# A QUANTIZED STATE INTEGRATOR WITH SECOND ORDER ERRORS OVER MONOTONIC SEGMENTS

Rasika Mahawattege

Department of Mathematical Sciences
The University of Memphis
Memphis, TN, USA
rasikamahawattege@gmail.com

James Nutaro

Computational Sciences & Engineering Division
Oak Ridge National Laboratory Oak Ridge, TN, USA
nutarojj@ornl.gov

## ABSTRACT

We introduce a novel quantized state integration method that is analogous to the explicit, second order Runge-Kutta technique. A feature of this method is that, for a single differential equation over an interval where the solution is monotonic, it exhibits an error proportional to the square of the integration quantum. We offer a theoretical proof of this behavior and demonstrate it with a numerical example. At the same time, an extension of the method to a system with input causes the errors to become proportional to the integration quantum. The latter observation fits established theory; the former appears to be a new result that might offer new insights into the construction of quantized state integration methods.

**Keywords:** quantized state systems, numerical integration, discrete event simulation

## 1 INTRODUCTION

Quantized state integration methods constitute a family of numerical techniques for solving differential equations. The distinguishing feature of these methods is that the state is made discrete and time continuous, whereas discrete time methods (like the familiar Euler or Runge-Kutta methods) make time discrete and the state is continuous. The quantized state methods are particularly attractive for simulation problems that exhibit heterogeneous rates of activity in time and space. For these types of systems, the methods naturally and efficiently focus computation in areas of high activity (Muzy, Jammalamadaka, Zeigler, and Nutaro 2011). Illustrations of this advantage have appeared in numerous domains of application; see, e.g., Nutaro (2007), Muzy, Innocenti, Aiello, Santucci, and Wainer (2002), Nutaro, Zeigler, Jammalamadaka, and Akerkar (2003), Bergero, Casella, Kofman, and Fernández (2018).

In a well know result by Kofman (Kofman 2005, Cellier and Kofman 2006) it was shown that the errors exhibited by a quantized state numerical method are proportional to the spacing $D$ of discrete points in the state space; $D$ is the integration quantum. Nonetheless, it has been observed that in some special circumstances this error bound may be improved upon to obtain errors proportional to the square of the integration quantum (Nutaro 2003, Nutaro 2005). Though the linear proportionality result of Kofman holds in the general case, these special circumstances are worth examining for the insight they may offer regarding the construction of new schemes.

Towards this end, we introduce a novel quantized state integration method that is analogous to the explicit, second order Runge-Kutta (RK2) technique. For the special case of a single equation describing a monotonic

segment, we offer a proof that this quantized RK2 method (QRK2) exhibits a global error proportional to the square of the integration quantum and a truncation error proportional to its cube. The result is illustrated with numerical examples. We also offer an empirical demonstration that the linear error bound holds under other circumstances for the same method.

We call this a special case because the method, as formulated here, has errors proportional to $D^2$ only under the circumstances considered. However, this special result illuminates two fundamental and challenging sources of error for quantized integration methods: the type of information communicated at solution points and the overshooting of inflection points. The former can likely be addressed by communicating higher order derivatives. The latter poses an interesting question for future research.

After introducing the new method, proof, and demonstrations, we conclude with a series of conjectures that extrapolate from our results and may offer fertile ground for future work. Moreover, the proposed method is simple to implement and appears to offer some practical improvement over the numerical errors of the QSS1 method (Kofman and Junco 2001, Cellier and Kofman 2006, Nutaro 2007). This combination of simplicity and the potential for reduced errors may make the QRK2 method attractive for some applications.

## 2 SIMULATING A SINGLE ORDINARY DIFFERENTIAL EQUATION

Consider an initial value problem

$$\dot{x}(t) = f(x(t)); \quad x(t_0) = x_0 \tag{1}$$

over an interval in which $|f(x(t))|$ is not zero and the sign of $f(x(t))$ is constant. Several quantized state integration methods for (1) can be obtained from the Taylor series

$$x(t+h) = x(t) + h\dot{x}(t) + \frac{h^2\ddot{x}(t)}{2} + \sum_{n=3}^{\infty} \frac{h^n x^{(n)}(t)}{n!} \; . \tag{2}$$

The QSS1 method is obtained by discarding all terms in (2) after the first derivative which gives

$$x(t+h) = x(t) + h\dot{x}(t) \; .$$

By choosing the integration quantum $D$ the sequence of points in the state space that we calculate will be such that $D = |x(t+h) - x(t)|$. Hence, the interval $h$ between these solution points is approximated by

$$h = \frac{D}{|\dot{x}(t)|} \; .$$

If $\dot{x}(t) = -x(t) < 0$ over the interval of interest, then we calculate a solution $x(0)$, $x(0) - D$, $x(0) - 2D$, etc. at time points $D/|x(0)|$, $D/|x(0) - D|$, $D/|x(0) - 2D|$, and so forth.

