to unit 2 exceeds a level $k$ $(< K)$.

Note that the policy is characterized by two parameters $(T, k)$ with $0 < T < \infty$, $0 < k < K$. When these two parameters assume their upper limits, then there is no preventive replacement and the system is replaced only on failure.

Let $C(T, k)$ denote the expected cost per unit time for infinite operation. Then the optimal parameters of the policy are $T^*$ and $k^*$ which yield a minimum value for $C(T, k)$.

We have a family of 1-parameter policies by allowing two of the parameter to assume their upper limits. As a result, we have the following two 1-parameter policies:

**Policy 1a:** $T \to \infty$. In this case, the policy is characterized by $k$.

**Policy 1b:** $k \to K$. In this case, the policy is characterized by $T$.

For the analysis of these policies, we make the following simplifying assumptions:

**1)** The failures of unit 1 and 2 are detected immediately.

**2)** The damage to unit 2 is measured after each failure of unit 2.

**3)** The time to repair unit 1 and replace the system is small so that they can be approximated as being zero. In other words, the repair or replacements are instantaneous.

**4)** The cost of each minimal repair for unit 1 is $c_m$. The cost of each failure [preventive] replacement for the system is $c_f$ $[c_p]$ with $c_f > c_p > c_m$.

Finally, for a continuous distribution function $G(x)$, let $\overline{G}(x) = 1 - G(x) = P_r\{Y_j > x\}$ and $g(x) = dG(x)/dx$, are the survivor and density functions associated with $G(x)$ respectively.

# 3. ANALYSIS OF THE 2 PARAMETER POLICY

## 3.1 THE EXPECTED COST PER UNIT TIME

Note that the system gets renewed with each failure or preventive replacement. As a result, the time interval between two successive renewals defines a cycle for a renewal process. From the renewal reward theorem [13] $C(T, k)$, the expected cost per unit time for infinite time operation, can be expressed as the ratio of the expected cycle cost and the expected cycle length. We proceed to obtain the expressions for these two quantities.

The probability $\alpha(K)$, that the system is replaced at failure of unit 2 (due to total damage exceeding $K$), is given by

$$\begin{aligned} \alpha(K) &= \sum_{j=1}^{\infty} \sum_{l=1}^{j} \mathrm{P_r}\{N(T) = j, Z_{l-1} \leq k, Z_l > K\} \\ &= \sum_{j=0}^{\infty} H_{j+1}(T) \int_0^k \overline{G}(K - x) dG^{(j)}(x). \end{aligned} \tag{4}$$

The probability $\beta(T)$, that the system is replaced preventively at age $T$, is given by

$$\beta(T) = \sum_{j=0}^{\infty} \mathrm{P_r}\{N(T) = j, Z_j \leq k\} = \sum_{j=0}^{\infty} F_j(T) G^{(j)}(k). \tag{5}$$

Finally, the probability $\gamma(k)$, that the system is replaced preventively when the total damage of unit 2 exceeds $k$ and is less than or equal to $K$, is given by

$$\begin{aligned} \delta(k) &= \sum_{j=1}^{\infty} \sum_{l=1}^{j} \mathrm{P_r}\{N(T) = j, Z_{l-1} \leq k < Z_l \leq K\} \\ &= \sum_{j=0}^{\infty} H_{j+1}(T) \int_0^k [G(K-x) - G(k-x)] dG^{(j)}(x). \end{aligned} \tag{6}$$

It is easily seen that eqn(4)+eqn(5)+eqn(6)=1.

The expected cost per cycle is made up of corrective and preventive maintenance cost. Corrective maintenance means the failure replacement of unit 2 and minimal repair of unit 1. The expected number of minimal repairs over a cycle, $\delta(T,k)$, is given by

$$\begin{aligned} \delta(T,k) &= \sum_{j=0}^{\infty} j F_j(T) G^{(j)}(k) + \sum_{j=1}^{\infty} j H_{j+1}(T) \int_0^k \overline{G}(K-x) dG^{(j)}(x) \\ &+ \sum_{j=1}^{\infty} j H_{j+1}(T) \int_0^k [G(K-x) - G(k-x)] dG^{(j)}(x) \\ &= \sum_{j=1}^{\infty} H_j(T) G^{(j)}(k). \end{aligned} \tag{7}$$

Using the above expressions, the expected cycle cost is given by

$$A(T,k) = c_f \alpha(K) + c_p[\beta(T) + \gamma(k)] + c_m \delta(T,k), \tag{8}$$

where $c_p$ and $c_f$ are the preventive replacement cost and failure replacement cost for the system, and $c_m$ is the cost of each minimal repair for unit 1.

The expected cycle length is given by

$$\begin{aligned} T \sum_{j=0}^{\infty} F_j(T) G^{(j)}(k) &+ \sum_{j=1}^{\infty} \int_0^T t dH_j(t) \int_0^k \overline{G}(K-x) dG^{(j-1)}(x) \\ &+ \sum_{j=1}^{\infty} \int_0^T t dH_j(t) \int_0^k [G(K-x) - G(k-x)] dG^{(j-1)}(x) \\ &= \sum_{j=0}^{\infty} G^{(j)}(k) \int_0^T F_j(t) dt. \end{aligned} \tag{9}$$

From eqns(8) and (9) and renewal reward theorem, the expected cost per unit time, for infinite operation, is given by

$$C(T,k) = \frac{A(T,k)}{\sum_{j=0}^{\infty} G^{(j)}(k) \int_0^T F_j(t)dt}. \tag{10}$$

## 3.2 OPTIMAL POLICY FOR 2-PARAMETERS

$T^*$ and $k^*$ are the values which minimizes $C(T,k)$ given by eqn(10). The optimal $T^*$ and $k^*$ can be obtained from the first order conditions, i.e., setting the derivatives of $C(T,k)$ with respect to $T$ and $k$ to zero. We assume that $r(t) = \lambda$. This implies that failures of unit 1 occur according to a stationary Poisson process. Differentiating $C(T,k)$ with respect to $T$ and setting it to zero, yields

$$\sum_{j=0}^{\infty} H_{j+1}(T)B_j(k) - \sum_{j=0}^{\infty} H_{j+1}(T)G^{(j)}(k) \frac{\sum_{j=0}^{\infty} F_j(T)B_j(k)}{\sum_{j=0}^{\infty} F_j(T)G^{(j)}(k)} = c_p, \tag{11}$$

where

$$B_j(k) \equiv \int_0^k [(c_f - c_p)G(K-x) - c_m G(k-x)]dG^{(j)}(x). \tag{12}$$

Denote the left-hand side of eqn(11) by $J(T;k)$.

