C-030

ロールバックの制約を考慮した最適非連携チェックポイント間隔に関する検討

On the Optimal Checkpoint Interval in Uncoordinated Checkpointing with Bound Rollbacks

大原 衛[†] 新井 雅之[†] 福本 聡[†] 岩崎 一彦[†]
Mamoru Ohara Masayuki Arai Satoshi Fukumoto Kazuhiko Iwasaki

1. まえがき

非連携チェックポインティングでは，通常運用時のチェックポインティングオーバヘッドを低減するために，プロセスは協調をおこなわず，独自にチェックポイントを生成する．リカバリ時には，各プロセスが保持する複数のチェックポイントから無矛盾な組み合わせ（リカバリライン）を探索し，ロールバックをおこなう．この際にドミノ効果が生じ，ロールバックが連鎖的に発生する場合がある．ロールバック間隔は無制限に大きくなり，リカバリオーバヘッドが制御できない．これを防ぐために，実際的な応用では追加的なしくみによってロールバックに制約を与える．従来の解析モデルでは，このような制約は考慮されていない．

本稿では，ロールバック間隔に制約をもったシステムの総期待オーバヘッドを解析的に評価し，これを最小化するチェックポイント間隔について述べる．シミュレーションによって，これらの制約が厳しい場合にも，本稿の解析モデルが最適チェックポイント間隔を与えることなどが示される．

2. 関連研究

非連携チェックポインティングでは，各プロセスが独自にチェックポイント生成時機を決定する．このためのアルゴリズムとして，各プロセスが一定数Tのイベントを実行する毎に状態を保存する定期的チェックポインティング（Periodical checkpointing）がよく用いられる．定数Tは，チェックポイント間隔と呼ばれる．

チェックポイント間隔$T>1$のとき，必ずしもリカバリポイントにチェックポイントが存在しない．このため，プロセスはリカバリ時に，まずリカバリポイント直前のチェックポイント取得時の状態を再構築し，このチェックポイントとリカバリポイント間のイベントを再実行することで，リカバリポイントへロールバックする．Linらは，このような手法について，通常時オーバヘッドとリカバリ時のオーバヘッドを含む総オーバヘッドを解析的に評価し，その最大・最小値および最適チェックポイント間隔を導いた[1]．

Solimanらは，定期的チェックポインティングに加えて，イベント実行前後の状態差分を保存し，リカバリを効率化するハイブリッド状態保存手法（HSS）を提案し，Linらと同様の手法で最適チェックポイント間隔を求めた[2]．HSSにおけるリカバリでは，プロセスはリカバリポイントに最も近いチェックポイントのデータを読み込み，これに状態差分を順次適用することでロールバックをおこなう．リカバリポイントがその直前のチェックポイントから離れていても，直後のチェックポイントから状態を再構築できるため，効率的にロールバックできる．

3. 総期待オーバヘッドの解析

本節では，プロセスの保持できるチェックポイント数と最大ロールバック間隔に制約がある場合のHSSの総期待オーバヘッドを解析的に評価し，最適チェックポイント間隔を導く．各プロセスが保持できるチェックポイントの最大数をN，最大ロールバック間隔をLとする．

まず，通常運用時に1イベント当たりに付加されるオーバヘッドを見積もる．1イベント当たりの通常時オーバヘッドは，チェックポイント間隔をT，チェックポイントデータ取得のオーバヘッドをC，1イベント当たりの状態差分生成オーバヘッドをδとすると，$C/T+\delta$で与えられる．

次に，1回のロールバックリカバリに伴うオーバヘッドを導く．障害発生時点とリカバリポイント間の間隔をロールバック間隔と呼び，確率変数Xで表す．ロールバック間隔Xの確率分布関数を$f(x)$とする．ロールバック間隔がXであるときのロールバックオーバヘッドを$r(X)$とすると，ロールバック1回あたりの期待オーバヘッドは，$R(T)=\sum_{x=0}^{L} r(x)f(x)$として得られる．

