Revised Optimal Checkpointing Strategies in Database Availability Modeling and Their Comparison

土肥 正[†], 海生 直人[‡], 尾崎 俊治[†]
DOHI Tadashi, KAIO Naoto, OSAKI Shunji
[†]広島大学工学部, [‡]広島修道大学経済科学部

Abstract: In this paper, a revised checkpoint institution method is proposed for a database availability model. In the method, the checkpointing is always performed after rollback recovery actions when the system failure occurs as well as in the periodic (pre-scheduled) checkpoint interval. The revised optimal checkpointing strategies which maximize the system availability are derived and compared with existing ones numerically.

Keywords: optimal checkpointing, rollback recovery, database availability modeling, data maintainability, stochastic models, comparison.

1. Introduction

Checkpointing and rollback recovery are commonly used techniques for improving the reliability/maintainability of fault-tolerant computing systems. Especially, in file or database systems, checkpoint generations play an important role to limit the amount of data processing for the recovery actions after system failures occur. If the generations of checkpoints are frequently executed, a larger overhead for them will be incurred. Conversely, if only a few checkpoints are generated, the rollback recovery (RR) actions after system failures will require a large overhead. Hence, it is very important, from the practical point of view, to determine the optimal checkpoint interval taking account of the trade-off between the two kinds of overheads above.

A number of efforts of determining the optimal checkpoint sequence have been devoted in the literature. First, Young [1] obtained the optimal checkpoint interval for the computation restart after system failures. Chandy *et al.* [2, 3], Gelenbe *et al.* [4, 5], Baccelli [6], Kulkarni *et al.* [7], Nicola and Van Spanje [8] and Grassi *et al.* [9] discussed several evaluation models for database recovery and proposed the optimal checkpoint interval which maximizes or minimizes the system availability or the expected overhead during the normal operation. Sumita, Kaio and Goes [10], Goes and Sumita [11] and Goes [12] extended mathematically the standard checkpoint model by Gelenbe [5] and gave more general models in that non-homogeneous Poisson failures are incorporated with the RR procedure dependent of the cumulative operation time since the last checkpoint. Gelenbe and Hernandez [13] and Dohi *et al.* [14] presented methods for obtaining the optimal checkpoint interval of a different transaction processing computer system subject to time-dependent failures.

This paper deals with a similar stochastic database availability model to Sumuita, Kaio and Goes [11], but assumes that the checkpointing is always performed after RR actions when the system failure occurs as well as in the periodic (pre-scheduled) checkpoint interval. We derive the optimal checkpointing strategies under the different checkpoint institution method, and compare with the existing ones in the literature [1, 3, 5]. This new policy will make an overhead itself by checkpointing increase, but provide intuitively satisfied RR effects.

2. Existing Models without Immediate Checkpointing After RR (Model 1)

Following Gelenbe [5], consider a probabilistic model for a RR with checkpoint generations. The possible realization of the stochastic system under consideration is depicted in Fig. 1. System failures occur according to a homogeneous Poisson process whose intensity is λ (> 0). More specifically, let $X(t)$ be the cumulative operation time for the database system at time t since the last checkpoint. Then the probability that a system failure on the primary memory

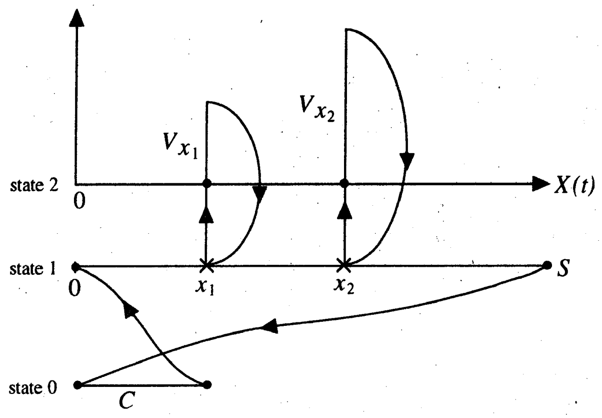

Figure 1: Possible realization of Model 1.

occurs in the time interval $(x, x + \Delta)$ is given by $\lambda x \Delta + o(\Delta)$. Upon a failure, a RR takes place where the information of transactions saved at the last checkpoint creation and recorded in the log are used for restoring the database to a usable state. The length of the RR is assumed to depend on the number of transactions executed, *i.e.*, on the value of $X(t)$ at the time of failure.

