Available online at www.sciencedirect.com

SCIENCE DIRECT®

Future Generation Computer Systems 22 (2006) 303–312

FGCS
FUTURE GENERATION COMPUTER SYSTEMS

www.elsevier.com/locate/fgcs

A higher order estimate of the optimum checkpoint interval for restart dumps

J.T. Daly

Los Alamos National Laboratory, M/S T080, Los Alamos, NM 87545, USA

Available online 28 December 2004

Abstract

This paper examines methods of approximating the optimum checkpoint restart strategy for minimizing application run time on a system exhibiting Poisson single component failures. Two different models will be developed and compared. We will begin with a simplified cost function that yields a first-order model. Then we will derive a more complete cost function and demonstrate a perturbation solution that provides accurate high order approximations to the optimum checkpoint interval.

Keywords: Optimal checkpointing; Poisson failures; Perturbation series; Lambert function

1. Introduction

In this study we will expand upon the work begun by Young [1] and continued by Daly [2] in refining a model for quantifying the optimum restart interval that minimizes the total application run time. Our goal is to derive a result in terms of a simple analytic approximation, easily accessible to the application user, with well-defined error bounds. Our strategy for optimizing the compute interval between dumps τ is to generate a cost function $T_w(\tau)$, the total wall clock time to complete the execution of an application, and to determine a unique minimum. Heuristically speaking, this cost function will look like

$$T_w(\tau) = \text{solve time} + \text{dump time} + \text{rework time} + \text{restart time}. \quad (1)$$

Solve time is defined as time spent on actual computational cycles towards a final solution. For a system with no interrupts, the wall clock time $T_w(\tau)$ consists entirely of solve time. Dump time is the overhead spent writing out the checkpoint files required to restart the application after an interrupt. Rework time is the amount of wall clock time lost when an application is killed by an interrupt prior to completing a restart dump. It is the amount of time elapsed since the last restart dump was successfully written. Restart time is the time required before an application is able to resume

E-mail address: jtd@lanl.gov.

0167-739X/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2004.11.016

real computational work. It includes both the application initialization and any system overhead associated with restarting a calculation after an interrupt.

2. A first-order model

Young [1] proposed $\tau_{\text{opt}} = \sqrt{2\delta M}$ as a useful approximation of the optimum checkpoint interval, where δ is the time to write a checkpoint file, M is the mean time to interrupt (MTTI) for the system, and τ_{opt} is the optimum compute time between writing checkpoint files. We will start with a first-order derivation that produces a result similar to Young's estimate before moving on to a more complete model. To help us consider how the wall clock time relates to the solve time, consider Fig. 1, which provides a conceptual view of an application run encountering a single interrupt.

2.1. The basic cost function

Referring to Fig. 1, it becomes straightforward to construct a basic cost function for total wall clock time. Solve time will be defined as T_{s}, which is equal to $N\tau$, where N is the number of passed segments required to complete a calculation. Dump time will be $(N - 1)\delta$, where one is subtracted because there is no dump on the last segment. For rework time we will assume that, on average, interrupts occur some fraction $\phi(\tau + \delta)$ of the way through a segment. (For the moment, we will make the simplifying assumption that interrupts never occur during problem restart. This assumption will be relaxed when we develop our new cost function in Section 3.) This implies that, over a large number of failures, the amount of work completed in a segment prior to the failure, equivalent to the amount of rework for that segment, will be $\phi(\tau + \delta)$ times the segment length. Finally, the restart time is simply $Rn(\tau)$, the product of the amount of time required to restart and the total number of failures. Combining all of these terms, we construct our basic cost function.

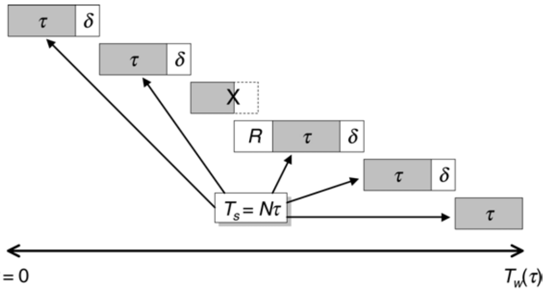


Fig. 1. The application time line broken into five passed compute segments and one failed compute segment designated by X. An application run is complete when the accumulated computation time τ of all of the passed segments is equal to the total solution time T_{s} for the application.

$$T_w(\tau) = T_{\text{s}} + \left(\frac{T_{\text{s}}}{\tau} - 1\right)\delta + [\tau + \delta]\phi(\tau + \delta)n(\tau) + Rn(\tau). \qquad (2)$$

2.2. Determining the number of interrupts

The simplest useful life distribution model for mechanical and electrical equipment for which the mean time to interrupt is known to be M is described by an exponential model [3] whose probability density function is

$$f(t) = \frac{1}{M}\,\mathrm{e}^{-t/M}. \qquad (3)$$

The probability of an interrupt occurring before time Δt in such a system is given by the cumulative distribution function

$$P(t \le \Delta t) = \int_0^{\Delta t} \frac{1}{M}\,\mathrm{e}^{-t/M}\,\mathrm{d}t = 1 - \mathrm{e}^{-\Delta t/M}. \qquad (4)$$

From this we see that the probability of successfully computing for a time Δt without an interrupt will be

$$P(t > \Delta t) = 1 - P(t \le \Delta t) = \mathrm{e}^{-\Delta t/M}. \qquad (5)$$

Therefore, the average number of attempts needed to complete N calculations of duration Δt is

$$\text{number of attempts} = \frac{N}{P(t > \Delta t)} = N\,\mathrm{e}^{\Delta t/M}. \qquad (6)$$

Finally, the total number of interrupts is the number of attempts minus the number of successes.

$$n(\Delta t) = \frac{N}{P(t > \Delta t)} - N = N(\mathrm{e}^{\Delta t/M} - 1). \qquad (7)$$

Notice that this assumes that we will never have more than a single failure in any given compute segment. This assumption will also be relaxed in our new model.

