JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 1: 170~175, Jan. 2022

즉각적 오류 감지가 가능한 경우의 체크포인팅 모형 분석

이유태*

Analysis of Checkpointing Model with Instantaneous Error Detection

Yutae Lee*

*Professor, Department of Information and Communications Engineering, Dong-eui University, Busan, 47340 Korea

요 약

고성능 컴퓨팅 분야에서 오류의 영향을 완화하기 위해 사후 장애 관리 기법이 필요하다. 일반적인 오류 복구 기법은 체크포인트 기법이다. 이 기법은 체크포인트를 설정해서 주기적으로 응용 프로그램의 상태를 저장했다가, 오류가 발생했을 때 오류 발생 이전 상태로 시스템을 복구하는 것이다. 본 논문에서는 오류 발생 시간이 독립이고 동일한 일반적인 분포를 따른다는 가정에서 즉각적으로 오류를 감지하는 경우의 체크포인팅 모형을 분석한다. 두 체크포인트 사이에 많아야 하나의 오류만 발생한다는 가정을 제거한다. 체크포인트 발생 시간, 고장 시간, 복구 시간 등이 주어질 때, 시스템의 신뢰도를 유도한다. 또한, 오류 발생 시간이 지수 분포를 따르는 경우에 최적의 체크 포인팅 시간 간격을 구한다.

ABSTRACT

Reactive failure management techniques are required to mitigate the impact of errors in high performance computing. Checkpoint is the standard recovery technique for coping with errors. An application employing checkpoints periodically saves its state, so that when an error occurs while some task is executing, the application is rolled back to its last checkpointed task and resumes execution from that task onward. In this paper, assuming the time-to-errors are independent each other and generally distributed, we analyze the checkpointing model with instantaneous error detection. The conventional assumption that two or more errors do not take place between two consecutive checkpoints is removed. Given the checkpointing time, down-time, and recovery time, we derive the reliability of the checkpointing model. When the time-to-error follows an exponential distribution, we obtain the optimal checkpointing interval to achieve the maximum reliability.

키워드 : 체크포인팅, 즉각적 오류 감지, 신뢰도, 수학적 분석

Keywords : Checkpointing, Instantaneous error detection, Reliability, Mathematical analysis

Received 17 November 2021, **Revised** 18 November 2021, **Accepted** 23 November 2021

*** Corresponding Author** Yutae Lee(E-mail:ylee@deu.ac.kr, Tel:+82-51-890-1682)
Professor, Department of Information and Communications Engineering, Dong-eui University, Busan, 47340 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.1.170 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

고성능 컴퓨팅 분야는 계산 능력의 향상으로 의학, 생물학, 화학, 항공 우주 과학 등 다양한 연구 분야에 도움이 되고 있다. 이 분야의 주요한 목표 중 하나는 효율적인 작업 실행을 보장하는 것으로 이는 전체 실행 시간을 줄이는 것에 해당한다. 신뢰할 수 있는 작업 실행도 성능에 결정적이다. 실제로 대규모 플랫폼에는 더욱더 오류가 발생하기 쉽다. 각각의 컴퓨팅 자원의 신뢰도가 높다 할지라도 전체 플랫폼은 오류가 빈번히 발생할 수 있다. 예를 들어, 각각의 컴퓨팅 자원의 MTBE(Mean Time Between Errors)가 10년이라고 하자. 이는 평균적으로 10년에 한 번의 오류가 발생한다는 의미이다. 어떤 플랫폼이 이러한 컴퓨팅 자원 10만개로 구성되었다면, 이 플랫폼은 평균적으로 50분마다 한 번의 오류가 발생할 것이다. 따라서 오류의 영향을 완화하기 위해 사후 장애 관리 기법이 필요하다.

일반적인 오류 복구 기법으로 체크포인트 기법이 있다[1,2]. 이 기법은 체크포인트를 설정해서 주기적으로 응용 프로그램의 상태를 저장했다가, 오류가 발생했을 때 시스템을 오류 발생 이전 생태로 복구하는 것이다[3,4]. 본 논문에서는 오류가 감지되었을 때, 고장 시간(down time)과 복구 시간(recovery time) 후에 가장 최근의 체크포인트 지점에서 작업을 다시 실행한다고 가정한다.