We, extensively, employ a generic random variable V_x denoting the length of the RR given that a failure occured at time t with $X(t) = x$ (see [10]). The distribution of V_x is denoted by $B_x(y) = \Pr\{V_x \leq y\}$. Intervals between two consecutive checkpoints are determined by the total operation time in the interval excluding rollback periods. The i-th checkpoint is generated as soon as the total operation time since the $(i-1)$st checkpoint reaches the length S_i $(i = 1, 2, \cdots)$. Assume that S_i $(i = 1, 2, \cdots)$ constitute a sequence of random variables with common distribution $A(x) = \Pr\{S_i \leq x\}$. Times (overheads) required for creating checkpoints also form a sequence of i.i.d. random variables C_i $(i = 1, 2, \cdots)$ with $W(z) = \Pr\{C_i \leq z\}$.

Let S_i $(i = 1, 2, \cdots)$ be the actual time interval between the $(i-1)$st and the i-th checkpoints. Then, since S_i $(i = 1, 2, \cdots)$ is a sequence of i.i.d. random variables, checkpoints are clearly regenerative points. From the well-known renewal argument, it is sufficient to consider the system behaviour for one cycle and we drop the discrete time index i $(i = 1, 2, \cdots)$ in the following discussion. Since the one cycle is defined as the time period between commencing at the end of one checkpoint and ending another checkpoint, the mean time of one cycle is $\mathrm{E}_A[S] + \mathrm{E}_W[C] + R_1$, where R_1 is the total mean time required by the RR and is expressed by

$$R_1 = \int_0^\infty dA(x) \int_0^x \lambda \mathrm{E}_B[V_s] ds. \tag{1}$$

Then, the system availability (ergodic availability) is formulated as

$$\Pi_1 = \frac{\mathrm{E}_A[S]}{\mathrm{E}_A[S] + \mathrm{E}_W[C] + R_1}. \tag{2}$$

From Eq.(2), the problem is to seek the optimal checkpoint strategy which maximizes Π_1.

Let us derive the optimal checkpoint interval. Define the following functions;

$$h(x) \equiv \lambda \mathrm{E}_B[V_x], \tag{3}$$

$$H(x) \equiv \int_0^x h(s) ds. \tag{4}$$

The function $h(x)$ is called *switching function* in this paper and satisfies the following relationship;

$$\begin{aligned}\mathrm{E}_A[H(S)] &= R_1 = \int_0^\infty dA(x)\int_0^x h(s)ds \\ &= \int_0^\infty H(x)dA(x). \end{aligned} \tag{5}$$

Further, we define a set of all $A(x)$s with fixed expectation $T \in [0,\infty)$ as J_T. The following result proved by Gelenbe [5] will be useful to characterize the condition on the existence of the optimal checkpoint interval.

Lemma 2.1: Let J_T be a set of all $A(x)$s with fixed expectation $T \in [0,\infty)$. If the function $h(x)$ is increasing, then the element $A(x)$ of the set J_T which maximizes the system availability Π_1 is

$$A(x) = U(x-T) = \begin{cases} 1 & \text{if } x \geq T \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $U(\cdot)$ is the unit function.

From Lemma 2.1, the randomized policy $A(x)$ is translated to the constant policy T and we can replace Π_1 and $\mathrm{E}_A[S]$ to $\Pi_1(T)$ and T, respectively, that is, the problem can be reduced to the following algebraic one:

$$\max_{0<T<\infty} : \Pi_1(T) = \frac{T}{T + \mathrm{E}_W[C] + H(T)}. \tag{7}$$

Now, we specify the random variable V_x denoting the length of the RR. In accordance with the literature [1-14], put

$$\mathrm{E}_B[V_x] = \alpha x + \beta, \quad \alpha \geq 0, \beta \geq 0, \tag{8}$$

where the first term denotes the mean time necessary to re-execute transactions which were processed in time interval $[0,x]$ and the second term is a fixed time associated with reloading the information stored at the checkpoint back into primary memory. From Eq. (3), the switching function becomes

$$h(x) = \lambda(\alpha x + \beta). \tag{9}$$

Theorem 2.2: (i) Suppose that the $h(x)$ is strictly increasing. Then, there exists a finite and unique optimal checkpoint interval T^* ($0 < T^* < \infty$) satisfying the nonlinear equation and the corresponding system availability is

$$\Pi_1(T^*) = 1/\{1 + h(T^*)\}, \tag{10}$$

$$\mathrm{E}_W[C] + H(T^*) = T^* h(T^*). \tag{11}$$

(ii) *Suppose that the function $h(x)$ is constant ($\alpha = 0$). Then, the optimal checkpoint interval is $T^* \to \infty$, i.e., no checkpoint should be generated and the corresponding system availability is*

$$\Pi_1(\infty) = 1/\{1 + h(\infty)\}. \tag{12}$$

Next let us consider the problem to choice the parameters α and β. Following Young [1], Chandy *et al.* [3] and Gelenbe [5], we introduce three methods to evaluate the mean RR period for different operation circumstance.