2.3. First-order assumptions

Assuming interrupts arrive according to a Poisson process (see [4–6]), we want to linearize our cost function by making the simplifying assumption that the exponential term in Eq. (7) is small, which means that $\tau + \delta \ll M$. In that case the exponential term behaves basically linearly, and we can use Eq. (7) to approximate the expected number of failures as

$$n(\tau) = \frac{T_s}{\tau}(e^{(\tau+\delta)/M} - 1) \cong \frac{T_s}{\tau}\left(\frac{\tau+\delta}{M}\right) \quad \text{for } \frac{\tau+\delta}{M} \ll 1. \tag{8}$$

We still need to address the issue of the fraction of rework. In the case that the compute segment is significantly short compared to the MTTI for the system, which we already assumed when we linearized the exponential in Eq. (8), it turns out to be a good estimate that, on the average, interrupts will occur halfway through the compute interval. (This can be neatly demonstrated by taking the $\phi(\tau)$ derived in Eq. (15) and looking at the limit as $\Delta t/M$ goes to zero.) So we will use the following as the fraction of rework in our first-order model:

$$\phi(\tau) = \tfrac{1}{2}. \tag{9}$$

Substituting the terms for $n(\tau)$ and $\phi(\tau)$ into our cost function from Eq. (2) gives

$$T_w(\tau) = T_s + \left(\frac{T_s}{\tau} - 1\right)\delta + \left[\tfrac{1}{2}(\tau+\delta) + R\right]\frac{T_s}{\tau}\left(\frac{\tau+\delta}{M}\right). \tag{10}$$

Finally, from the results of Ling et al. [7], we know that, for a system with a Poisson failure distribution in which no failures occur during checkpointing and recovery, the optimal checkpointing strategy will be equidistant. Therefore, our model will be solving for a single value of τ_{opt}. Aperiodic checkpointing strategies are not considered.

2.4. Solving the first-order model

Eq. (10) will be our cost function for the first-order model. We are interested in finding a unique minimum for values of $\tau > 0$. To do this we consider solutions of the first derivative with respect to τ that are equal to zero.

$$-\frac{\delta T_s}{\tau^2} + \frac{T_s}{M}\left(\frac{1}{2} - \frac{\delta^2}{2\tau^2} - \frac{\delta R}{\tau^2}\right) = -\frac{1}{\tau^2}(2\delta M + 2\delta R + \delta^2) + 1 = 0. \tag{11}$$

Thus the minimization problem reduces to a simple quadratic form. Assuming that the delta squared term is negligible (the assumption we made when we expanded the exponential failure term in Eq. (8)), we recover Young's original solution with an added term for the restart overhead.

$$\tau_{\text{opt}} = \sqrt{2\delta(M+R)} \quad \text{for } \tau + \delta \ll M. \tag{12}$$

Notice that this solution is identical to Young's [1] except for the presence of the restart time R under the radical because Young did not include the restart time in his derivation. (If we assume that $R = 0$, then we recover Young's solution exactly.) With this as our starting point, we will now develop and solve a new cost function that relaxes our first-order assumptions.

3. Developing a new cost function

Let us reconsider two assumptions that turn out not to be very accurate for small M. The first was made in Eq. (9), where the fraction of a segment requiring rework was approximated as one-half. In fact, that was a reasonable approximation for large M, but as M approaches $\tau + \delta$, the fraction of rework drops off precipitously because the expected point of failure occurs before the end of the $\tau + \delta$ length segment. This means that the beginning of the segment will see far more failures than the end of the segment.

The second problematic assumption associated with our first-order model is that the segment size for a failure is always assumed to be $\tau + \delta$, which means a failure never occurs in a segment of length $R + \tau + \delta$. If we encounter a failure during a restart segment, then the contribution to wall clock time is the expected rework time for the $R + \tau + \delta$ segment.

3.1. Fractional rework

To better grasp how these expected failures are behaving, consider that the probability of a failure occurring halfway through any arbitrary compute segment is actually the sum of the probabilities of the failure occurring halfway through the first segment plus the probability of it occurring halfway through the second interval and so forth. In other words, the probability density function describing the probability of failure at a time t in any arbitrary compute segment of length Δt will be

$$g(t) = \frac{1}{M}\,\mathrm{e}^{-t/M} + \frac{1}{M}\,\mathrm{e}^{-(t+\Delta t)/M} + \frac{1}{M}\,\mathrm{e}^{-(t+2\Delta t)/M} + \cdots = \frac{\mathrm{e}^{-t/M}}{M(1-\mathrm{e}^{-\Delta t/M})}. \quad (13)$$

Therefore, the expected point of failure for a random variable T in the range $0 \le T \le \Delta t$ in terms of the probability density function $g(t)$ will be given by

$$E(\Delta t) = \int_0^{\Delta t} t g(t)\,\mathrm{d}t = \frac{\int_0^{\Delta t} t\,\mathrm{e}^{-t/M}\,\mathrm{d}t}{M(1-\mathrm{e}^{-\Delta t/M})} = \frac{M\,\mathrm{e}^{\Delta t/M} - M - \Delta t}{\mathrm{e}^{\Delta t/M} - 1} = M + \frac{\Delta t}{1-\mathrm{e}^{\Delta t/M}}. \quad (14)$$