즉각적인 오류 감지는 컴퓨팅 자원에서의 장애 발생 등과 같은 경우에 나타날 수 있다. 최적의 체크포인팅 시간 간격을 주요 파라미터인 고장 시간, 체크포인트 발생 시간, 복구 시간 등의 함수로 나타낼 수 있다. Young [5]은 오류 발생 시간(time-to-error)이 지수 분포를 따른다는 가정을 하고, 최적의 체크포인팅 시간 간격에 대한 first order 근삿값을 처음으로 제시하였다. Young [5]은 이 연구에서 고장 시간과 복구 시간을 고려하지 않았다. 나중에 Daly [6]는 복구 시간을 고려하는 등 Young[5]의 결과를 개선하였다. 하지만, Young[5]과 Daly[6]의 결과에 오류가 있다는 것이 알려졌다. 오류 발생 시간이 임의의 확률 분포에 대한 dynamic programming heuristic은 Bouguerra et al.[7]에 의해 제안되었다.

본 논문은 오류 발생 시간이 독립이고 동일한 일반적인 분포를 따른다는 가정에서 즉각적으로 오류를 감지하는 경우의 체크포인팅 모형을 분석한다. 본 논문에서는 두 체크포인트 사이에 많아야 하나의 오류만 발생한다는 기존 연구들에서의 가정을 제거하였다. 체크포인트 발생 시간, 고장 시간, 복구 시간 등이 주어질 때, 시스템의 신뢰도를 유도하고, 오류 발생 시간이 지수 분포를 따르는 경우에 최적의 체크 포인팅 시간 간격을 구하였다.

Ⅱ. 체크포인트 모형

즉각적인 오류 감지(instantaneous error detection)를 하는 체크포인팅(checkpointing) 모형을 먼저 소개한다. 그림 1은 즉각적인 오류 감지를 하는 체크포인팅 모형을 나타낸다. 두 연속된 체크포인팅 시간 사이의 실제 유효한 작업 시간을 τ라 하자. 즉, 성공적인 체크포인팅이 완료된 후 시스템이 오류 없이 τ라는 시간 동안 정상적으로 동작한 후 다음 체크포인팅을 실행한다. 시간 τ 동안 작업을 성공적으로 수행하는데 필요한 전체 시간을 T라 하자. 이 시간 T에는 성공적으로 작업을 수행한 후 체크포인팅을 하는 시간을 포함한다. 시간 τ는 두 연속된 체크포인트 사이에 수행 작업이 신뢰할 만한 시간의 양을 나타내고, T는 두 체크포인트 사이의 전체 시간으로 오류 발생으로 인한 고장 시간과 복구 시간을 포함한다.

체크포인트를 생성하는데 걸리는 시간을 C라 하자. 오류가 없다면, 시간 T는 시간 τ와 체크포인트 발생 시간 C를 더한 것과 같다. 작업은 일련의 예상치 못한 오류에 의해 중단될 수 있다. 오류 발생 시간 X는 일반적인 분포를 따르는 것으로 하고, 확률 변수 X의 누적 확률 분포 함수를 $F_X(x)$라 하자. 오류가 발생했을 때, 고