各プロセスがLに対して十分な数のチェックポイントを保持できる場合，すなわち，最適チェックポイント間隔T^*に対し，$L<NT^*$が成り立つ場合，常にリカバリポイントの近くにチェックポイントが存在する．$L=NT^*$を満たすLをL_{bound}で表す．このとき，1回のロールバックに必要なオーバヘッドは，$r(X)=C+\delta/4$で近似できる．ただし，障害発生時点およびリカバリポイントは，チェックポイント間隔のちょうど中間時点であるとし，チェックポイントデータの読み込みおよび状態差分の適用に必要なオーバヘッドは，それぞれの生成オーバヘッドと等しいとした．

逆に，チェックポイント数Nに対して，最大ロールバック間隔Lが相対的に大きい場合，リカバリポイントがプロセスの保持する最も古いチェックポイントよりもかなり以前の時点となる場合がある．このような場合には，プロセスは最も古いチェックポイントを取得した時点の状態を再構築し，ここから状態差分を用いてリカバリポイントまで遡る．このため，$r(X)$は以下のように表される．

$$r(X)=\begin{cases} \delta T/8 & 0\le X<T/4 \\ C+\delta T/8 & T/4\le X<T/2 \\ C+\delta T/4 & T/2\le X<(N-\frac{1}{2})T \\ C+\delta\{X-(N-\frac{1}{2})T\} & (N-\frac{1}{2})T\le X\le L. \end{cases} \quad (1)$$

プロセスは，リカバリポイントへロールバックした後，

[†]東京都立大学大学院工学研究科，Graduate School of Engineering, Tokyo Metropolitan University

障害発生時点までイベントを再実行する．このための期待オーバヘッドは，$(1+C/T+\delta)E[X]$ である．

これらを合わせて，1 イベント当たりの総期待オーバヘッドは，

$$H(T)=\frac{C}{T}+\delta+\lambda\left\{\left(1+\frac{C}{T}+\delta\right)E[X]+R(T)\right\} \quad (2)$$

として与えられる．ここで，λ は 1 イベント当たりのロールバックリカバリ発生数である．$H(T)$ を最小化する最適チェックポイント間隔 T^* は，$H(T+1)-H(T)=0$ を解いて得られる．

Soliman らの解析では，ロールバック間隔は幾何分布に従うと仮定された．同様に，本研究では $f(x)$ を $[0,L]$ に成分を持つ片側の切れた幾何分布と仮定する．すなわち，$f(x)=g(x)/\sum_{y=0}^{L}g(y)$ である．ただし，$g(x)=p(1-p)^x$ とした．

以上を用いて最適チェックポイント間隔を数値的に求めることができるが，陽に解を得るのは容易ではない．本稿では，さらに $f(x)$ を一様分布で近似することで，T^* を陽に求める．次節の数値例から，一様分布による近似は，L が小さい場合にはよい近似解を与えることが示される．一様分布 $f(x)=1/L$ を用いて，$L<L_{bound}$ のときの最適チェックポイント間隔

$$T^*=\frac{-1+\sqrt{1+\frac{16}{\delta\lambda}C\left(1+\frac{\lambda L}{2}\right)}}{2} \quad (3)$$

を得る．これから，

$$L_{bound}=\frac{N}{2}\left\{\frac{2}{\delta}CN-1+\sqrt{\left(1-\frac{2}{\delta}CN\right)^2+\frac{16C}{\delta\lambda}}\right\} \quad (4)$$

である．また，$L\geq L_{bound}$ のとき，

$$T^* = \left\lceil \omega^m\sqrt[3]{-\frac{Q}{2}+\sqrt{\frac{Q^2}{4}+\frac{P^3}{27}}} + \omega^{3-m}\sqrt[3]{-\frac{Q}{2}-\sqrt{\frac{Q^2}{4}+\frac{P^3}{27}}}-\frac{A_2}{3}\right\rceil \quad (5)$$