So, the expected fraction of rework $\phi(\Delta t)$ over a time interval Δt will be $E(\Delta t)$, derived in Eq. (14), divided by the length of the interval

$$\phi(\Delta t) = \frac{M}{\Delta t} + \frac{1}{1-\mathrm{e}^{\Delta t/M}}. \quad (15)$$

3.2. Failures during restart segments

Previously, in Eq. (2), we estimated the contribution of restart and rework to our cost function as

$$\tfrac{1}{2}(\tau+\delta)n(\tau) + Rn(\tau). \quad (16)$$

Based on the expected point of failure given by Eq. (14), the distinction between failed solve-dump segments and restart-solve-dump segments can be incorporated into Eq. (16) by noticing that the time for rework and restart depends on the length of the interval between checkpoints.

There is no additional restart cost associated with computing a segment of length $R+\tau+\delta$ because its restart is included as part of its rework. If we define $P(\tau)$ as the probability of successfully completing an interval of length $R+\tau+\delta$ without an interrupt, then the expected number of failures during segments beginning with a restart will be the total number of interrupts $n(\tau)$ times $1-P(\tau)$ and the contribution of restart and rework to the cost function can be reformulated as follows:

$$\{E(\tau+\delta)+R\}P(\tau)n(\tau) + E(R+\tau+\delta)[1-P(\tau)]n(\tau). \quad (17)$$

3.3. Multiple failures in a compute segment

In order to allow for the possibility of multiple restarts in a single compute segment, we will redefine $n(\tau)$. Instead of estimating the total number of failures by dividing the number of compute segments by the probability of a failure in each segment, we will now use the total wall clock time divided by the mean time between failures. After replacing $n(\tau)$ by $T_w(\tau)/M$, our new model can be expressed as

$$T_w(\tau) = \frac{M(T_\mathrm{s} - \delta + \delta T_\mathrm{s}/\tau)}{M - \{E(\tau+\delta)+R\}P(\tau) - E(R+\tau+\delta)[1-P(\tau)]}, \quad (18)$$

where

$$E(\tau+\delta) = M + \frac{\tau+\delta}{1-\mathrm{e}^{(\tau+\delta)/M}},$$

$$E(R+\tau+\delta) = M + \frac{R+\tau+\delta}{1-\mathrm{e}^{(R+\tau+\delta)/M}},$$

$$P(\tau) = \mathrm{e}^{-(R+\tau+\delta)/M}.$$

4. Solving the new model

Using Eq. (18) as our starting point, we discover after some algebraic manipulation that the new model reduces conveniently to

$$T_w(\tau) = M\,\mathrm{e}^{R/M}(\mathrm{e}^{(\tau+\delta)/M} - 1)\left(\frac{T_\mathrm{s}}{\tau} - \frac{\delta}{\tau+\delta}\right). \quad (19)$$

Now assume that the write time for a checkpoint file is much less than the total application solve time, and

Eq. (19) simplifies even further to

$$T_w(\tau) = M\,\mathrm{e}^{R/M}(\mathrm{e}^{(\tau+\delta)/M}-1)\frac{T_\mathrm{s}}{\tau} \quad \text{for } \delta \ll T_\mathrm{s}. \qquad (20)$$

To find the extrema of Eq. (20), we will consider values of $\tau > 0$ such that the derivative with respect to τ is zero.

$$\begin{aligned}\frac{\mathrm{d}T_w(\tau)}{\mathrm{d}\tau} &= (\tau - M)\,\mathrm{e}^{(\tau+\delta)/M} + M \\ &= 0 \Rightarrow \left(1-\frac{\tau}{M}\right)\,\mathrm{e}^{(\tau+\delta)/M} = 1. \qquad (21)\end{aligned}$$

For positive values of τ, this will have only a single zero because the second derivative increases monotonically. Notice that, unlike in the first-order model, R has disappeared from the solution for the optimum when we provide for the possibility of failures during the restart segments of the calculation.

4.1. An exact solution

Nondimensionalize Eq. (21) by choosing $\xi = \sqrt{\delta/2M}$ and $\eta = (\tau+\delta)/M$, and

$$\mathrm{e}^{-\eta} + \eta = 2\xi^2 + 1. \qquad (22)$$

We may express the solution to equations of this form in terms of Lambert's function $W(z)$, which satisfies

$$W(z)\,\mathrm{e}^{W(z)} = z \qquad (23)$$

in the complex plane. If we restrict it to real values $\Re(z) = x \geq -1/e$ and require $W(x) \geq -1$, then the Lambert function is single valued. We can solve Eq. (22) as follows:

$$\begin{aligned}\mathrm{e}^{-\eta} + \eta = 2\xi^2 + 1 &\Rightarrow \mathrm{e}^{\mathrm{e}^{-\eta}}\,\mathrm{e}^{\eta} \\ &= \mathrm{e}^{2\xi^2+1} \Rightarrow -\mathrm{e}^{-\eta}\,\mathrm{e}^{-\mathrm{e}^{-\eta}} \\ &= -\mathrm{e}^{-2\xi^2-1} \Rightarrow -\mathrm{e}^{-\eta} = W(-\mathrm{e}^{-2\xi^2-1}).\end{aligned} \qquad (24)$$

In this case, $x = -\mathrm{e}^{-2\xi^2-1}$. Since $\xi \geq 0$, we have that $x \geq -1/e$, so we can write the solution simply in terms of a single-valued Lambert function on the principal branch as

$$\eta = 2\xi^2 + 1 + W(-\mathrm{e}^{-2\xi^2-1}). \qquad (25)$$

We now have an exact solution to our optima problem, but since that solution cannot be written in terms of elementary functions, it is not particularly useful as a simple engineering estimate unless we happen to have some means handy for numerically approximating $W(\xi)$.