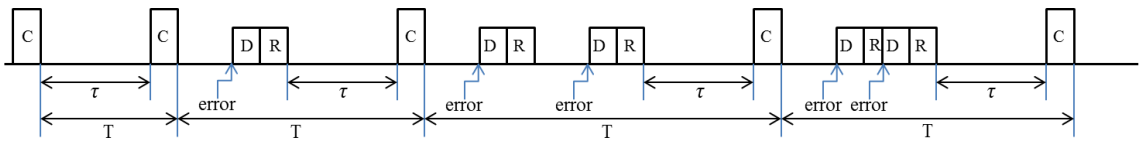

Fig. 1 Checkpointing model with instantaneous error detection

장 시간과 복구 시간이 필요하다. 이에 필요한 시간을 각각 D와 R로 표시한다. 고장 시간 D는 오류 발생 후에 작업 프로세서를 재활(rejuvenation)하는데 반드시 필요한 시간으로, 오류가 발생한 프로세서를 중단시키고 체크포인트 이미지를 적재할 새 프로세서를 복원하는 시간을 포함할 수 있다. 복구 시간 R은 체크포인트 시점에 저장되어 있던 정보를 주 기억 장치에 다시 로드하는데 요구되는 시간이다. 시간 τ, C, D, R을 상수로 가정한다. 복구 시간동안 오류가 발생하지 않는다면, 복구 시간이 끝난 후 시스템은 정상적으로 동작하며 오류가 발생하기 전 마지막 체크포인트 시점부터 다시 시작하게 된다. 체크포인트 발생 시간과 복구 시간에도 오류가 일어날 수 있지만, 고장 시간에는 추가적인 오류가 발생하지 않는다고 가정한다.

Ⅲ. 체크포인팅 모형의 신뢰도 분석

시각 t_n을 n번째 체크포인트가 생성된 시각이라 하고, 마지막 오류 발생 후부터 시각 t까지 생성된 체크포인트 수를 $S(t)$라 하자. 시각열 $\{t_n, n=1,2,...\}$는 확률과정 $\{S(t), t\geq 0\}$에 내재되어 있는 마르코프 순간(Markov point)이며,

$$S_n \equiv S(t_n) \tag{1}$$

으로 정의하면 확률 과정 $\{S_n, n=1,2,...\}$은 내재 마르코프 연쇄(embedded Markov chain)가 된다. 확률 과정 $\overline{S}(t)$를

$$\overline{S}(t) \equiv S_{\sup\{n:t_n < t\}} \tag{2}$$

로 정의하면, $\{\overline{S}(t), t\geq 0\}$는 반 마르코프 확률 과정(semi-Markov process)이 된다. 확률 과정 $\overline{S}(t)$의 값은 임의의 한 마르코프 점에서 다음 마르코프 점까지 일정하게 유지된다.

확률 분포(probability distribution) $\{\pi_i, i=1,2,...\}$를 마르코프 연쇄 $\{S_n, n=1,2,...\}$의 정상상태에서의 확률 분포(steady-state probability distribution)라 하면, $\{\pi_i, i=1,2,...\}$는

$$\pi_j = \sum_{i=1}^{\infty} \pi_i p_{ij}, \quad \sum_{i=1}^{\infty} \pi_i = 1 \tag{3}$$

을 만족하는 유일한 해이다. 여기서 조건부 확률 $p_{ij} \equiv \mathrm{P}\{S_{n+1}=j | S_n = i\}$는 상태 전이 확률이며 n에 의존하지 않는다. 양의 정수 i에 대하여 $S_n = i$로 시작하는 n번째 체크포인터 구간에 오류가 발생했을 때 $S_n = i$에서 $S_{n+1}=1$로의 상태 전이가 일어나고, 오류가 발생하지 않았을 때 $S_{n+1}=i+1$로의 상태 전이가 일어난다. 간단한 계산을 통해 상태 전이 확률을 계산하면,

$$p_{ij} = \begin{cases} 1-q_{i,1}^{+}, i=1,2,...,j=1, \\ q_{i,1}^{+}, \quad i=1,2,...,j=i+1, \\ 0 \quad\quad i=1,2,...,j\neq 1, j\neq i+1 \end{cases} \tag{4}$$

이다. 여기서

$$q_{i,j}^{+} \equiv \mathrm{P}(X > R+(i+j)(\tau+C) \,|\, X > R+i(\tau+C))$$

$$= \frac{1-F_X(R+(i+j)(\tau+C))}{1-F_X(R+i(\tau+C))}, \quad i,j=1,2,...$$

이다. 식 (4)를 식 (3)에 대입하여 정리하면, 2보다 크거나 같은 정수 i에 대하여 다음 관계식을 얻을 수 있다:

$$\pi_i = \pi_{i-1} q_{i-1,1}^{+} = \pi_1 \prod_{j=2}^{i} q_{j-1,1}^{+} = \pi_1 q_{1,i-1}^{+} \tag{5}$$