Differentiating $C(T,k)$ with respect to $k$ and setting it to zero, yields

$$(c_f - c_p)\sum_{j=0}^{\infty} H_{j+1}(T) \int_0^k [G(K-x) - G(K-k)]dG^{(j)}(x)$$

$$+c_m \left[ \sum_{j=0}^{\infty} H_{j+1}(T)G^{(j)}(k) \frac{\sum_{j=1}^{\infty} H_j(T)g^{(j)}(k)}{\sum_{j=1}^{\infty} H_{j+1}(T)g^{(j)}(k)} - \sum_{j=1}^{\infty} H_j(T)G^{(j)}(k) \right] = c_p. \tag{13}$$

Denote the left-hand side of eqn(13) by $Q(k;T)$.

On comparing $J(T;k)$ with $Q(k;T)$, we see that $Q(k;T)$ is always greater than $J(T;k)$ for $({}^{\forall}T,\ (0 < T < \infty),\ {}^{\forall}k,\ (0 < k \leq K))$, as

$$J(T;k) - Q(k;T) =$$

$$(c_f - c_p)\sum_{j=0}^{\infty} H_{j+1}(T)G^{(j)}(k)\left\{\frac{\sum_{j=0}^{\infty} F_j(T)\int_0^k [G(K-k)-G(K-x)]dG^{(j)}(x)}{\sum_{j=0}^{\infty} F_j(T)G^{(j)}(k)}\right\}$$

$$+c_m\sum_{j=0}^{\infty} H_{j+1}(T)G^{(j)}(k)\left[\frac{\sum_{j=0}^{\infty} F_j(T)G^{(j+1)}(k)}{\sum_{j=0}^{\infty} F_j(T)G^{(j)}(k)} - \frac{\sum_{j=1}^{\infty} H_j(T)g^{(j)}(k)}{\sum_{j=1}^{\infty} H_{j+1}(T)g^{(j)}(k)}\right] < 0. \tag{14}$$

This implies that there does not exist $(T^*, k^*)$ which satisfies eqns(11) and (13) simultaneously.

# 4. 1-PARAMETER POLICIES

In this section, we consider the special case where one of the parameters assume their upper limits so that the policy is characterized by a single parameter. As indicated earlier, we have two different cases to consider.

## 4.1 POLICY 1a: CONTROL LIMIT POLICY ($k$)

The system is replaced preventively when the total damage of unit 2 exceeds $k$ or on system failure should it occur earlier. As a result, from eqn(10), the expected cost per unit time for infinite time operation is given by

$$C(\infty,k) \equiv \lim_{T\to\infty} C(T,k) = \frac{c_f + (c_p - c_f)\sum_{j=1}^{\infty}\int_0^k G(K-x)dG^{(j-1)}(x) + (c_f - c_p + c_m)M(k)}{\sum_{j=0}^{\infty} G^{(j)}(k)\int_0^{\infty} F_j(t)dt} \tag{15}$$

$k^*$, the optimal level $k$ which minimizes $C(\infty,k)$, can be obtained by differentiating $C(\infty,k)$ with respect to $k$ and setting it equal to zero. This yields

$$\sum_{j=1}^{\infty}\int_0^k [G(K-x)-G(K-k)]dG^{(j-1)}(x) = \frac{c_p - c_m}{c_f - c_p}. \tag{16}$$

Denote the left-hand side of eqn(16) by $V(k)$. Note that

$$\frac{dV(k)}{dk} = g(K-k)[1+M(k)] > 0. \tag{17}$$

where $M(K) \equiv \sum_{j=1}^{\infty} G^{(j)}(K)$. Note that

$$\lim_{k\to 0} V(k) = 0 \text{ and } \lim_{k\to K} V(k) = M(K). \tag{18}$$

As a result, if

$$M(K) > \frac{c_p - c_m}{c_f - c_p}, \tag{19}$$

then there exists a finite and unique $k^*$ which satisfies eqn(16). In this case, the optimal expected cost per unit time is given by

$$C(\infty, \infty, k^*) = \lambda[(c_f - c_p)\overline{G}(K - k^*) + c_m], \tag{20}$$

If eqn(16) is not satisfied for $0 < k < K$ then $k^* = K$. This implies that the optimal policy is no preventive replacement, and in this case the expected cost per unit time is given by

$$C(\infty, K) = \lambda \left[\frac{c_f + c_m M(K)}{1 + M(K)}\right]. \tag{21}$$

## 4.2 POLICY 1b: AGE POLICY ($T$)

The system is replaced preventively at time $T$ or on system failure should it occur earlier. As a result, from eqn(11), the expected cost per unit time for infinite time operation is given by

$$C(T, K) \equiv \lim_{k \to K} C(T, k) = \frac{c_f + (c_p - c_f)\sum_{j=0}^{\infty} F_j(T)G^{(j)}(K) + c_m \sum_{j=1}^{\infty} H_j(T)G^{(j)}(K)}{\sum_{j=0}^{\infty} G^{(j)}(K) \int_0^T F_j(t)dt}. \tag{22}$$

We assume that $r(t) = \lambda$. Differentiating $C(T, K)$ with respect to $T$ and setting it to zero, yields

$$\sum_{j=1}^{\infty} H_j(T)G^{(j)}(K) - Q(T)\sum_{j=0}^{\infty} H_{j+1}(T)G^{(j)}(K) = \frac{c_p}{c_f - c_p - c_m}, \tag{23}$$

where

$$Q(T) = \left[\frac{\sum_{j=0}^{\infty} F_j(T)G^{(j+1)}(K)}{\sum_{j=0}^{\infty} F_j(T)G^{(j)}(K)}\right]. \tag{24}$$

We need to consider the following three cases.

**(i)** $c_f - c_p - c_m > 0$ ; Let $U(T)$ denote the left-hand side of eqn(23) and let $Q(\infty) = \lim_{T \to \infty} Q(T)$. If $Q(T)$ is strictly decreasing then $U(T)$ is strictly increasing from 0 to $U(\infty)$. As a result, if

$$\lim_{T \to \infty} U(T) = M(K) - Q(\infty)[1 + M(K)] > \frac{c_p}{c_f - c_p - c_m}, \tag{25}$$

then there exists a finite and unique $T^*$ which satisfies eqn(23). If $G^{(j+1)}(K)/G^{(j)}(K)$ is strictly decreasing in $j$ then $Q(T)$ is strictly decreasing in $T$.

The optimal expected cost per unit time is given by

$$C(T^*, \infty, K) = \lambda[c_f - c_p - (c_f - c_p - c_m)Q(T^*)]. \tag{26}$$

If eqn(23) is not satisfied for $0 < T < \infty$ then $T^* = \infty$ and the expected cost per unit time is given by eqn(21).