4.2. Asymptotic analysis and a perturbation solution

While it is true that the Lambert function has a corresponding Taylor series that converges quite nicely for small values of ξ, it is still rather cumbersome to deal with even the first few terms of the expansion when attempting to do a quick and easy calculation. Besides, the Taylor series is not at all a reliable approximation as ξ approaches one. On the other hand, perturbation series often converge even when the small parameter becomes large.

Remember that ξ and η are non-negative, which means that $0 < \mathrm{e}^{-\eta} \leq 1$. For $\xi \gg 1$, the linear term η dominates the exponential term, which tends toward zero. This leads to the result

$$\lim_{\xi\to\infty} \eta(\xi) = 2\xi^2 + 1 \Rightarrow \tau_\mathrm{opt} \cong M. \qquad (26)$$

For $\xi \ll 1$, η is no longer the dominant term, and $\mathrm{e}^{-\eta} \cong 1 - \eta + \eta^2/2$, which gives us

$$\lim_{\xi\to 0} \eta(\xi) = 2\xi \Rightarrow \tau_\mathrm{opt} \cong \sqrt{2\delta M} - \delta. \qquad (27)$$

Since we know the asymptotic limit as $\xi \to 0$, we will pick ξ as our small parameter. Expand $\mathrm{e}^{-\eta}$ as a Taylor series, and substitute a perturbation type solution, designated by a tilde, of the form

$$\tilde{\eta}(\xi) = \sum_{n=0}^{\infty} a_n \xi^n. \qquad (28)$$

From Eq. (27), we already know that $\tilde{\eta} = 2\xi$ to highest order in ξ. Thus we have values for the first two terms of the perturbation series: $a_0 = 0$ and $a_1 = 2$. To get the remaining terms, expand the exponential from Eq. (22) by substituting $\tilde{\eta}$ and equating terms of like order.

$$\begin{aligned}&\left(1 - \tilde{\eta} + \frac{\tilde{\eta}^2}{2} + \frac{\tilde{\eta}^3}{6} + \frac{\tilde{\eta}^4}{24} + \frac{\tilde{\eta}^5}{120}\right) - \tilde{\eta} \\ &\quad = 1 + 2\xi^2 + \left(2a_2 - \tfrac{4}{3}\right)\xi^3\end{aligned}$$

$$+\left(2a_3-2a_2+\tfrac{1}{2}a_2^2+\tfrac{2}{3}\right)\xi^4 + \left(2a_4-2a_3+a_2a_3+\tfrac{4}{3}a_2-a_2^2-\tfrac{4}{15}\right)\xi^5 + \mathrm{O}(\xi^6)=2\xi^2+1. \qquad (29)$$

Now solve for the individual terms of the perturbation series by setting each coefficient separately equal to zero.

$$2a_2-\tfrac{4}{3}=0,\qquad 2a_3-2a_2+\tfrac{1}{2}a_2^2+\tfrac{2}{3}=0,$$
$$2a_4-2a_3+a_2a_3+\tfrac{4}{3}a_2-a_2^2-\tfrac{4}{15}=0. \qquad (30)$$

Thus, we get the next three terms of our perturbation series solution $\tilde{\eta}$ as

$$a_2=\tfrac{2}{3},\qquad a_3=\tfrac{2}{9},\qquad a_4=\tfrac{8}{135}. \qquad (31)$$

Based on the asymptotic analysis in Eqs. (26) and (27), we have developed solutions for the asymptotic limits as ξ becomes very small or large. In most cases, we expect that $\delta < M$, which implies $\xi < 1$. Because of this we are particularly interested in the accuracy of our solution to Eq. (22) for small ξ, which is why we developed the perturbation solution. This means that we can express the approximate solution $\tilde{\eta}$ in terms of two functions of ξ as