확률 π_1을 정규화 조건으로 구하면

$$\pi_1 = \frac{1}{1+\sum_{j=1}^{\infty} q_{1,j}^{+}} \tag{6}$$

이 된다.

$$1+\sum_{j=1}^{\infty} q_{1,j}^{+} \tag{7}$$
$$= \sum_{j=0}^{\infty} \mathrm{P}(X > R+(j+1)(\tau+C) \,|\, X > R+\tau+C)$$
$$= \mathrm{E}\left(\left\lceil \frac{X-(R+\tau+C)}{\tau+C} \right\rceil \,\middle|\, X > R+\tau+C \right)$$

이므로

$$\pi_1 = \frac{1}{\mathrm{E}\left(\left\lceil \frac{X-(R+\tau+C)}{\tau+C} \right\rceil \,\middle|\, X > R+\tau+C \right)} \tag{8}$$

이다. 여기서 $\lceil a \rceil$는 a보다 크거나 같은 가장 작은 정수를 나타낸다.

반 마르코프 확률 과정 $\{\overline{S}(t), t\geq 0\}$가 상태 i에 머무는 시간을 T_i라 하자. 확률 변수 T_i의 평균 $\mathrm{E}(T_i)$는

$$\mathrm{E}(T_i) = \tau + C + (1-q_{i,1}^{+}) \times \tag{9}$$

$$\left[\begin{array}{l}\dfrac{F_X(R+\tau+C)}{1-F_X(R+\tau+C)}\{\mathrm{E}(X|X\le R+\tau+C)+D\}+R \\ \\ +D+\mathrm{E}(X-(R+i(\tau+C)) \\ \qquad\qquad |R+i(\tau+C)<X\le R+(i+1)(\tau+C))\end{array}\right]$$

이다. 반 마르코프 확률 과정 $\{\overline{S}(t), t\ge 0\}$가 임의의 상태에 머무는 시간 T의 평균 $\mathrm{E}(T)$는

$$\mathrm{E}(T)=\sum_{n=1}^{\infty}\pi_i\mathrm{E}(T_i) \tag{10}$$

로 구할 수 있다.

체크포인팅 시간 사이의 실제 유효한 작업 시간 τ에 대한 체크포인팅 모형의 신뢰도를 $R(\tau)$라 하면, 신뢰도 $R(\tau)$는

$$R(\tau)=\frac{\tau}{\mathrm{E}(T)} \tag{11}$$

로부터 구할 수 있다.

Ⅳ. 오류 발생이 지수 분포를 따르는 경우의 체크포인팅 모형 최적화 분석

본 절에서는 오류 발생 시간 X가 지수 분포를 따르는 경우를 다룬다. 지수 분포의 무 기억성에 의해 확률 변수 T의 확률 분포는 확률 변수 T_i의 확률 분포와 같다. 확률 변수 X가 파라미터가 λ인 지수 분포를 따른다면, 임의의 상수 a에 대해

$$\mathrm{E}(X|X\le a)=\frac{1-(1+\lambda a)e^{-\lambda a}}{\lambda(1-e^{-\lambda a})} \tag{12}$$

이고,

$$\mathrm{E}(T)=\left(\frac{1}{\lambda}+D\right)\frac{1-e^{-\lambda(\tau+C)}}{e^{-\lambda(\tau+C+R)}} \tag{13}$$

이다. 따라서 정상 상태에서의 신뢰도 $R(\tau)$는

$$R(\tau)=\frac{\lambda\tau e^{-\lambda(\tau+C+R)}}{(1+\lambda D)\left[1-e^{-\lambda(\tau+C)}\right]} \tag{14}$$

이다.