**(ii)** $c_f - c_p - c_m < 0$ ; If $U(T)$ is increasing then $C(T,K)$ is decreasing in $T$; therefore the optimal $T^* \to \infty$. The expected cost per unit time is given by eqn(21). When $U(T)$ is not increasing, we need to use a numerical method to obtain $T^*$.

**(iii)** $c_f - c_p - c_m = 0$ ; In this case, $C(T,K)$ can be rewritten as follows.

$$C(T,K) = \frac{c_f + c_m \left[\sum_{j=1}^{\infty} H_j(T) G^{(j)}(K) - \sum_{j=0}^{\infty} F_j(T) G^{(j)}(K)\right]}{\sum_{j=0}^{\infty} G^{(j)}(K) \int_0^T F_j(t)dt}. \tag{27}$$

Differentiating $C(T,K)$ with respect to $T$, we have

$$\frac{dC(T,K)}{dT} = \frac{(c_m - c_f) \sum_{j=0}^{\infty} F_j(T) G^{(j)}(K)}{\left[\sum_{j=0}^{\infty} G^{(j)}(K) \int_0^T F_j(t)dt\right]^2}. \tag{28}$$

Since $c_m < c_f$, $C(T,K)$ is decreasing in $T$. Therefore, the optimal $T^* \to \infty$. In other words, no preventive replacement is the optimal policy. The expected cost per unit time is given by eqn(21).

## 5. NUMERICAL EXAMPLES

Let $G(x)$ be an exponential distribution with mean $1/\mu$, i.e., $G(x) = 1 - e^{-\mu x}$.

For Policy 1a, the optimal $k^*$ can be obtained from (16) and this can be rewritten as

$$\overline{G}(K - k^*) M(k^*) = \frac{c_p - c_m}{c_f - c_p}. \tag{29}$$

For Policy 1b, $T^*$ to be finite and unique requires (25) to be satisfied and this can be rewritten as

$$M(K) > \frac{c_p}{c_f - c_p - c_m}. \tag{30}$$

Note that $G^{(j+1)}(x)/G^{(j)}(x)$ is strictly decreasing in $j$ when $G(x)$ is an exponential distribution. Therefore, if eqn(30) is satisfied then a finite $T^*$ exists and is unique.

We assume the following values for the model parameters, $\lambda = 1$ (Mean time to failure for unit 1 is 1.) $\mu = 1$ (Mean damage caused to unit 2 by each unit 1 failure is 1.)

Let $c_m = 1$. We consider a range of value for $c_p$ (varying from 2 to 30) and $c_f$ (varying from 10 to 50). Table 1 and 3 give the optimal $k^*$ (for Policy 1a) and $T^*$ (for Policy 1b) for $c_p = 5$; two values of $K$ (= 100 & 200) and a range of value for $c_f$. Similar results for $c_f = 50$ and $c_p$ varying are given in Tables 2 and 4. Also the optimal expected costs per unit time are given.

Table 1 . Policy 1a
OPTIMAL $k^*$, $C(\infty, k^*)$
$\lambda = 1, \mu = 1, c_p = 5, c_m = 1$

| | $K = 100$ | | $K = 200$ | |
|---|---|---|---|---|
| $c_f$ | $k^*$ | $C(\infty, k^*)$ | $k^*$ | $C(\infty, k^*)$ |
| 10 | 95.2 | 1.0419 | 194.5 | 1.0205 |
| 20 | 94.1 | 1.0424 | 193.4 | 1.0206 |
| 30 | 93.6 | 1.0427 | 192.9 | 1.0207 |
| 40 | 93.2 | 1.0428 | 192.5 | 1.0207 |
| 50 | 93.0 | 1.0429 | 192.3 | 1.0207 |

Table 2 . Policy 1a
OPTIMAL $k^*$, $C(\infty, k^*)$
$\lambda = 1, \mu = 1, c_f = 50, c_m = 1$

| | $K = 100$ | | $K = 200$ | |
|---|---|---|---|---|
| $c_p$ | $k^*$ | $C(\infty, k^*)$ | $k^*$ | $C(\infty, k^*)$ |
| 2 | 91.6 | 1.0109 | 190.8 | 1.0052 |
| 8 | 93.6 | 1.0746 | 192.9 | 1.0362 |
| 15 | 94.5 | 1.1479 | 193.8 | 1.0721 |
| 20 | 94.9 | 1.1999 | 194.2 | 1.0977 |
| 30 | 95.8 | 1.3026 | 195.0 | 1.1486 |

Table 3 . Policy 1b
OPTIMAL $T^*$, $C(T^*, K)$
$\lambda = 1, \mu = 1, c_p = 5, c_m = 1$

| | $K = 100$ | | $K = 200$ | |
|---|---|---|---|---|
| $c_f$ | $T^*$ | $C(T^*, K)$ | $T^*$ | $C(T^*, K)$ |
| 10 | 83.8 | 1.065 | 171.9 | 1.030 |
| 20 | 74.9 | 1.071 | 160.0 | 1.032 |
| 30 | 71.9 | 1.074 | 155.9 | 1.033 |
| 40 | 70.1 | 1.075 | 153.5 | 1.033 |
| 50 | 68.9 | 1.077 | 151.8 | 1.034 |

Table 4 . Policy 1b
OPTIMAL $T^*$, $C(T^*, K)$
$\lambda = 1, \mu = 1, c_f = 50, c_m = 1$

| | $K = 100$ | | $K = 200$ | |
|---|---|---|---|---|
| $c_p$ | $T^*$ | $C(T^*, K)$ | $T^*$ | $C(T^*, K)$ |
| 2 | 64.7 | 1.032 | 145.8 | 1.014 |
| 8 | 71.5 | 1.119 | 155.5 | 1.053 |
| 15 | 76.2 | 1.212 | 161.8 | 1.097 |
| 20 | 79.2 | 1.273 | 165.8 | 1.126 |
| 30 | 86.0 | 1.384 | 174.6 | 1.182 |

The optimal results for 2-parameter policy is identical to Policy 1a.

One would expect the optimal expected cost per unit time for the two parameter policy to be smaller than that for the one parameter policies. The numerical results indicate that this is not so. The reason for this apparent counter unintuitive result is as follows.

The state of component 2 is best indicated by the total damage ($Z(t)$). If this information is not available then the age ($t$) of component 2 is the best indicator. In other words, given $Z(t)$, then $t$ does not provide any extra information.

$$I(Z(t), t) = I(Z(t)) \tag{31}$$

where I represents the information about the state of component 2. Also, it is worth noting that

$$I(t) \subset I(Z(t))$$

This can be seen from Tables 1-4 where for a given set of parameter values, the optimal expected cost per unit time for Policy 1a which is based on $Z(t)$ is smaller than that for Policy 1b.