The QRK2 method is obtained by discarding the summation terms in (2) after the second derivative. We replace $\ddot{x}(t)$ with the finite difference approximation

$$\ddot{x}(t) = \frac{\dot{x}(t+h) - \dot{x}(t)}{h}$$

and obtain from the Taylor series

$$x(t+h) = x(t) + \frac{h}{2}(\dot{x}(t) + \dot{x}(t+h)) \; . \tag{3}$$

Using $D = |x(t+h) - x(t)|$ we rearrange (3) to obtain

$$h = \frac{2D}{|\dot{x}(t) + \dot{x}(t+h)|} \,. \tag{4}$$

Without loss of generality, assume $f(x(t)) < 0$. The calculated solution at $t+h$ has the value of $x$ decreased by $D$. From this we have

$$\dot{x}(t+h) = f(x(t+h)) = f(x(t) - D) \,. \tag{5}$$

Now $h$ can be calculated using (4) and (5) to be

$$h = \frac{2D}{|f(x(t)) + f(x(t) - D)|} \,. \tag{6}$$

A simulation of $x(t)$ is accomplished with an iterative procedure

$$x(t_{n+1}) = x(t_n) - D; \quad t_{n+1} = \sum_{i=1}^{n+1} h_i; \quad h_i = \frac{2D}{|f(x(t_{i-1})) + f(x(t_{i-1}) - D)|} \,. \tag{7}$$

Before proceeding to an analysis of the error introduced by this simulation procedure, it is illuminating to examine the truncation error for the equation $\dot{x} = -ax$. The first four terms of the Taylor series are

$$x(t+h) = x(t) - ahx(t) + \frac{1}{2}h^2a^2x(t) - \frac{1}{6}h^3a^3x(t) + \cdots \tag{8}$$

When $x(t) > D$ the absolute value in the denominator of (6) can be removed to simplify the expression and the step in time is

$$h = \frac{2D}{2ax(t) - aD} \,. \tag{9}$$

Substituting (9) into the Taylor series (8) gives us

$$x(t+h) = x(t) - a\frac{2D}{2ax(t) - aD}x(t) + \frac{1}{2}\frac{4D^2}{(2ax(t) - aD)^2}a^2x(t) - \frac{1}{6}\frac{8D^3}{(2ax(t) - aD)^3}a^3x(t) + \cdots$$

The error $\varepsilon$ committed in a single step is the difference between the true solution $x(t+h)$ at time $t+h$ and the computed solution $x(t) - D$ so $\varepsilon = x(t+h) - (x(t) - D))$ and in terms of the Taylor series,

$$\varepsilon = D - a\frac{2D}{2ax(t) - aD}x(t) + \frac{1}{2}\frac{4D^2}{(2ax(t) - aD)^2}a^2x(t) - \frac{1}{6}\frac{8D^3}{(2ax(t) - aD)^3}a^3x(t) + \cdots \tag{10}$$

The $a$ terms throughout (10) cancel and may be dropped. Then multiplying as needed by $(2x(t) - D)$, we may transform the first, second, and third terms containing $D$ and $D^2$ into the ratio of polynomials

$$\frac{D(2x(t) - D)^2 - 2Dx(t)(2x(t) - D) + 2D^2x(t)}{(2x(t) - D)^2} \,. \tag{11}$$

Examining the terms in the numerator of (11) we find that

$$D(2x(t) - D)^2 = 4Dx(t)^2 + D^3 - 4D^2x(t)$$
$$-2Dx(t)(2x(t) - D) + 2D^2x(t) = -4Dx(t)^2 + 4D^2x(t)$$

and so the the error in (10) reduces to

$$\varepsilon = \frac{D^3}{(2x(t) - D)^2} - \frac{1}{6}\frac{8D^3}{(2ax(t) - aD)^3}a^3x(t) + \cdots$$

Hence, the largest term in the truncation error depends on $D^3$ and we may expect the error over any monotonic segment to depend on $D^2$. This intuition can be confirmed with an essentially geometric argument.