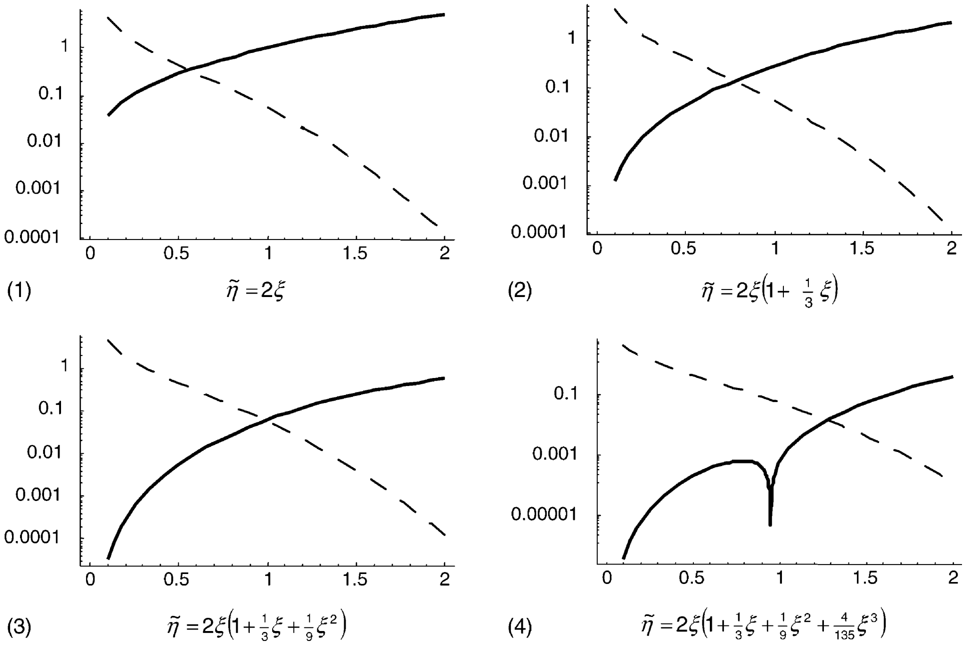
$$\tilde{\eta}=\begin{cases}2\xi\left(1+\tfrac{1}{3}\xi+\tfrac{1}{9}\xi^2+\tfrac{4}{135}\xi^3+\cdots\right) & \text{for } \xi<\xi_0,\\ 2\xi^2+1 & \text{for } \xi\geq\xi_0.\end{cases} \qquad (32)$$

All that remains is to pick the reference value ξ_0 such that the difference between the exact solution for η, given by Eq. (25), and the approximate solution $\tilde{\eta}$ is as small as possible over the entire range of $\xi > 0$.

5. Results

5.1. Determining the reference value from the relative error

Since we are able to calculate the exact value of the optimum checkpoint interval from Eq. (25), a relative error seems an intuitive metric for quantifying the accuracy of our approximation. Let us consider the



(1) $\tilde{\eta}=2\xi$

(2) $\tilde{\eta}=2\xi\left(1+\tfrac{1}{3}\xi\right)$

(3) $\tilde{\eta}=2\xi\left(1+\tfrac{1}{3}\xi+\tfrac{1}{9}\xi^2\right)$

(4) $\tilde{\eta}=2\xi\left(1+\tfrac{1}{3}\xi+\tfrac{1}{9}\xi^2+\tfrac{4}{135}\xi^3\right)$

Fig. 2. The relative error for four different perturbation solutions $\tilde{\eta}$, represented by solid lines, plotted as a function of the nondimensionalized dump time ξ. The relative error for the asymptotic solution $\tilde{\eta}=2\xi^2+1$ is shown with a dashed line in each plot.

relative error between the restart interval τ_{opt} and our approximation $\tilde{\tau}_{\text{opt}}$, which we may rewrite in terms of our nondimensional parameters ξ and η as

$$\varepsilon = \frac{|\tau_{\text{opt}} - \tilde{\tau}_{\text{opt}}|}{\tau_{\text{opt}}} = \frac{|\eta - \tilde{\eta}|}{\eta - \delta/M} = \frac{|\eta - \tilde{\eta}|}{\eta - 2\xi^2}. \tag{33}$$

Using our results from Eq. (25), we can write the relative error as

$$\varepsilon = \frac{|2\xi^2 + 1 + W(-\mathrm{e}^{-2\xi^2-1}) - \tilde{\eta}|}{1 + W(-\mathrm{e}^{-2\xi^2-1})}. \tag{34}$$

Plotting the relative error ε as a function of the nondimensionalized dump time ξ for various values of $\tilde{\eta}$ gives us the results shown in Fig. 2. Notice that, as more terms are added to the perturbation series, the error associated with the perturbation solution decreases and the point of intersection ξ_0 between the perturbation and asymptotic errors moves to the right. The relative error is minimized by choosing ξ_0 as the point of intersection between the errors associated with the perturbation solution and the asymptotic solution. Those minima are given in tabular form in Table 1.

Table 1
The optimal reference value ξ_0 associated with applying different numbers of terms from the perturbation solution $\tilde{\eta}$ in Eq. (32) is shown along with the maximum relative error in τ_{opt} corresponding to that solution

Perturbation solution, $\tilde{\eta}$	Reference value, ξ_0	Maximum relative error, ε
2ξ	0.5594	0.340
$2\xi\left(1 + \frac{1}{3}\xi\right)$	0.7751	0.143
$2\xi\left(1 + \frac{1}{3}\xi + \frac{1}{9}\xi^2\right)$	0.9861	0.059
$2\xi\left(1 + \frac{1}{3}\xi + \frac{1}{9}\xi^2 + \frac{4}{135}\xi^3\right)$	1.2767	0.015