The results of section 3 showed that 2-parameter policy is no better than the better of Policy 1a. Since Policy 1a is better than Policy 1b, we see that Policy 2a does not perform better than Policy 1a. This is to be expected since the age of component provides no new information.

# 6. CONCLUSION

In this paper we considered a two component system where component 1 failures occur according to a Poisson process and cause damage to component 2. The damage is accumulated and component 2 fails when the total damage exceeds a specified limit.

We derived an expression for the expected cost per unit time for a two parameter policy. The policy reduces to two one-parameter polies as special cases. We give analytical characterization to obtain the optimal parameter value for these special cases. In the process we obtained an apparent counter unintuitive result and gave an explanation for it.

The results of the paper highlight indicate an important issue, i.e., increasing the number of parameters does not necessary lead to lower expected costs. The parameters used provide information about the state of one or more components of the system.

The important issue is whether the information provided by a parameter (A) is contained in that provided by another parameter (B). If so, then the parameter (A) provide no new information and hence will not lead to lower expected costs.

This issue has not received sufficient attention in the maintenance literature as attested by the number of two parameter policies that have been developed which perform no better than the one parameter policies. There is scope for further study of this issue.

## REFERENCES


1. Valdez-Flores, C. and Feldman, R. M., A Survey of Preventive Maintenance Models for Stochastically Deteriorating Single-Unit Systems, *Naval. Res. Logist. Quart.* **36**. 419-446 (1989).
2. Thomas, L.C., A Survey of Maintenance and Replacement Models for Maintainability and Reliability of Multi-Item Systems, *Rel. Eng.* **16**. 297-309 (1986).
3. Cho, D. and Parlar, M., A Survey of Maintenance Models for Multi-Unit Systems, *European J. Operational Research.* **51**. 1-23 (1991).
4. Dekker, R., Applications of Maintenance and Optimization Models: A Review and Analysis, *Rel. Eng. and Sys. Safety.* **51**. 229-240 (1996).
5. Shaked, M. and Shanthikumar, J. G., Reliability and Maintenability, *In Stochastic Models*, Heymen.D.P and Sobel.M.J. Eds., North Holland, Amsterdam. (1990).
6. Özekici, S., Optimal Periodic Replacement of Multicomponent Reliability Systems, *Operations Research.* **36**. 542-552 (1988)
7. Murthy, D. N. P. and Nguyen, D. G., Study of Two-Component System with Failure Interaction, *Naval. Res. Logist. Quart.* **32**. 239-247 (1985).
8. Murthy, D. N. P. and Wilson, R. J., Parameter Estimation in Multi-Component System with Failure Interaction, *Stochastic Models and Data Analysis.* **10**. pp.47-60 (1994)
9. Nakagawa, T. and Murthy, D. N. P., Optimal Replacement Policies for a Two-Unit System with Failure Interactions, *RAIRO* . **27**. 427-438 (1993).
10. Barlow, R. E. and Proschan, F., Mathematical Theory of Reliability, Wiley, New York (1965)
11. Taylor, H. M., Optimal Replacement under Additive Damage and Other Failure Models, *Naval. Res. Logist. Quart.* **22**. 1-18 (1975)
12. Abdel Hameed, M and Shimi, I. N., Optimal Replacement of Damage Devices, *J. Appl. Prob.*. **15**. 153-161 (1978)
13. Ross, S.M., Applied Probability Models with Optimization Applications, Holden-Day San Francisco, (1970)

# ON RELATIONSHIP BETWEEN SOFTWARE AVAILABILITY MEASUREMENT AND THE NUMBER OF RESTORATIONS

KOICHI TOKUNO and SHIGERU YAMADA

Department of Social Systems Engineering, Faculty of Engineering, Tottori University,
4-101, Koyama, Tottori-shi, 680-8552 Japan

toku@sse.tottori-u.ac.jp / yamada@sse.tottori-u.ac.jp

**Abstract**—In this paper, we propose a software availability model considering the number of restoration actions. We correlate the failure and restoration characteristics of the software system with the cumulative number of corrected faults. Furthermore, we consider an imperfect debugging environment where the detected faults are not always corrected and removed from the system. The time-dependent behavior of the system alternating between up and down states is described by a Markov process. From this model, we can derive quantitative measures for software availability assessment based on the number of restoration actions. Finally, we show numerical examples of software availability analysis.

**Keywords**—Software availability, Imperfect debugging, Software reliability growth, Markov process, Quantitative assessment.

## 1. INTRODUCTION

Many methodologies for software reliability measurement and assessment have been discussed for the last few decades [1]–[4]. A mathematical software reliability model is often called a **software reliability growth model** (SRGM); this describes a software fault-detection or a software failure-occurrence phenomenon during the testing phase of software development process and the operation phase. A software failure is defined as an unacceptable departure from program operation caused by a fault remaining in the software system. This model is available for measuring and assessing the degree of achievement of software reliability, deciding the time to software release for operational use, and estimating the maintenance cost for faults undetected during the testing phase.

Most of SRGMs so far provide quantitative software reliability measures for developers. However, it begins to be necessary to assess software systems from the viewpoint of customers. In particular, recent systems are required nonstop operation and utilities. One of the customer-oriented attributes is **software availability** [5]–[7]; this is defined as the attribute that the software systems are performing at a given time point, according to the specification, under the specified environment. In other words, it represents the property that the systems are in available states whenever the customers want to use them. Few mathematical models for evaluating software availability are proposed.

In this paper, we construct a software availability model. Quantitative measures on reliability derived from previous SRGMs, such as the mean time between software failures and the software reliability representing the probability that the system can continue to operate for a given time period, are often provided as the functions of the number of software failures or fault detections, and useful for seizing the relationship between the number of detected faults and software reliability growth. On the other hand, there are scarcely software availability measures for explicitly understanding the relation with the number of restoration actions. Here we discuss software availability

measurement considering the number of restoration actions. The time-dependent behavior of the software system is described by a Markov process [8]. The software failure and the restoration characteristics are correlated with the cumulative number of corrected faults. Furthermore, we also describe the imperfect debugging environment where the debugging activities corresponding to software failure-occurrences are not always performed for certain [9]. The assumptions and modeling are detailed in Sect. 2. Derivation of the stochastic quantities for software availability measurement is presented in Sect. 3. Numerical illustrations of software availability analysis are shown in Sect. 4. Finally, concluding remarks of this paper are summarized in Sect. 5.