Method 1 [1]: $\alpha = l$,

Method 2 [3]: $\alpha \approx \rho l$,

Method 3 [5]: $\alpha \approx l(1 - p_1^*)$,

where

$$p_1^* = 1 - \rho/\Pi_1(T) \tag{13}$$

is the ergodic conditional probability that the database is idle given that the system is operating (see Gelenbe [5]), l (> 0) is the rate of transactions to be re-executed after any failure and ρ (> 0) is the traffic intensity. Without loss of generality, we suppose the heavy traffic condition, *i.e.* $\rho < 1$. More concretely, consider the following situation; if transactions arrive at the system according to a homogeneous Poisson process with intensity ν (> 0), processing requirements of transactions for both initial processing and reprocessing are i.i.d. having a common exponential distribution with mean $1/\kappa (> 0)$, where $\rho = \nu/\kappa$. For example, Gelenbe [5] derived an approximate optimal checkpoint interval applying Method 3 as follows:

$$T^* \approx \frac{\mathrm{E}_W[C]}{1+\beta\lambda}\Big\{\Big(1 + \frac{2(1+\beta\lambda)}{\rho\lambda l \mathrm{E}_W[C]}\Big)^{1/2} - 1\Big\}. \tag{14}$$

In this way, Method 1 should be used when the system is continuously operating (heavy loaded). Methods 2 and 3 present the intermittent behaviour of the database system based on the arrival-service (queueing) process of transactions. These methods should be used corresponding to the situation to be modeled.

3. Revised Models with Immediate Checkpointing After RR (Model 2)

In this section, we suppose that the checkpointing is always performed after RR actions when the system failure occurs as well as in the periodic (pre-scheduled) checkpoint interval. In other words, the checkpoint interval is periodically determined through operation period measured by the calendar time in the previous models, but the present models take account of the age of the database system. Figure 2 illustrates the possible realization of the system behaviour. In a fashion similar to Model 1, the one cycle is defined as the time period between commencing at the end of one checkpoint and ending another checkpoint. The mean operating time for one cycle is

$$Q = \int_0^\infty e^{-\lambda x}\overline{A}(x)dx, \tag{15}$$

and the mean time length of one cycle is $Q + \mathrm{E}_W[C] + R_2$, where

$$R_2 = \int_0^\infty e^{-\lambda x}\lambda \mathrm{E}[V_x]\overline{A}(x)dx. \tag{16}$$

Then the system availability is defined by

$$\Pi_2 = \frac{Q}{Q + \mathrm{E}_W[C] + R_2}. \tag{17}$$

Lemma 3.1: Let J_T be a set of all $A(x)$s with fixed expectation $T \in [0, \infty)$. Then, the element $A(x)$ of the set J_T which maximizes the system availability Π_2 is $A(x) = U(x - T)$.

Next, we calculate new checkpointing strategies for Model 2. Define

$$q(T) = \int_0^T \mathrm{E}[V_x]\lambda e^{-\lambda x}dx$$

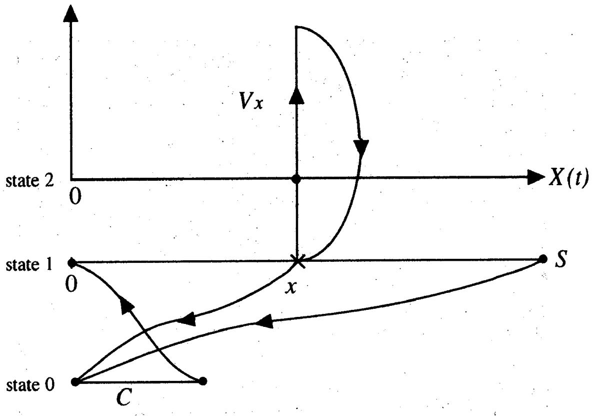

Figure 2: Possible realization of Model 2.

$$+\mathrm{E}_W[C] - h(T)\int_0^T e^{-\lambda x}dx \tag{18}$$

and $\mathrm{E}_B[V_x] = \alpha x + \beta$.