신뢰도를 최대로 하는 τ 값을 구하기 위해 $R(\tau)$를 τ에 대해 미분하면,

$$\frac{d}{d\tau}R(\tau)=\frac{\lambda e^{-\lambda(\tau+C+R)}\left[1-\lambda\tau-e^{-\lambda(\tau+C)}\right]}{(1+\lambda D)\left[1-e^{-\lambda(\tau+C)}\right]^2} \tag{15}$$

이다. 따라서 정상 상태에서의 신뢰도 $R(\tau)$를 최대로 하는 τ를 τ_{opt}라 하면, τ_{opt}는

$$1-\lambda\tau-e^{-\lambda(\tau+C)}=0 \tag{16}$$

의 해이다. 이 식의 해를 Lambert 함수 $L(z)$로 나타낼 수 있다. Lambert 함수 $L(z)$는

$$L(z)e^{L(z)}=z \tag{17}$$

를 만족한다. Lambert 함수 $L(z)$의 정의역을 $-\frac{1}{e}$보다 크거나 같은 실수로 제한하고 $L(z)\ge -1$이면 Lambert 함수 $L(z)$는 single-valued 함수이다. 식 (17)는

$$(\lambda\tau-1)e^{\lambda\tau-1}=-e^{-\lambda C-1} \tag{18}$$

로 나타낼 수 있고

$$-e^{-\lambda C-1}\ge-\frac{1}{e} \tag{19}$$

이므로, $L(-e^{-\lambda C-1})$은 single-valued 이다. 따라서 식 (18)의 해를 single-valued Lambert 함수로 다음과 같이 나타낼 수 있다:

$$\tau_{opt}=\frac{1}{\lambda}\left[1+L(-e^{-\lambda C-1})\right] \tag{20}$$

Ⅴ. 수치 해석

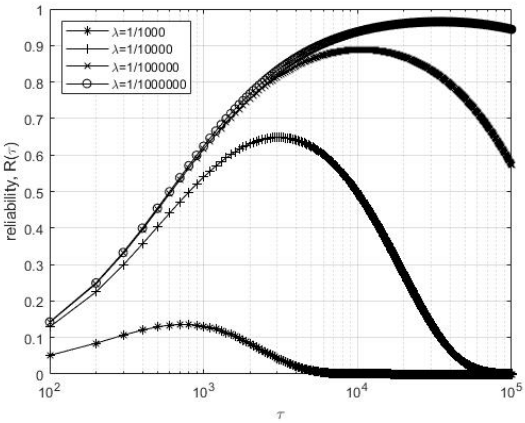

Fig. 2 Reliability as a function of τ

본 절에서는 오류 발생 시간 X가 지수 분포를 따르는 경우에 대한 몇 가지 수치적인 예를 제공한다. 먼저 τ 값에 따라 신뢰도 $R(\tau)$를 비교한다(그림 2). 이 결과를 위해 C와 R은 600초이고 D는 60초라 하고, $\frac{1}{\lambda}$은 1000초, 10000초, 100000초, 1000000초인 경우를 다룬다.

Table. 1 Optimal checkpoint interval and reliability

$\frac{1}{\lambda}$ (sec)	$C=R=100$ sec, $D=10$ sec		$C=R=500$ sec, $D=50$ sec		$C=R=900$ sec, $D=90$ sec	
	τ_{opt} (sec)	$R(\tau_{opt})$	τ_{opt} (sec)	$R(\tau_{opt})$	τ_{opt} (sec)	$R(\tau_{opt})$
1000	383	55.26%	698	17.43%	821	6.67%
2000	568	67.79%	1102	34.10%	1351	19.80%
4000	829	77.12%	1682	50.52%	2120	36.71%
8000	1199	83.85%	2505	64.12%	3220	52.80%
16000	1723	88.62%	3674	74.44%	4784	65.89%
32000	2464	91.98%	5329	81.93%	7002	75.74%
64000	3511	94.35%	7670	87.26%	10142	82.86%
128000	4993	96.02%	10983	91.03%	14585	87.92%
256000	7089	97.19%	15668	93.68%	20870	91.49%
512000	10053	98.02%	22295	95.54%	29761	94.01%
1024000	14244	98.60%	31668	96.86%	42335	95.77%
2048000	20172	99.01%	44922	97.78%	60117	97.02%
4096000	28555	99.30%	63667	98.43%	85266	97.89%

각각의 경우에 대해 τ값이 커짐에 따라 신뢰도 $R(\tau)$는 커지다가 최고점에 도달한 후 줄어드는 형태를 보인다. 또한, 임의의 τ에 대하여, $\frac{1}{\lambda}$이 커짐에 따라 $R(\tau)$의 값은 커지며 $R(\tau)$를 최대로 만드는 최적의 τ 값도 커진다는 것을 알 수 있다.