## 2. MODEL DESCRIPTION

The following assumptions are made for software availability modeling:

A1. The software system is unavailable and starts to be restored as soon as a software failure occurs, and the system can not operate until the restoration action is complete.

A2. The restoration action implies the debugging activity; this is performed perfectly with probability $a$ $(0 < a \leq 1)$ and imperfectly with probability $b(= 1 - a)$. We call $a$ the perfect debugging rate [9]. One fault is corrected and removed from the software system when the debugging activity is perfect.

A3. The next time intervals of software failures and restorations when $n$ faults have already been corrected from the system, follow exponential distributions with means $1/\lambda_n$ and $1/\mu_n$, respectively.

A4. The probability that two or more software failures occur simultaneously is negligible.

Consider a stochastic process $\{X(t),\ t \geq 0\}$ whose state space is $(\boldsymbol{W},\ \boldsymbol{R})$, where up state vector $\boldsymbol{W} = \{W_n;\ n = 0,\ 1,\ 2,\ \ldots\}$ and down state vector $\boldsymbol{R} = \{R_n;\ n = 0,\ 1,\ 2,\ \ldots\}$ [10]. Then, the events $\{X(t) = W_n\}$ and $\{X(t) = R_n\}$ mean that the system is operating and inoperable due to the restoration action at time point $t$, when $n$ faults have already been corrected, respectively.

From assumption [2], when the restoration action has been complete in $\{X(t) = R_n\}$,

$$X(t) = \begin{cases} W_n & \text{(with probability } b\text{)} \\ W_{n+1} & \text{(with probability } a\text{).} \end{cases} \tag{1}$$

We use Moranda's model [11] to describe the software failure-occurrence phenomenon, i.e., when $n$ faults have been corrected, the hazard rate $\lambda_n$ is given by

$$\lambda_n = Dk^n \quad (n = 0,\ 1,\ 2,\ \ldots;\ D > 0,\ 0 < k < 1), \tag{2}$$

where $D$ and $k$ are the initial hazard rate and the decreasing ratio of the hazard rate, respectively. The expression of (2) comes from the viewpoint that software reliability depends on the debugging efforts, not the residual fault content. We do not note how many faults remain in the software system. Equation (2) describes a software failure-occurrence phenomenon where a software system has high frequency of software failure-occurrence during the early stage of the testing or the operation phase and it gradually decreases after then [4], [9]. Early software availability models such as those of Okumoto and Goel [12] and Kim et al. [13] often assume that the hazard rate is proportional to the residual fault content and decreases by a constant amount with the perfect debugging, i.e., $\lambda_n$ is described as

$$\lambda_n = \phi(N - n) \quad (n = 0,\ 1,\ 2,\ \ldots,\ N - 1;\ N > 0,\ \phi > 0), \tag{3}$$

where $N$ and $\phi$ are the initial fault content and the hazard rate per fault remaining in the system, respectively [14]. $X(t)$ forms a finite-state Markov process if (3) is applied to $\lambda_n$.

Next, we describe the time-dependent behavior of the restoration action. The restoration action for software systems includes not only the data recovery and the program reload but also the debugging activities for manifested faults. From the viewpoint of the fault complexity, there are cases where the faults detected during the early stage of the testing or the operation phase have low complexity and are easily corrected/removed, and as the testing is in progress, detected faults have higher complexity and are more difficult of correction/removal [1], [15]. In the above case, it is appropriate that the mean restoration-time becomes longer with the increasing number of corrected faults. Accordingly, we express $\mu_n$ as follows:

$$\mu_n = Er^n \quad (n = 0,\ 1,\ 2,\ \ldots;\ E > 0,\ 0 < r \leq 1), \tag{4}$$

where $E$ and $r$ are the initial restoration rate and the decreasing ratio of the restoration rate, respectively. In the case of $r = 1$, i.e., $\mu_n = E$ means that the complexity of each fault is random [10].

Let $Q_{A,B}(\tau)$ $(A,\ B \in (\boldsymbol{W},\ \boldsymbol{R}))$ denote the one-step transition probability that after making a transition into state $A$, the process $\{X(t),\ t \geq 0\}$ makes a transition into state $B$ by time $\tau$. The expressions for $Q_{A,B}(\tau)$'s are given as follows:

$$Q_{W_n,R_n}(\tau) = 1 - e^{-\lambda_n \tau}, \tag{5}$$

$$Q_{R_n,W_{n+1}}(\tau) = a(1 - e^{-\mu_n \tau}), \tag{6}$$

$$Q_{R_n,W_n}(\tau) = b(1 - e^{-\mu_n \tau}). \tag{7}$$

The sample state transition diagram of $X(t)$ is illustrated in Fig. 1.

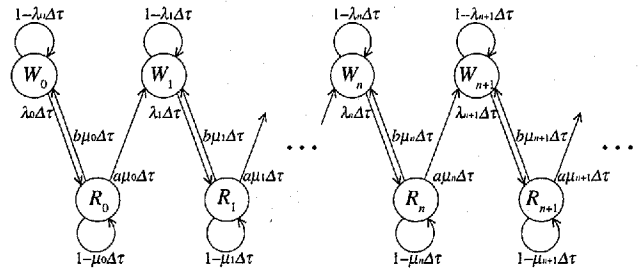


Figure 1: A diagrammatic representation of state transitions between $X(t)$'s.

# 3. DERIVATION OF SOFTWARE AVAILABILITY MEASURES

## 3.1 DISTRIBUTION OF TRANSITION TIME BETWEEN UP STATES

Let $S_{i,n}$ $(i \leq n)$ be the random variable representing the transition time from state $W_i$ to state $W_n$, and $G_{i,n}(t)$ be the distribution function of $S_{i,n}$, respectively. Then, we obtain the following renewal equation with respect to $G_{i,n}(t)$:

$$G_{i,n}(t) = Q_{W_i,R_i} * Q_{R_i,W_{i+1}} * G_{i+1,n}(t) + Q_{W_i,R_i} * Q_{R_i,W_i} * G_{i,n}(t), \tag{8}$$

where $*$ denotes a Stieltjes convolution and $G_{n,n}(t) = 1(t)$ (step function, $n = 0,\ 1,\ 2,\ \ldots$).