**Theorem 1.** *Consider an Initial Value Problem in the form*

$$\dot{x}(t) = f(x(t)); \quad x(t_0) = x_0$$

*over a finite interval of time and its approximate solution values $x_0$, $x_1$, ..., $x_k$ calculated with the iterative procedure* (7)*. We assume the following:*

1. *$f(x)$ is continuous and has the same sign for all $x$ in the interval $[x_0, x_k]$ and*
2. *there exists $m > 0$ such that $m < |f(x)|$ for all $x \in [x_0, x_k]$.*

*There exits constant $C > 0$ such that the error $\varepsilon$ in the approximate solution is bounded by $CD^2$.*

*Proof.* The proof is by analogy to the second order Runge-Kutta (RK2) numerical method. The RK2 method starts at the point $(x_0, t_0)$ and then uses steps of size $\Delta t$ in time to construct a numerical solution as subsequent points $(x_n, t_n)$ using the iteration

$$x_{n+1} = x_n + \frac{\Delta t}{2}(f(x_n) + f(x_n + \Delta t f(x_n))) \; ; t_n = t_0 + n\Delta t \; . \tag{12}$$

For this procedure choose $\Delta t$ such that

$$\Delta t = \frac{D}{m} \; .$$

Over the segment of interest $m < |f(x)|$ and $f(x)$ has the same sign everywhere. Hence, the largest $h_i$ in the iteration (7) is bounded from above by

$$h_i < \frac{2D}{2m} = \Delta t \; .$$

The RK2 method (12) simulates the system by averaging the slopes at points $x_n$ and $x_n + \Delta t f(x_n)$ and then stepping a distance in time $\Delta t$ along the line defined by that slope to move from $x_n$ to $x_{n+1}$. Without loss of generality we may assume $f(x) < 0$. In this case, the quantized state method simulates the system by using the averaged slopes at points $x_n$ and $x_n - D$. This second point is identical to the second point used by RK2 if $D/|f(x_n)|$ is chosen for its step in time as

$$\frac{D}{|f(x_n)|} f(x_n) = -D \; .$$

Consequently, over this monotonic segment the quantized state method acts identically to RK2, but always with a step in time that is less than $\Delta t$ as $D/|f(x_n)| < D/m = \Delta t$. It is well known that there exits a constant $K$ such that the error incurred by the RK2 method is bounded by $K\Delta t^2$. Therefore the error $\varepsilon$ is such that

$$\varepsilon < K\Delta t^2 = K\left(\frac{D}{m}\right)^2 = \frac{K}{m^2} D^2$$

and picking $C = K/m^2$ completes the proof. □

## 3 SYSTEMS OF EQUATIONS

For a system of equations we cannot expect errors proportional to $D^2$ everywhere. Nonetheless, we expect that a bound on the error in proportion to $D$ will hold for the proposed method numerical method (Cellier

and Kofman 2006, Kofman 2005). From this perspective, the proposed QRK2 technique may offer a simple, practical improvement over the QSS1 method that calculates its advance in time as $D/|f(x_n)|$.

QRK2 is identical to QSS1 except as follows; the discussion here is limited to a two state variable system, with states $x$ and $y$, but can be generalized in a natural way. Our realization of QRK2 in software builds upon the tutorial by Nutaro (2007). We describe the scheme for $x$; it is formulated for $y$ in the same way. At internal and confluent events (see Nutaro (2007) for definitions), the time advance $h$ for $x$ is calculated using its present value and the most recently received update $\tilde{y}$ of state variable $y$ as

$$
\begin{aligned}
k_1 &= f(x_n, \tilde{y}) \\
\Delta &= D\frac{k_1}{|k_1|} \\
k_2 &= f(x_n + \Delta, \tilde{y}) \\
h &= \begin{cases} \infty & \text{if } k_1 = 0 \\ \dfrac{2D}{|k_1 + k_2|} & \text{if } k_1 k_2 \geq 0 \\ \dfrac{2k_1 D}{(k_2 - k_1)|k_1 + k_2|} & \text{if } k_1 k_2 < 0 \end{cases}
\end{aligned}
$$