として求まる．ただし，m は 1,2, または 3 であり，

$$A_2 = \frac{\delta\left(1+8L-20N-16LN+24N^2\right)-4C}{2\delta\left(8N^2-4N-1\right)}, \quad (6)$$

$$A_1 = \frac{\delta\left(3+8L+12N-16LN+8N^2\right)-4C}{2\delta\left(8N^2-4N-1\right)}, \quad (7)$$

$$A_0 = -\frac{4C(L+1)\left\{2+\lambda(L+1)\right\}}{\lambda\delta\left(8N^2-4N-1\right)}, \quad (8)$$

$$P = -\frac{1}{3}A_2^3+A_1, \quad (9)$$

$$Q = \frac{2A_2^3}{27}-\frac{A_2A_1}{3}+A_0, \quad (10)$$

$$\omega = \frac{-1+\sqrt{3}i}{2} \quad (11)$$

である．

4. 数値例

与えられたチェックポイント間隔 T および N,L,C,δ,λ,p に対して，総オーバヘッドを求めるシミュレータを作成した．図 1 は，$N=5,C=2.7,\delta=0.9,\lambda=0.001,p=0.0015$ を設定した際のシミュレーション結果である．図で 'HSS' は Soliman らの，'proposal' は本稿の解析モデルによる最適チェックポイント間隔を設定したものである．また，'optimal' は，シミュレータに $T=1,2,\cdots$ を順次設定し，最適総オーバヘッドを求めたものである．

図から，チェックポイント数 N，最大ロールバック間隔 L を考慮しない Soliman らの解析は，必ずしもオーバヘッドを最小化するチェックポイント間隔を与えていないことが分かる．これに対して，本稿の解析は $L<2000$ では最適オーバヘッドを与える．一方，L が大きい場合には，一様分布による幾何分布の近似は適当でなくなる．また，同様のパラメータで $N>20$ と比較的多くのチェックポイントを保持できるような場合には，Soliman らの解析も最適オーバヘッドを与えた．

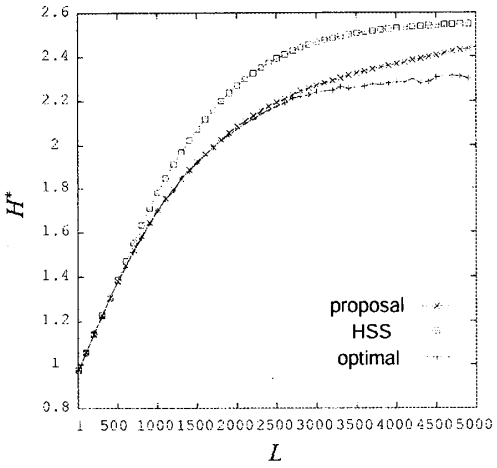

図 1: $N=5,p=0.0015$ における総期待オーバヘッド

5. まとめ

本稿では，プロセスが保持できるチェックポイント数と最大ロールバック間隔に制約がある場合における，非連携チェックポインティング手法の最適チェックポイント間隔を求めた．解を陽に得るために，ロールバック間隔の確率分布を一様分布で近似した．

シミュレーションから，これらの制約を考慮しない従来の解析モデルは必ずしも最適チェックポイント間隔を与えないのに対して，本稿の解析モデルがよりよいチェックポイント間隔を与えることが示された．

参考文献

[1] Y. Lin, B. Preiss, W. Loucks, and E. Lazowska, "Selecting the Checkpoint Interval in Time Warp Simulation," Proc. Workshop on Parallel and Distributed Simulation (PADS) 1993, pp. 3–10, 1993.

[2] H. M. Soliman and A. S. Elmaghraby, "An Analytical Model for Hybrid Checkpointing in Time Warp Distributed Simulation," *IEEE Trans. Parallel and Distributed Systems*, Vol. 9, No. 10, pp. 947–951, Oct. 1998.