We use Laplace-Stieltjes (L-S) transforms [8] to solve (8), where the L-S transform of $G_{i,n}(t)$ is defined as

$$\widetilde{G}_{i,n}(s) \equiv \int_0^\infty e^{-st} dG_{i,n}(t). \tag{9}$$

Substituting the L-S transforms of (5)–(7) into that of (8) yields

$$\begin{aligned} \widetilde{G}_{i,n}(s) &= \frac{a\lambda_i\mu_i}{s^2 + (\lambda_i + \mu_i)s + a\lambda_i\mu_i}\widetilde{G}_{i+1,n}(s) \\ &= \frac{x_i y_i}{(s + x_i)(s + y_i)}\widetilde{G}_{i+1,n}(s), \end{aligned} \tag{10}$$

where

$$\left.\begin{matrix} x_i \\ y_i \end{matrix}\right\} = \frac{1}{2}\left[(\lambda_i + \mu_i) \pm \sqrt{(\lambda_i + \mu_i)^2 - 4a\lambda_i\mu_i}\right] \quad \text{(double signs in same order)}. \tag{11}$$

By solving (10) recursively, we obtain $\widetilde{G}_{i,n}(s)$ as

$$\begin{aligned} \widetilde{G}_{i,n}(s) &= \prod_{m=i}^{n-1} \frac{x_m y_m}{(s + x_m)(s + y_m)} \\ &= \sum_{m=i}^{n-1} \left[\frac{A^1_{i,n}(m)x_m}{s + x_m} + \frac{A^2_{i,n}(m)y_m}{s + y_m}\right], \end{aligned} \tag{12}$$

where

$$A^1_{i,n}(m) = \frac{\displaystyle\prod_{j=i}^{n-1} x_j y_j}{x_m \displaystyle\prod_{\substack{j=i \\ j\neq m}}^{n-1}(x_j - x_m)\prod_{j=i}^{n-1}(y_j - x_m)} \quad (m = i,\ i+1,\ \ldots,\ n-1), \tag{13}$$

$$A^2_{i,n}(m) = \frac{\displaystyle\prod_{j=i}^{n-1} x_j y_j}{y_m \displaystyle\prod_{\substack{j=i \\ j\neq m}}^{n-1}(y_j - y_m)\prod_{j=i}^{n-1}(x_j - y_m)} \quad (m = i,\ i+1,\ \ldots,\ n-1), \tag{14}$$

respectively. By inverting (12), we obtain the distribution function of $S_{i,n}$ as

$$\begin{aligned} G_{i,n}(t) &\equiv \Pr\{S_{i,n} \le t\} \\ &= 1 - \sum_{m=i}^{n-1}\left[A^1_{i,n}(m)e^{-x_m t} + A^2_{i,n}(m)e^{-y_m t}\right] \quad (n = 1,\ 2,\ \ldots;\ i = 0,\ 1,\ 2,\ \ldots,\ n). \end{aligned} \tag{15}$$

It is noted that

$$\sum_{m=i}^{n-1} \left[ A_{i,n}^{1}(m) + A_{i,n}^{2}(m) \right] = 1. \tag{16}$$

Furthermore, the expectation and the variance of $S_{i,n}$ are given by

$$\mathrm{E}[S_{i,n}] = \sum_{m=i}^{n-1} \left( \frac{1}{x_m} + \frac{1}{y_m} \right), \tag{17}$$

$$\mathrm{Var}[S_{i,n}] = \sum_{m=i}^{n-1} \left( \frac{1}{{x_m}^2} + \frac{1}{{y_m}^2} \right), \tag{18}$$

respectively.

## 3.2 STATE OCCUPANCY PROBABILITY

Let $P_{A,B}(t)$ $(A,\ B \in (\boldsymbol{W},\ \boldsymbol{R}))$ be the state occupancy probability that the system is in state $B$ at time point $t$ on the condition that the system was in state $A$ at time point $t = 0$, i.e.,

$$P_{A,B}(t) \equiv \Pr\{X(t) = B | X(0) = A\} \quad (A,\ B \in (\boldsymbol{W},\ \boldsymbol{R})). \tag{19}$$

We obtain the following renewal equations with respect to $P_{W_i,W_n}(t)$:

$$P_{W_i,W_n}(t) = G_{i,n} * P_{W_n,W_n}(t), \tag{20}$$

$$P_{W_n,W_n}(t) = e^{-\lambda_n t} + Q_{W_n,R_n} * Q_{R_n,W_n} * P_{W_n,W_n}(t). \tag{21}$$

From (21), the L-S transform of $P_{W_n,W_n}(t)$ is given by

$$\begin{aligned} \widetilde{P}_{W_n,W_n}(s) &= \frac{s(s+\mu_n)}{(s+x_n)(s+y_n)} \\ &= \left( \frac{s}{a\lambda_n} + \frac{s^2}{a\lambda_n\mu_n} \right) \frac{x_n y_n}{(s+x_n)(s+y_n)}. \end{aligned} \tag{22}$$

Substituting (22) into the L-S transform of (20) yields

$$\widetilde{P}_{W_i,W_n}(s) = \frac{s\widetilde{G}_{i,n+1}(s)}{a\lambda_n} + \frac{s^2\widetilde{G}_{i,n+1}(s)}{a\lambda_n\mu_n}. \tag{23}$$

By inverting (23), $P_{W_i,W_n}(t)$ is obtained as

$$\begin{aligned} P_{W_i,W_n}(t) &\equiv \Pr\{X(t) = W_n | X(0) = W_i\} \\ &= \frac{g_{i,n+1}(t)}{a\lambda_n} + \frac{g'_{i,n+1}(t)}{a\lambda_n\mu_n}, \end{aligned} \tag{24}$$

where $g_{i,n}(t)$ is the probability density function associated with $G_{i,n}(t)$ and $g'_{i,n}(t) \equiv dg_{i,n}(t)/dt$.

Using the similar procedure for the derivation of $P_{W_i,W_n}(t)$, we obtain the following renewal equations with respect to $P_{W_i,R_n}(t)$:

$$P_{W_i,R_n}(t) = G_{i,n} * Q_{W_n,R_n} * P_{R_n,R_n}(t), \tag{25}$$

$$P_{R_n,R_n}(t) = e^{-\mu_n t} + Q_{R_n,W_n} * Q_{W_n,R_n} * P_{R_n,R_n}(t). \tag{26}$$

Substituting the L-S transform of (26) into that of (25) yields

$$\widetilde{P}_{W_i,R_n}(s) = \frac{s\widetilde{G}_{i,n+1}(s)}{a\mu_n}. \tag{27}$$

By inverting (27), $P_{W_i,R_n}(t)$ is obtained as

$$\begin{aligned} P_{W_i,R_n}(t) &\equiv \Pr\{X(t) = R_n | X(0) = W_i\} \\ &= \frac{g_{i,n+1}(t)}{a\mu_n}. \end{aligned} \tag{28}$$