The third case for $h$ accounts for the step spanning an inflection point. In this case, the approximate point of the zero crossing lies at the solution for $d$ in a linear model of $f$. This zero crossing occurs when

$$k_1\left(1 - \frac{d}{D}\right) + k_2\frac{d}{D} = 0\,.$$

At an external event, the solution is advanced as with the RK2 scheme using the elapsed time $e$ for the step size. If $\tilde{y}$ is the prior communicated value of $y$ and $\bar{y}$ is the newly communicated value, then the new value $x_{n+1}$ of the state variable $x$ is calculated as

$$
\begin{aligned}
k_1 &= f(x_n, \tilde{y}) \\
k_2 &= f(x_n + ek_1, \bar{y}) \\
x_{n+1} &= x_n + \frac{e}{2}(k_1 + k_2)
\end{aligned}
$$

The time advance $h$ is calculated as before but with $D$ replaced by the distance remaining to the next quantum level and $\bar{y}$ is assigned to $\tilde{y}$. However, should the external event cause $x_{n+1}$ to move past the next quantum boundary, then the time advance is zero.

We make a preliminary exploration of this idea with the system of equations

$$
\begin{aligned}
\dot{x} &= -2x + y \\
\dot{y} &= -y
\end{aligned}
$$

which has the solution

$$
\begin{aligned}
x(t) &= x(0)e^{-2t} + y(0)(e^{-2t} - e^{-t})/3 \\
y(t) &= y(0)e^{-t}
\end{aligned}
$$

The initial conditions are $x(0) = 0$ and $y(0) = \sqrt{2}$.

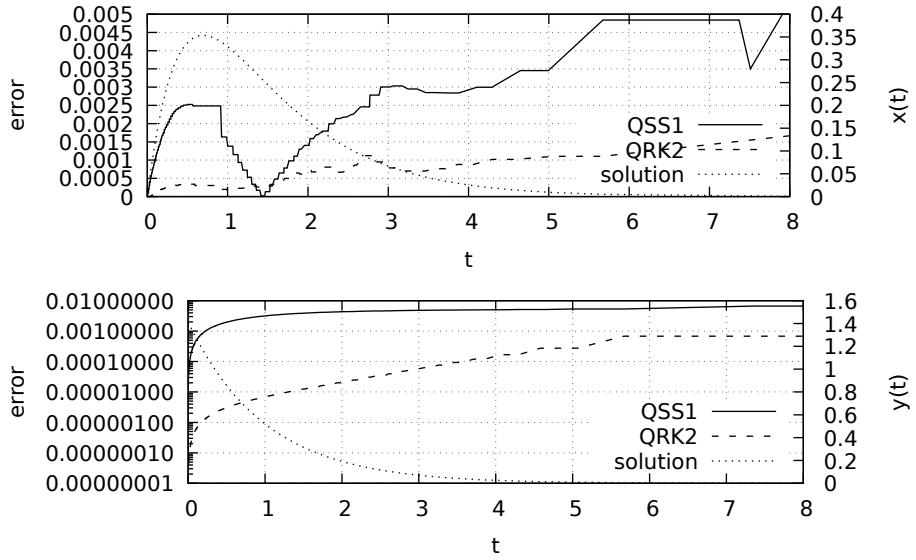


Figure 1: Plots of the exact solution to the system of test equations and the errors incurred by the QSS1 and QRK2 methods at each calculation point.

Figure 1 shows plots of the solution to this system and the errors at each calculation point using the QSS1 and QRK2 methods with $D = 0.01$. The solution for $y(t)$ satisfies the conditions of our theorem when the numerical solution is sufficiently far from zero. This is apparent in the very small errors (show on a log scale) of the QRK2 trajectory. The solution to $x(t)$ is not monotonic everywhere and the input from $y$ is quantized. Nonetheless, it exhibits errors that are generally improved over the QSS1 method.

Figure 2 shows the error of the first calculated points past $t = 15$ for $x$ and $y$. It is clear from the plot that the error in $y$ follows a curve proportional to $D^2$, as anticipated by the theorem. The error for $x$ is proportional to $D$, in agreement with the more general result.