Theorem 3.2: Suppose that the function $h(x)$ is strictly increasing. (i) If $q(\infty) < 0$, then there exists a finite and unique optimal checkpoint interval T^* $(0 < T^* < \infty)$ satisfying the nonlinear equation $q(T^*) = 0$ and the corresponding system availability is

$$\Pi_2(T^*) = 1/\{1 + h(T^*)\}. \tag{19}$$

(ii) *If $q(\infty) \geq 0$, then the optimal checkpoint interval is $T^* \to \infty$.*

Similar to Section 2, we consider three cases:

Method 1: $\alpha = l$,

Method 2: $\alpha \approx \rho l$,

Method 3: $\alpha \approx l(1 - p_2^*)$,

where

$$p_2^* = 1 - \rho/\Pi_2(T) \tag{20}$$

is the ergodic conditional probability that the database is idle given that the system is operating for the present model.

4. Performance Comparison

Two models, Model 1 and Model 2, are compared numerically for each Method. In this section, we describe the numerical results on Method 3. Figure 3 presents the dependence of the optimal checkpoint interval in the failure rate λ. It is shown that the optimal checkpoint interval for Model 2 is always larger than that for Model 1. Also, Fig. 4 illustrates the corresponding behaviour of the system availability for varying λ. From these results, Model 1 without immediate checkpointing after system failures generates larger system availability than Model 2.

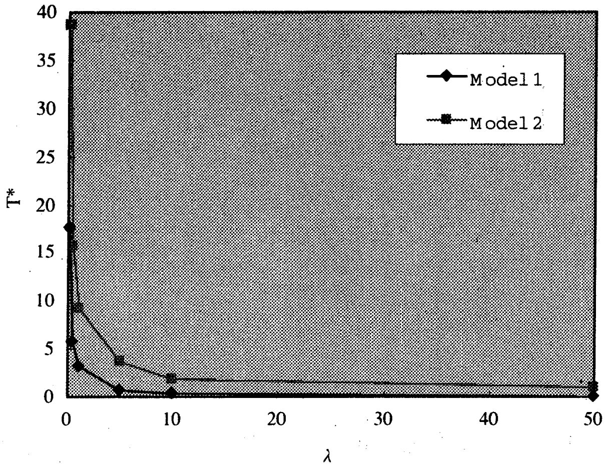

Figure 3: Behaviour of the optimal checkpoint interval for varying failure rate.

On the other hand, the behaviour of the ergodic probability that the system is in the rollback recovery is shown in Fig. 5. This result tells us that the recovery loss probability in the steady state for Model 1 is larger than Model 2 as the failure rate is increasing, when the optimal checkpoint interval maximizing the system availability is generated. Similarly, Fig. 6 is the befaviour of the probability that the checkpointing is executed in the steady state. Of our interest is the system operation for such a catastrophic case. From Figs. 4, 5 and 6, the optimal checkpointing strategies based on Model 1 makes the system availability increase, but the main contribution is to control an overhead by checkpointing itself. Conversely, Model 2 generates the frequent generations of checkpointing, but reduce the recovery loss probability.

Finally, if the data maintainability is the most important factor for file or database systems, Model 2 is preferable as a checkpoint institution method. If we would like to reduce the overhead by checkpointing as much as possible, the existing institution method such as Model 1 should be executed. Of course, both methods (Models) strictly maximize the corresponding system availabilities.

5. Concluding Remarks

In this paper, we have considered a revised checkpoint institution method. It is shown numerically that the model proposed reduces a rollback recovery probability in the steady state effectively. In addition, this new model describes checkpoint generations from the practical point of view and, at the same time, satisfies our intuition. The basic idea is due to Goes and Sumita [11], but it should be noted that their model is questionable since the checkpointing is executed only when the system failure does not occur during pre-determined time interval. That is to say, the existing methods by Gelenbe [5] *et al.* as well as Goes and Sumita [11] also determine the checkpointing schedule by only the calendar time, but our model is based on both the system age and the calendar time.

In future, further merits on the model proposed should be investigated. In particular, the comparison of checkpoint institution methods should be made under different cost criteria.

References

[1] J. W. Young, "A first order approximation to the optimum checkpoint interval", *Commun. ACM*, **17**, 9, 530-531 (1974).

[2] K. M. Chandy, "A survey of analytic models of roll-back and recovery strategies", *Computer*, **8**, 5, 40-47 (1975).