### 3.3 SOFTWARE AVAILABILITY

Once we specify integer $i$, the following equation holds for arbitrary time $t$:

$$\sum_{n=i}^{\infty} [P_{W_i,W_n}(t) + P_{W_i,R_n}(t)] = 1. \tag{29}$$

Here we consider the relationship between the number of the restoration actions and software availability measurement. Let $l = 0, 1, 2, \ldots$ denote the number of the restoration actions. Furthermore, we introduce the binary indicator variable $I_A(t)$ taking the value 1 (0) if the system is operating (inoperable) at time point $t$, given that it was in state $A \in (\boldsymbol{W}, \boldsymbol{R})$ at time point $t = 0$, respectively. Then $A_i(t) \equiv \Pr\{I_{W_i}(t) = 1\}$ $(i = 0, 1, 2, \ldots)$ denotes the instantaneous software availability given that the system was in state $W_i$ at time point $t = 0$, i.e.,

$$\begin{aligned} A_i(t) &= \sum_{n=i}^{\infty} P_{W_i,W_n}(t) \\ &= 1 - \sum_{n=i}^{\infty} P_{W_i,R_n}(t), \end{aligned} \tag{30}$$

(see Fig. 2). It is noted that the cumulative number of corrected faults at the completion of the $l$-th restoration action, $C_l$, is not explicitly observed since imperfect debugging is assumed throughout this paper. However, $C_l$ follows the binomial distribution whose probability mass function is given by

$$\Pr\{C_l = i\} = \binom{l}{i} a^i b^{l-i} \quad (i = 0, 1, 2, \ldots, l), \tag{31}$$

where $\binom{l}{i} \equiv l!/[(l-i)!i!]$ denotes a binomial coefficient. Accordingly, the instantaneous software availability after the completion of the $l$-th restoration action is given by

$$A(t; l) = \sum_{i=0}^{l} \Pr\{C_l = i\} A_i(t), \tag{32}$$

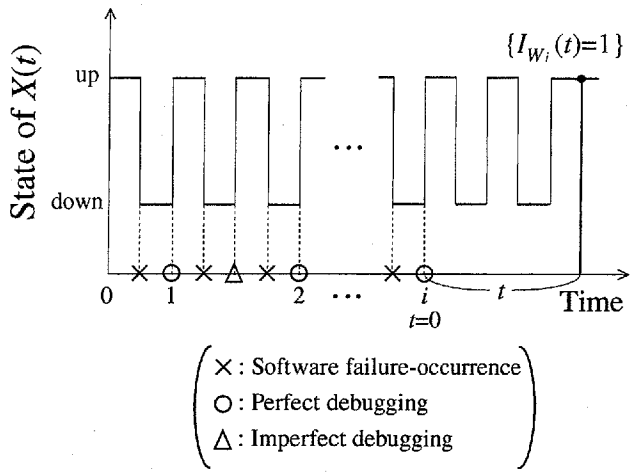


Figure 2: Sample behavior of the system and event $\{I_{W_i}(t) = 1\}$.

which represents the probability that the system is operating at time point $t$, given that the $l$-th restoration action was complete at time point $t = 0$. Furthermore, the average software availability after the completion of the $l$-th restoration action is given by

$$A_{av}(t; l) = \frac{1}{t} \int_0^t A(x; l) dx, \tag{33}$$

which represents the ratio of the system's operating time to the time interval $(0, t]$, given that the $l$-th restoration action was complete at time point $t = 0$. In particular using (28), we can express (32) and (33) as

$$A(t; l) = 1 - \sum_{i=0}^{l} \binom{l}{i} a^i b^{l-i} \sum_{n=i}^{\infty} \frac{g_{i,n+1}(t)}{a\mu_n}, \tag{34}$$

$$A_{av}(t; l) = 1 - \frac{1}{t} \sum_{i=0}^{l} \binom{l}{i} a^i b^{l-i} \sum_{n=i}^{\infty} \frac{G_{i,n+1}(t)}{a\mu_n}, \tag{35}$$

respectively.

# 4. NUMERICAL EXAMPLES

Using the software availability model discussed above, we show numerical illustrations for software availability measurement and assessment.

We define the maintenance factor as

$$\begin{aligned} \rho_n &\equiv \lambda_n/\mu_n \\ &= Cv^n \quad (C \equiv D/E,\ v \equiv k/r), \end{aligned} \tag{36}$$

where we call $C$ and $v$ the initial maintenance factor and the availability improvement parameter, respectively.

Figure 3 shows the time-dependent behavior of the average software availability, $A_{av}(t;l)$ in (35) for various numbers of the restoration actions, $l$, in the case of $v < 1$. This figure indicates that software availability drops rapidly immediately after the beginning of re-operation and then gradually increases. We can also see that software availability improves with the increasing number of the restoration actions.

Figures 4 and 5 show the instantaneous software availability, $A(t;l)$ in (34) for various values of perfect debugging rate, $a$, in the cases of $v < 1$ and $v > 1$, respectively. These figures tell us that the software availability becomes higher (lower) as the perfect debugging rate becomes larger when $v < 1$ ($v > 1$). The case of $v > 1$ may be a paradoxical result that the software availability decreases more slowly with decreasing $a$. This reasoning is that software availability is related to the ratio of the software failure time to the restoration time rather than the software failure time itself, i.e., $\rho_n$ increases more slowly with decreasing $a$ since smaller $a$ means that the cumulative number of corrected faults is more difficult to increase.

In the case of $v < 1$, we can find the minimum number of restoration actions, $l_{min}$, satisfying that the minimum value of $A(t;l)$ or $A_{av}(t;l)$ exceeds the prespecified availability objective, $\alpha$. Table 1 summarizes $l_{min}$'s on $A(t;l)$ and $A_{av}(t;l)$ for various values of $a$, in the case of $\alpha = 0.95$. As shown in Table 1, the higher certainty of debugging attains the objective of software availability earlier.

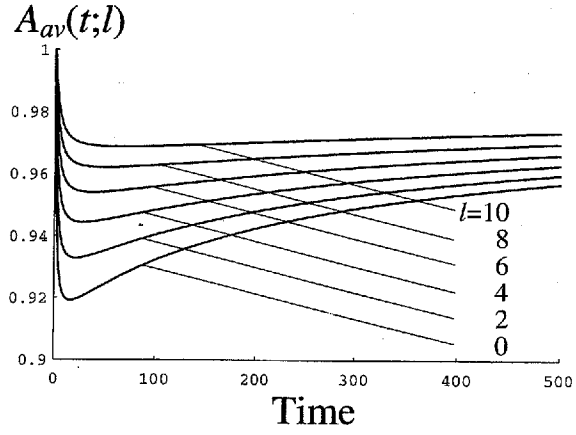


Figure 3: Dependence of number of restoration actions on $A_{av}(t;l)$ ($a = 0.9$, $D = 0.1$, $k = 0.8$, $E = 1.0$, $r = 0.9$).