## 4 CONCLUSIONS

For a single system over a monotonic segment the QRK2 method exhibits error proportional to $D^2$. For a system of equations, the QRK2 method communicates the value of the state variable at each threshold crossing. The error in the receiver's approximation is proportional to $D$ between updates, with the constant of proportionality depending on the partial derivative of $f$ with respect to the arriving variable. This error contributes to the overall integration error being proportional to $D$, which we see for $x$ in the test problem.

It is conceivable that this contribution to the overall error could be eliminated by transmitting other information. For example, the QSS2 method (Kofman 2002) communicates the state variable value and its first derivative at each quantum crossing. In principle, this could permit second order accurate extrapolations of the communicated value in time and thereby reduce this contributor to the overall error.

Let us suppose that this can be accomplished, and that by doing so we are able to obtain errors proportional to $D^2$ for systems of equations over monotonic segments. Then the remaining difficulty appears to be at inflection points where the sign of the derivative changes. Suppose that the derivative at this point switches from positive to negative. At such a point, we increase the state variable by a positive quantity bounded by $D$ and may overshoot the inflection point by the same amount. Hence, we induce an error bounded by $D$ rather than $D^2$.

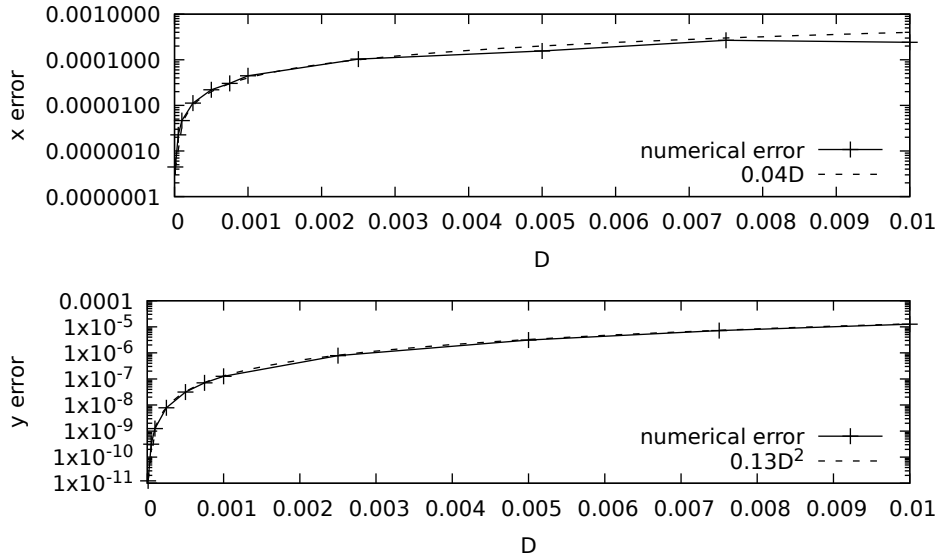


Figure 2: Numerical error as a function of $D$ for the system of test equations.

The linear implicit QSS method (Migoni, Bortolotto, Kofman, and Cellier 2013, Pietro, Migoni, and Kofman 2019) offers an example of detecting an inflection point as part of a quantized state numerical integation. However, the purpose in this case is to more quickly simulate stiff systems by avoiding a rapid passing back and forth across the inflection point. The error induced by this method remains proportional to $D$. A similar effort to account for inflection points is included in the QRK2 method by approximating $f(x,y)$ with a line in $x$ but still using the single communicated value of $y$. Therefore, it is unlikely to offer a solution to the problem of inflection points.

Another avenue for investigation is the link between time stepped and quantized state methods for numerical integration. The proof in this paper relies explicitly on translating properties of the RK2 method into bounds on the error of the QRK2 method. In prior work, a similar but more thoroughly investigated link was found between the explicit Euler method and the QSS1 method (Nutaro and Zeigler 2007). There it was found that the resultant of the discrete event system that defines QSS1 satisfies the stability criteria of explicit Euler. Continued investigation of these relationships could yield important new insights.

Future work to obtain quantized state methods with errors proportional to $D^2$ everywhere might begin with an investigation of the above questions. Moreover, it appears that such a method, if one exists, would necessarily be found outside the framework of perturbed linear systems that gives us well established theoretical error bounds for the known QSS variants; again, see (Kofman 2005, Cellier and Kofman 2006). On the one hand, this may indicate that such methods don't exist. At the same time, the demonstration of $D^2$ errors in a special case, one not anticipated by existing theory, suggests that something interesting remains to be discovered.