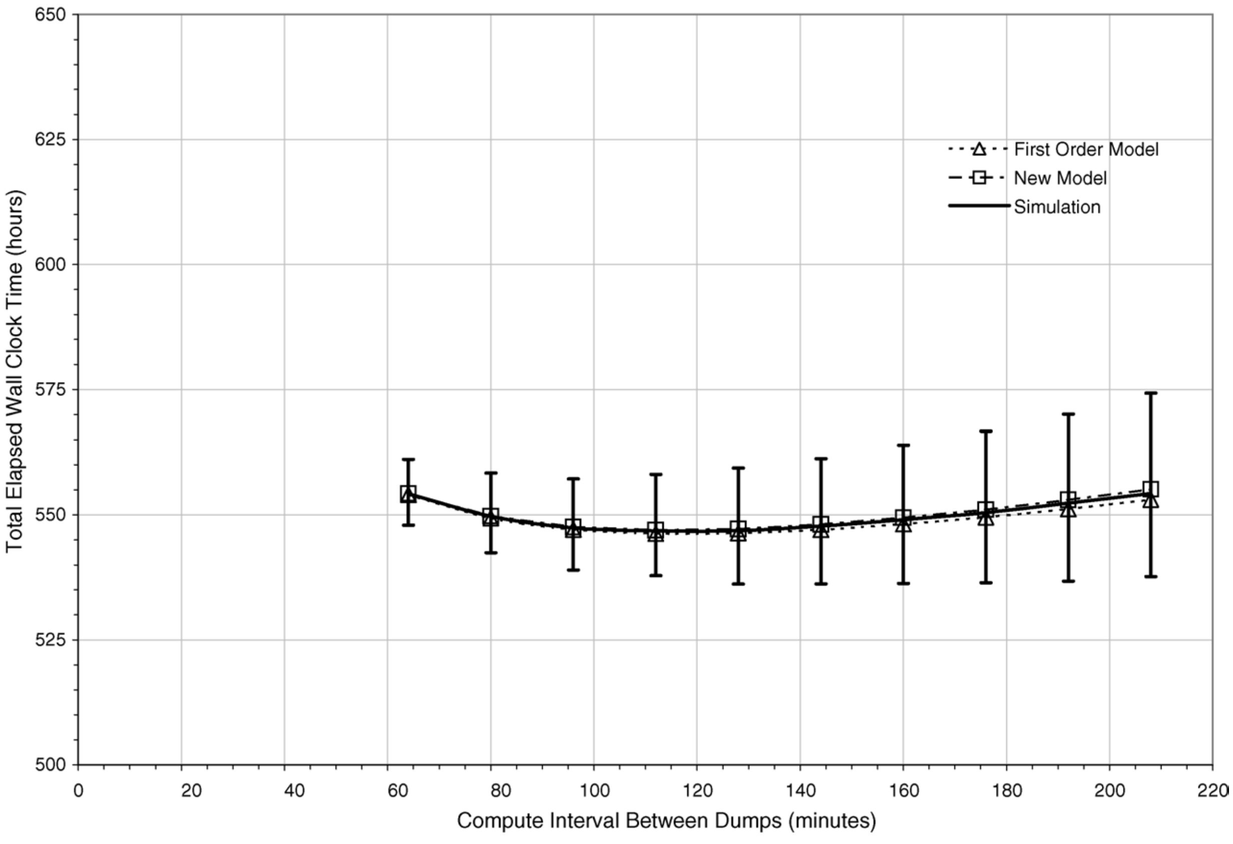


Fig. 3. Comparison of model and simulation results for $M = 24\,\text{h}$, $T_s = 500\,\text{h}$, $R = 10\,\text{min}$, and $\delta = 5\,\text{min}$. The new model predicts $\tau_{\text{opt}} = 117\,\text{min}$.

These results give us a nice upper bound on the maximum error in the restart interval for our perturbation solution in Eq. (32). However, what we would ultimately like to know is how to pick our reference value so as to minimize the maximum relative error in total wall clock time. To do this we will need to consider $\mathfrak{E}$, the relative error in wall clock time, shown below.

$$\begin{aligned}\mathfrak{E} &= \frac{|T_w(\tau_{\text{opt}}) - T_w(\tilde{\tau}_{\text{opt}})|}{T_w(\tau_{\text{opt}})} \\ &= \frac{|(\mathrm{e}^{\eta}-1)/(\eta-2\xi^2) - (\mathrm{e}^{\tilde{\eta}}-1)/(\tilde{\eta}-2\xi^2)|}{(\mathrm{e}^{\eta}-1)/(\eta-2\xi^2)} \\ &= \left|1 - \frac{\eta-2\xi^2}{\tilde{\eta}-2\xi^2}\left(\frac{e^{\tilde{\eta}}-1}{e^{\eta}-1}\right)\right|. \qquad (35)\end{aligned}$$

Using the solution from Eq. (25), we can write the relative error in wall clock time as

$$\mathfrak{E} = \left|1 - \frac{1+W(-\mathrm{e}^{-2\xi^2-1})}{\tilde{\eta}-2\xi^2} \times \left(\frac{\mathrm{e}^{\tilde{\eta}}-1}{\mathrm{e}^{2\xi^2+1+W(-\mathrm{e}^{-2\xi^2-1})}-1}\right)\right|. \qquad (36)$$

Using the same method as we did for Eq. (34), we can once again tabulate a reference value that corresponds to the maximum value of $\mathfrak{E}$ for various choices of $\tilde{\eta}$. We will designate these reference values ξ_1 in order to distinguish them for the reference values corresponding to the relative error in τ_{opt}. The results are shown in Table 2.

Notice that as we continue to add terms to $\tilde{\eta}$, not only do the relative errors ε and $\mathfrak{E}$ continue to decrease, but the difference between their reference values $|\xi_1 - \xi_0|$ decreases as well. By the time we have approximated $\tilde{\eta}$ to fourth order in ξ, their respective reference values agree to the first four decimal places.