Table 1: $l_{min}$ on $A(t;l)$ and $A_{av}(t;l)$ ($\alpha = 0.95$; $D = 0.1$, $k = 0.8$, $E = 1.0$, $r = 0.9$).

| $a$ | $l_{min}$ on $A(t;l)$ | $l_{min}$ on $A_{av}(t;l)$ |
|---|---|---|
| 1.0 | 6 | 5 |
| 0.9 | 6 | 6 |
| 0.8 | 7 | 6 |
| 0.7 | 8 | 7 |
| 0.6 | 9 | 9 |

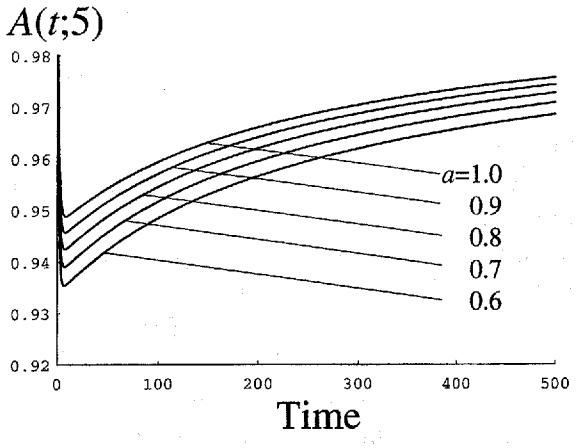


Figure 4: Dependence of perfect debugging rate on $A(t;l)$ in case of $v < 1$ ($l = 5,\ D = 0.1,\ k = 0.8,\ E = 1.0,\ r = 0.9$).

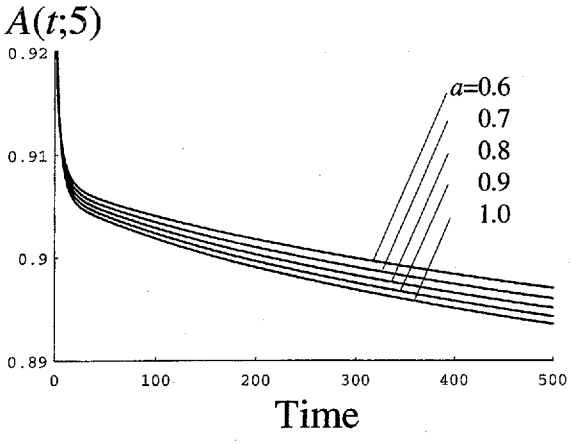


Figure 5: Dependence of perfect debugging rate on $A(t;l)$ in case of $v > 1$ ($l = 5,\ D = 0.1,\ k = 0.909,\ E = 1.0,\ r = 0.9$).

## 5. CONCLUDING REMARKS

In this paper, we have developed a stochastic model describing the relationship between the number of restoration actions and software availability measurement. We have used a Markov process for the description of the behavior of the system alternating between operable and inoperable states. We have derived the instantaneous and the average software availability considering the number of restoration actions. Numerical illustrations for software availability measurement have also been presented to show that these measures are very useful for software performance assessment. This model has been more generalized in terms of the imperfect debugging and the fault complexity than several previous models.

The unknown parameters must be estimated based on the actual data for assessing software availability with this model. But it is difficult to observe and collect the testing or the field data. In particular, it is necessary to equip the collection procedure of the restoration times. Establish of practical estimation of the model parameters remains a future study.

## ACKNOWLEDGMENTS

This work was supported in part by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, and Culture of Japan under Grant. No. 10680431.

## REFERENCES

1. J. D. Musa, A. Iannino, and K. Okumoto, *Software Reliability: Measurement, Prediction, Application*, McGraw-Hill, New York (1987).

2. S. Yamada, Software quality/reliability measurement and assessment: Software reliability growth models and data analysis, *J. Information Processing*, **14**, 254–266 (1991).

3. S. Yamada, *Software Reliability Models: Fundamentals and Applications*, JUSE Press, Tokyo (1994).

4. M. R. Lyu, ed., *Handbook of Software Reliability Engineering*, IEEE Computer Society Press, Los Alamitos, CA (1996).

5. J.-C. Laprie, K. Kanoun, C. Béounes, and M. Kaâniche, The KAT (Knowledge-Action-Transformation) approach to the modeling and evaluation of reliability and availability growth, *IEEE Trans. Software Eng.*, **17**, 370–382, (1991).

6. J.-C. Laprie and K. Kanoun, X-ware reliability and availability modeling, *IEEE Trans. Software Eng.*, **18**, 130–147, (1992).

7. K. Tokuno and S. Yamada, A summary of Markovian software availability modeling, In *Proc. Fifth ISSAT International Conf. Reliability and Quality in Design*, (Edited by H. Pham and M.-W. Lu), pp. 218–222, (1999).

8. S. Osaki, *Applied Stochastic System Modeling*, Springer-Verlag, Heidelberg, (1992).

9. S. Yamada, K. Tokuno, and S. Osaki, Software reliability measurement in imperfect debugging environment and its application, *Reliability Engineering and System Safety*, **40**, 139–147 (1993).

10. K. Tokuno and S. Yamada, Markovian software availability modeling for performance evaluation", In *Stochastic Modelling in Innovative Manufacturing: Proceedings, Cambridge, U.K., July 21-22, 1995*, (Edited by A. H. Christer, S. Osaki, and L. C. Thomas), pp. 246–256, Springer-Verlag, Berlin, (1997).

11. P. B. Moranda, Event-altered rate models for general reliability analysis, *IEEE Trans. Reliability*, **R-28**, 376–381 (1979).

12. K. Okumoto and A. L. Goel, Availability and other performance measures for system under imperfect maintenance, In *Proc. COMPSAC '78*, pp. 66–71, (1978).

13. J. H. Kim, Y. H. Kim, and C. J. Park, A modified Markov model for the estimation of computer software performance, *Operations Research Letters*, **1**, 253–257 (1982).

14. Z. Jelinski and P. B. Moranda, Software reliability research, In *Statistical Computer Performance Evaluation*, (Edited by W. Freiberger), pp. 465–484, Academic Press, New York, (1972).

15. Y. Nakagawa and I. Takenaka, Error complexity model for software reliability estimation, *Trans. IEICE*, **J74-D-I**, 379–386 (1991).