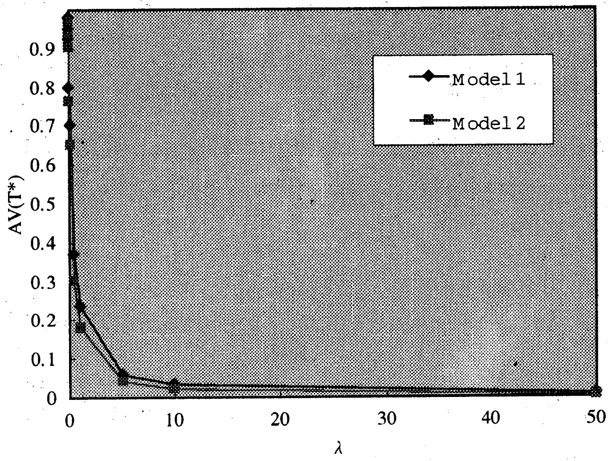

Figure 4: Behaviour of the system availability for varying failure rate.

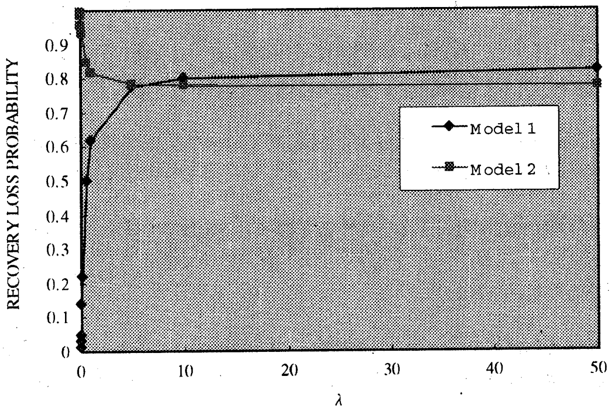

Figure 5: Behaviour of the recovery loss probability for varying failure rate.

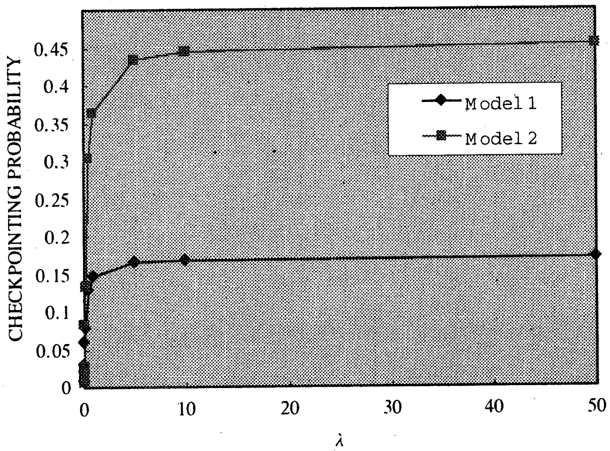

Figure 6: Behaviour of the checkpointing probability for varying failure rate.

[3] K. M. Chandy, J. C. Browne, C. W. Dissly and W. R. Uhrig, "Analytic models for rollback and recovery strategies", *IEEE Trans. Soft. Eng.*, **SE-1**, 1, 100-110 (1975).

[4] E. Gelenbe and D. Derochette, "Performance of rollback recovery systems under intermittent failures", *Commnun. ACM*, **21**, 6, pp. 493-499 (1978).

[5] E. Gelenbe, "On the optimum checkpoint interval", *J. ACM*, **26**, 2, 259-270 (1979).

[6] F. Baccelli, "Analysis of a service facility with periodic checkpointing", *Acta Informatica*, **15**, 67-81 (1981).

[7] V. G. Kulkarni, V. F. Nicola and K. S. Trivedi, "Effects of checkpointing and queueing on program performance", *Commun. Statist. Stochastic Models*, **6**, 4, 615-648 (1990).

[8] V. F. Nicola and J. M. Van Spanje, "Comparative analysis of different models of checkpointing and recovery", *IEEE Trans. Soft. Eng.*, **SE-16**, 8, 807-821 (1990).

[9] V. Grassi, L. Donatiello and S. Tucci, "On the optimal checkpointing of critical tasks and transaction-oriented systems", *IEEE Trans. Soft. Eng.*, **SE-18**, 1, 72-77 (1992).

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

[11] P. B. Goes and U. Sumita, "Stochastic models for performance analysis of database recovery control", *IEEE Trans. Computers*, **C-44**, 561-576 (1996).

[12] P. B. Goes, "A stochastic model for performance evaluation of main memory resident database systems", *ORSA J. Computing*, **7**, 3, 269-282 (1997).

[13] E. Gelenbe and M. Hernandez, "Enhanced availability of transaction oriented systems using failure tests", *Proc. Int'l Symp. Software Reliability Eng.*, pp. 342-350, IEEE CS Press, Los Alamitos, California (1992).

[14] T. Dohi, T. Aoki, N. Kaio and S. Osaki, "Computational aspects of optimal checkpoint strategy in fault-tolerant database management", *IEICE Transactions on Fundamentals* (in press).