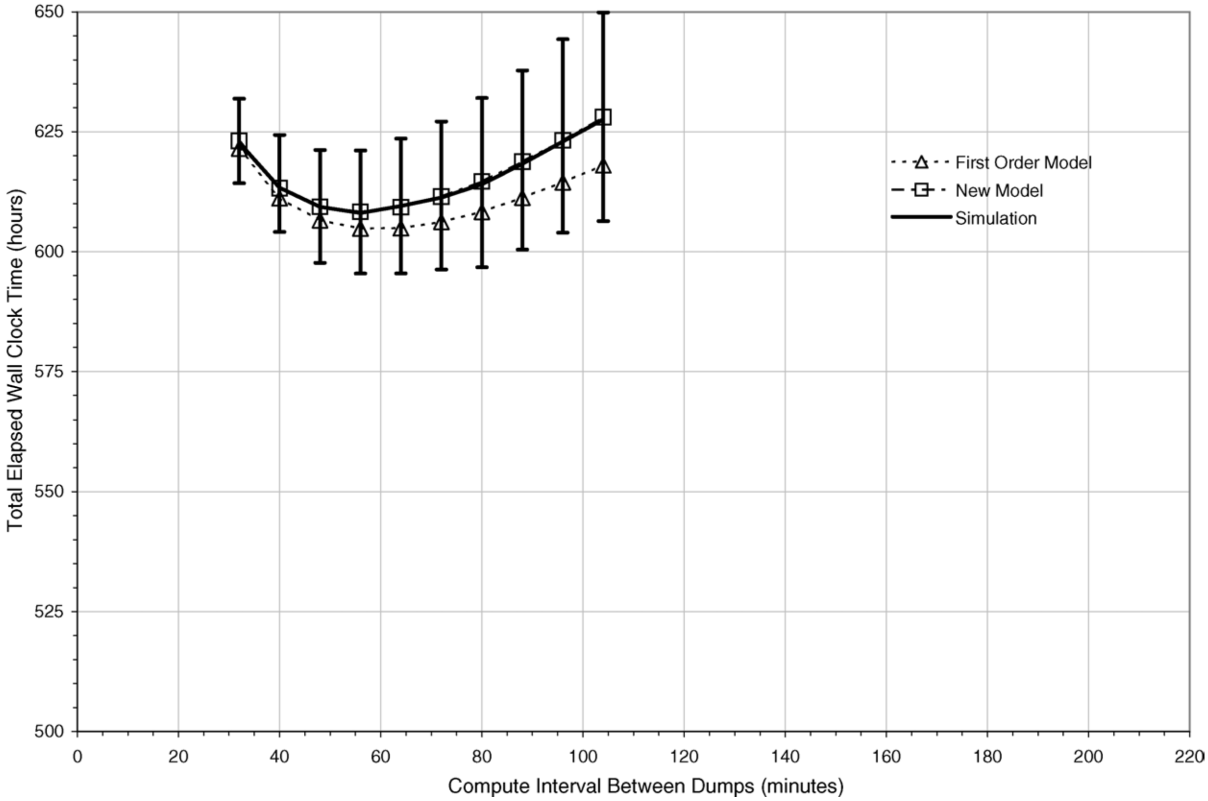


Fig. 4. Comparison of model and simulation results for $M = 6$ h, $T_s = 500$ h, $R = 10$ min, and $\delta = 5$ min. The new model predicts $\tau_{\text{opt}} = 57$ min.