다음으로 $\frac{1}{\lambda}$의 값에 따라 신뢰도 $R(\tau)$를 최대로 만드는 최적의 τ값인 τ_{opt}의 값과 그때의 최대 신뢰도 $R(\tau_{opt})$를 비교한다(표 1). 이 결과를 위해 $C=R=100$초, $D=10$초인 경우와 $C=R=500$초, $D=50$초인 경우, $C=R=900$초, $D=90$초인 경우를 다룬다. 각각의 경우에 $\frac{1}{\lambda}$이 커짐에 따라 τ_{opt} 값과 $R(\tau_{opt})$의 값이 커지는 것을 알 수 있다. 또한, 임의의 $\frac{1}{\lambda}$에 대하여, C, R, D의 값이 커짐에 따라, τ_{opt} 값이 커지지만, τ_{opt} 값의 증가율이 C, R, D 값의 증가율에는 미치지 못하는 것을 알 수 있다. C, R, D의 값이 커짐에 따라, $R(\tau_{opt})$의 값은 작아지는 것을 알 수 있다.

Ⅵ. 결 론

본 논문에서는 오류 발생 시간이 독립이고 동일한 일반적인 분포를 따른다는 가정에서 즉각적으로 오류를 감지하는 경우의 체크포인팅 모형을 분석하였다. 두 체크포인트 사이에 많아야 하나의 오류만 발생한다는 기존 논문에서의 가정을 제거하였다. 체크포인트 발생 시간, 고장 시간, 복구 시간 등이 주어졌을 때, 시스템의 신뢰도를 유도하고, 오류 발생 시간이 지수 분포를 따르는 경우에 최적의 체크 포인팅 시간 간격을 구하였다. 최적의 체크포인팅 시간 간격이 커짐에 따라 신뢰도가 위로 볼록한 형태의 곡선으로 나타나며, 체크포인트 발생 시간과 고장 시간 및 복구 시간이 커짐에 따라 최적의 체크포인팅 시간 간격이 줄어드는 것으로 나타났다.

REFERENCES

[1] A. Benoit, A. Cavelan, Y. Robert, and H. Sun, “Multi-level checkpointing and silent error detection for linear workflows,” *Journal of Computational Science*, vol. 28, pp. 398-415, Arp. 2017.

[2] Y. Du, L. Marchal, G. Pallez, and Y. Robert, “Optimal checking strategies for iterative applications,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 33,

no. 3, pp. 507-522, Mar. 2022.

[3] A. Benoit, A. Cavelan, F. Cappello, P. Raghavan, Y. Robert, and H. Sun, “Coping with silent and fail-stop errors at scale by combining replication and checkpointing,” *Journal of Parallel and Distributed Computing*, vol. 122, no. 1, pp. 209-225, Aug. 2018.

[4] Y. Lee, “Reliability analysis of checkpointing model with multiple verification mechanism,” *Bulletin of the Korean Mathematical Society*, vol. 56, no. 6, pp. 1435-1445, Nov. 2019.

[5] J. W. Young, “A first order approximation to the optimal checkpoint interval,” *Communications of the ACM*, vol. 17, no. 9, pp. 530-531, Sept. 1974.

[6] J. T. Daly, “A higher order estimate of the optimum checkpoint interval for restart dumps,” *Future Generation Computer Systems*, vol. 22, no. 3, pp. 303-312, 2004.

[7] M. S. Bouguerra, D. Trystram, and F. Wagner, “Complexity analysis of checkpoint scheduling with variable costs,” *IEEE Transactions on Computers*, vol. 62, no. 6, pp. 1269-1275, Mar. 2013.

이유태(Yutae Lee)

2001년 3월-현재 동의대학교 정보통신공학과 교수
1998년 3월-2001년 2월 한국전자통신연구원 선임연구원
1997년 8월 한국과학기술원 수학과 이학박사
※관심분야 : 정보 신선도 분석, 큐잉 이론, 빅데이터