## ACKNOWLEDGEMENTS

This manuscript has been authored by UT-Battelle, LLC under Contract No.DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication,acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of the manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public

access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

This research was supported in part by an appointment with the National Science Foundation (NSF) Mathematical Sciences Graduate Internship (MSGI) Program sponsored by the NSF Division of Mathematical Sciences. This program is administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and NSF. ORISE is managed for DOE by ORAU. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of NSF, ORAU/ORISE, or DOE.

## REFERENCES

Bergero, F. M., F. Casella, E. Kofman, and J. Fernández. 2018. "On the efficiency of quantization-based integration methods for building simulation". *Building Simulation* vol. 11 (2), pp. 405–418.

Cellier, F. E., and E. Kofman. 2006. *Continuous System Simulation*. Springer Link.

Kofman, E. 2002. "A Second-Order Approximation for DEVS Simulation of Continuous Systems". *SIMULATION* vol. 78 (2), pp. 76–89.

Kofman, E. 2005. "Non-conservative ultimate bound estimation in LTI perturbed systems". *Automatica* vol. 41 (10), pp. 1835–1838.

Kofman, E., and S. Junco. 2001. "Quantized-State Systems: A DEVS Approach for Continuous System Simulation". *Transactions of the Society for Computer Simulation International* vol. 18 (3), pp. 123–132.

Migoni, G., M. Bortolotto, E. Kofman, and F. E. Cellier. 2013. "Linearly implicit quantization-based integration methods for stiff ordinary differential equations". *Simulation Modelling Practice and Theory* vol. 35, pp. 118–136.

Muzy, A., E. Innocenti, A. Aiello, J.-F. Santucci, and G. Wainer. 2002. "Cell-DEVS quantization techniques in a fire spreading application". In *Proceedings of the Winter Simulation Conference*, Volume 1, pp. 542–549 vol.1.

Muzy, A., R. Jammalamadaka, B. P. Zeigler, and J. J. Nutaro. 2011. "The Activity-tracking paradigm in discrete-event modeling and simulation: The case of spatially continuous distributed systems". *SIMULATION* vol. 87 (5), pp. 449–464.

Nutaro, J. 2003. *Parallel discrete event simulation with application to continuous systems*. Ph. D. thesis, University of Arizona.

Nutaro, J. 2005. "Constructing Multi-Point Discrete Event Integration Schemes". In *Proceedings of the 37th Winter Simulation Conference*, WSC '05, pp. 267–273, IEEE.

Nutaro, J., and B. Zeigler. 2007. "On the stability and performance of discrete event methods for simulating continuous systems". *Journal of Computational Physics* vol. 227 (1), pp. 797–819.

Nutaro, J., B. P. Zeigler, R. Jammalamadaka, and S. Akerkar. 2003. "Discrete Event Solution of Gas Dynamics within the DEVS Framework". In *Computational Science — ICCS 2003*, edited by P. M. A. Sloot, D. Abramson, A. V. Bogdanov, Y. E. Gorbachev, J. J. Dongarra, and A. Y. Zomaya, pp. 319–328. Berlin, Heidelberg, Springer Berlin Heidelberg.

Nutaro, J. J. 2007. "Discrete-Event Simulation of Continuous Systems". In *Handbook of Dynamic System Modeling*, edited by P. Fishwick, Chapter 11. Chapman and Hall/CRC.

Pietro, F. D., G. Migoni, and E. Kofman. 2019. "Improving Linearly Implicit Quantized State System Methods". *SIMULATION* vol. 95 (2), pp. 127–144.

## AUTHOR BIOGRAPHIES

**RASIKA MAHAWATTEGE** is a recent PhD graduate from the Department of Math Sciences at the University of Mephis and will soon begin a post doctoral position at the University of Maryland. His research interests lie in applied mathematics, numerical methods, and fluid structure interactions. His email address is rmhwttge@memphis.edu.

**JAMES NUTARO** is Group Lead for the Computational Systems Engineering & Cybernetics Group at Oak Ridge National Laboratory. He holds a Ph.D. in Computer Engineering from the University of Arizona. His research interests discrete event systems, systems modeling and simulation, and hybrid dynamic systems. His email address is nutarojj@ornl.gov.