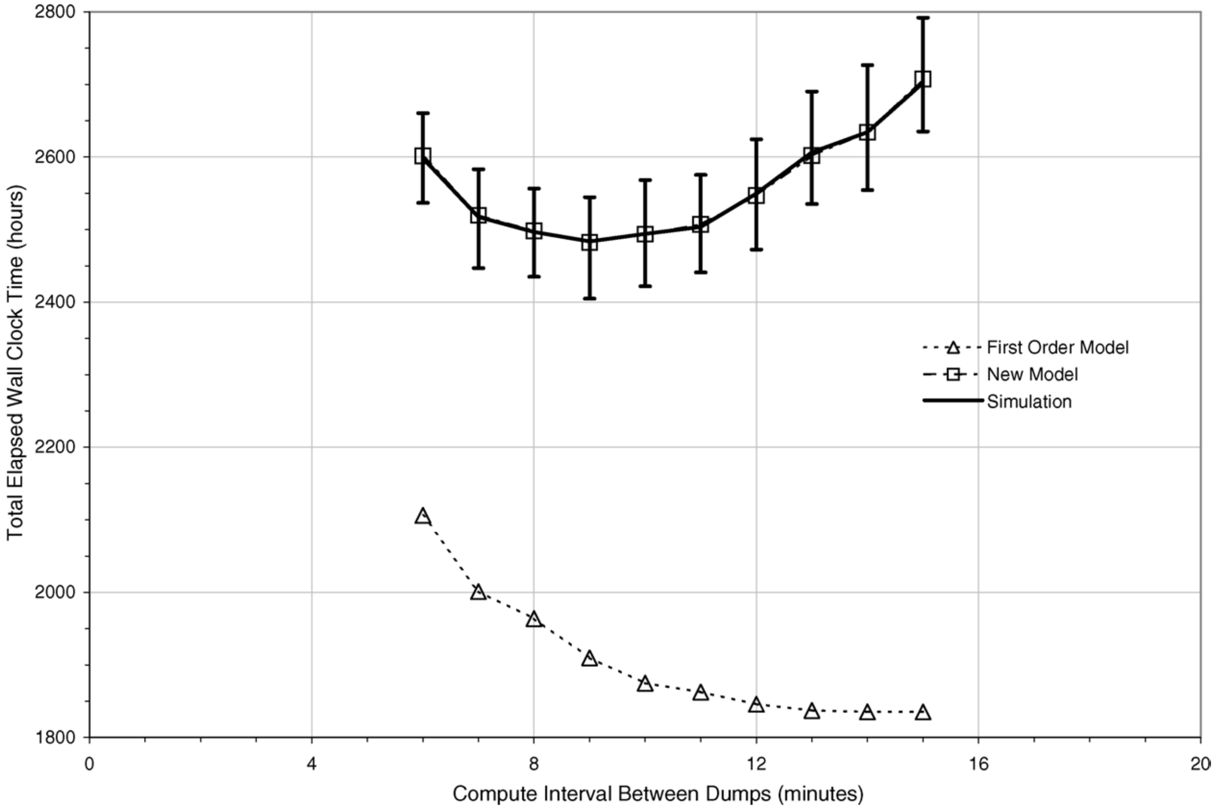


Fig. 5. Comparison of model and simulation results for $M = 15$ min, $T_s = 500$ h, $R = 10$ min, and $\delta = 5$ min. The new model predicts $\tau_{opt} = 9.1$ min.

5.2. Comparing models to simulation

In Figs. 3–5, we compare the wall clock times predicted by the first-order and new models to a simulated application execution. New model predictions of τ_{opt} are based on the first three terms of the perturbations solution.

Table 2
The optimal reference value ξ_1 associated with applying different numbers of terms from the perturbation solution $\tilde{\eta}$ in Eq. (32) is shown along with the maximum relative error in $T_w(\tau_{opt})$ corresponding to that solution

Perturbation solution, $\tilde{\eta}$	Reference value, ξ_1	Maximum relative error, $\mathfrak{E}$
2ξ	0.5218	0.0442
$2\xi\left(1+\frac{1}{3}\xi\right)$	0.7617	0.0090
$2\xi\left(1+\frac{1}{3}\xi+\frac{1}{9}\xi^2\right)$	0.9811	0.0016
$2\xi\left(1+\frac{1}{3}\xi+\frac{1}{9}\xi^2+\frac{4}{135}\xi^3\right)$	1.2767	0.0001

The simulation generates pseudo-random interrupts in an exponential distribution and determines the accumulated wall time required to complete N compute segments for the specified solve time. Each simulation is run up to 10,000 times per case, with the charts depicting the median simulation result as a bold line and the range in which the middle 95% of the data fell with error bars. The only parameter changing between runs is M.

Fig. 3 shows that for $\tau + \delta \ll M$, the first-order model yields a good estimate of both total wall clock time and the optimum checkpoint interval. As the mean time to interrupt decreases, Fig. 4 demonstrates that the first-order model is beginning to break down, though it still seems to be giving a reasonable estimate of the optimum restart interval. Finally, we see in Fig. 5 that, as the MTTI continues to decrease, the first-order model fails entirely while the new model continues to provide nearly exact agreement with the median values predicted by the simulation.

6. Conclusions

We found that even though the first-order model predicts a contribution of the restart time R to the selection of the optimum compute interval between checkpoints τ_{opt}, the higher order model demonstrates that in fact R has no contribution. Furthermore, we demonstrated that an excellent approximation to τ_{opt}, one that guarantees that the relative error in total problem-solution time of Eq. (36) never exceeds 0.2% of the exact solution time, is given by the first three terms of the perturbation solution.

$$\tilde{\tau}_{\mathrm{opt}} = \begin{cases} \sqrt{2\delta M}\left[1 + \frac{1}{3}\left(\frac{\delta}{2M}\right)^{1/2} \right. & \\ \left. \quad + \frac{1}{9}\left(\frac{\delta}{2M}\right)\right] - \delta & \text{for } \delta < 2M, \\ M & \text{for } \delta \geq 2M. \end{cases} \tag{37}$$

Finally, we observe that the maximum relative error in problem solution time associated with our lowest order perturbation solution, which is equivalent to Young's model, is less than 5%, even in the worst case when $\delta = M/2$. This is a good rule of thumb for most practical systems, when the value of δ is expected to be small compared to M.

$$\tilde{\tau}_{\mathrm{opt}} = \begin{cases} \sqrt{2\delta M} - \delta & \text{for } \delta < \frac{1}{2}M, \\ M & \text{for } \delta \geq \frac{1}{2}M. \end{cases} \tag{38}$$

References

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

[2] J. Daly, A model for predicting the optimum checkpoint interval for restart dumps, in: Proceedings of the ICCS 2003, LNCS 2660, vol. 4, 2003, pp. 3–12.

[3] NIST/SEMATECH e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook/.

[4] B. Dimitrov, Z. Khalil, N. Kolev, P. Petrov, On the optimal total processing time using checkpoints, IEEE Trans. Softw. Eng. 17 (1991) 436–442.

[5] S.W. Kwak, B.J. Choi, B.K. Kim, An optimal checkpointing-strategy for real-time control systems under transient faults, IEEE Trans. Reliab. 50 (2001) 293–301.

[6] N.H. Vaidya, Impact of checkpoint latency on overhead ratio of a checkpointing scheme, IEEE Trans. Comput. 46 (1997) 942–947.

[7] Y. Ling, J. Mi, X. Lin, A variational calculus approach to optimal checkpoint placement, IEEE Trans. Comput. 50 (2001) 699–708.

John Daly studied computational fluid dynamics with Antony Jameson at Princeton University. He is a technical staff member at the Los Alamos National Laboratory. His research interests include mathematical modeling and optimization of performance, reliability, and availability on systems where compute time is long compared to the mean time between interrupts. Prior to joining LANL he worked on the ASCI 20 TeraOp program as an application analyst for Raytheon Intelligence and Information Systems.