Quantum Simulation: Rare Event Simulation by means of Cloning and Thinning

R. G. Addie

University of Southern Queensland (addie@usq.edu.au)

Abstract: A method of rare event simulation, termed *quantum simulation* is introduced. The term *quantum simulation* is used here for this simulation method because the entire ensemble of simulations resembles the parallel universes model of quantum mechanics. Quantum simulation is a generalisation of the simulation methods known as *importance splitting* or *the restart method* and also of another rare event simulation method known as *importance sampling*. A general scheme for simulations made up of multiple threads in which each thread is assigned a weight is described and flexible rules which ensure that threads can be cloned or terminated at any time without introducing bias in simulation estimates is presented. Similar schemes known as the sequential Monte Carlo method, and the population Monte Carlo method have also been used in physics, control theory, and applied mathematics; the emphasis on rare event simulation in this paper distinguishes it from that work. A numerical example based on a simple queueing model with a Gaussian input process is used to illustrate the method and to compare approaches based on importance splitting and importance sampling.

Keywords: Importance Splitting; Importance Sampling; Simulation.

1. INTRODUCTION

Quantum simulation is a simulation method which makes use, in addition to conventional discrete event simulation procedures, the spontaneous generation of clones (copies) of simulation processes which then proceed with an independent random number stream. Processes are also thinned (killed) to ensure that the total number of processes stays within reasonable bounds, or, if desired, so that the total number of active processes at one time is fixed. The cloning and thinning rates will typically be state dependent, and the rates are chosen in such a way that events of greater interest occur more frequently, and therefore may be studied with greater accuracy. As a special case, cloning may occur *whenever* a process arrives in a certain state or set of states.

This method of simulation is similar to *importance splitting*, also known as the *Restart Method*, (Villén-Altamirano & Villén-Altamirano 1991, Akyamac, Haraszti & Townsend 1999, Görg & Füss 1999). In the Restart Method, simulations which arrive at certain boundaries in the simulation state space are restarted many times, to provide increased accuracy for estimates of probabilities in certain regions of the simulation state space.

The quantum simulation framework presented in this paper is distinguished from importance splitting in that there is no use made, in this paper, of any Markov assumption, nor any assumption of finiteness of the state space, and the flexibility with which cloning and splitting rates may be tuned to in order to adapt to individual problems is maximised. Furthermore importance sampling as well as splitting is naturally incorporated into the quantum simulation framework.

The concept of *importance sampling*, eg (Lassila & Virtamo 1999), is also closely related to quantum simulation. Importance sampling makes use of an analytic formula for a *change of measure* which transforms the model under consideration into one which can be simulated quickly. The change of measure is chosen so that any statistics observed on the modified simulation can be translated back to the original model. The efficiency of importance sampling is often remarkable, and in many cases optimal, but the situations where it is applicable are limited by the need to have an analytic formula for the change of measure. A quantum simulation in which there is always precisely one thread is precisely equivalent to an importance sampling simulation. However, there are plenty of situations where it is useful to have more than one non-independent thread, and so quantum simulation is significantly different from importance sampling.

The term *quantum simulation* is used here for this simulation method because the entire ensemble of simulations resembles the parallel universes model of quantum mechanics. Perhaps we really are part of a simulation set up so that the mice can find the answer to the ultimate question, or the question for the ultimate answer!? (as in Douglas Adams' famous radio series, the Hitchhikers Guide to the Galaxy).

Another family of methods which takes a broadly similar approach is that of Population Monte Carlo algorithms (Iba 2000). Indeed, the term Quantum Monte Carlo is sometimes used in relation to some of these algorithms, and they are applied in some cases to problems in Quantum Mechanics. A fair description of the present paper is that methods in the style of (Iba 2000) have been used for problems from and with the goals and the motivation in the framework of (Villén-Altamirano & Villén-Altamirano 1991, Akyamac et al. 1999, Görg & Füss 1999).

Aside from bringing the techniques of Population Monte Carlo algorithms into the domain of traffic simulation studies, some innovations of a more methodological nature have also been contributed:

(i) the method developed here makes no Markovian assumptions; it is essentially a simulation method, rather than an eigenvector computation, as in (Iba 2000);

(ii) the key defining property of the simulations in the present approach is the consistency property, stated below as (2). Making use of this key property as the defining property of the simulation technique reduces the need to define a range of techniques and facilitates the use of a broader range of methods.

It seems that the range of techniques which are valid and useful, and justified by reference to the consistency principle, is much richer than almost anyone would imagine at first glance. There is, for example, a widespread view that the importance sampling concept encompasses importance splitting, as well, perhaps, as all other useful rare event simulation techniques; the former view is implicit in (Haraszti & Townsend 1998) for example. However, a little experience with population Monte Carlo methods (which are, in a sense, the natural generalization of importance splitting) should convince most researchers that the idea that all rare event simulations can be viewed as importance sampling is only tenable in an extremely limited and technical sense, if at all.

The observed processes generated in a quantum simulation are not individually unbiased and neither are they independent from each other, so simple uniform averages should *not* be used to estimate parameters. An unbiased estimator of event probabilities, or expectations of random variables is always available, however, in a quantum simulation.

This unbiased estimator is produced in a standard manner for all event probabilities and expectations. Each *thread* in the simulation is accompanied by a number, its *weight*, denoted by $p_i(t)$ for the ith thread at time t, for example.

The estimation rule which applies to any quantum simulation is as follows. Suppose that the original process is X_t and the quantum simulation exhibits k threads, $X_t^{[1]}, \ldots, X_t^{[k]}$, with weights $p_1(t), \ldots, p_k(t)$ at time t. Then, for any function, f, on the state space of the simulation,

$$\sum_{i=1}^{k} E\left(p_i(t) f(X_t^{[i]})\right) = E\left(f(X_t)\right), \qquad (1)$$

for all $t > 0$. The function f may be defined on the entire path of the process, not just on its value at t. A more general statement of this rule is given later which makes this more explicit.

If there was precisely one thread, this rule would imply that the weight, $p_1(t)$ at time t, was precisely the Radon-Nikodym derivative of the probability measure of X_t with respect to that of $X_t^{[k]}$ at this particular point. This is the estimation rule which is used in an importance sampling simulation, so, a quantum simulation which has precisely one thread at any time is precisely equivalent to importance sampling.

2. QUANTUM SIMULATION

Definition 2.1 *A* quantum stochastic process *(QSP) is a collection of stochastic processes (which we shall call* threads*),* $\{X_t^{(i)}\}_{t\in[s_i,f_i]}$, $i \in I$, *together with their* weights, $p_i(t)$, $i \in I$, *and a* prior function, $\phi : I \to I$ *(which indicates, for each thread, which thread* precedes *it in time). The numbers s_i and f_i denote the times when thread i starts and finishes.*

These weights often have the property that

$$\sum_{i\in I \& s_i \le t < f_i} p_i(t) = 1, \qquad t \ge 0,$$

however, this is not essential.

The index set, I, is always finite and at any time, t, we expect the total number of *active* simulations to be significantly less than the total number of elements in I. When one stochastic process (or simulation) stops, in many cases, one or more other stochastic processes will continue from where this one left off. We therefore need a mapping, $\phi : I \to I$, which designates, for each *thread* of which *prior* thread this thread is a continuation. Thus, for any $i \in I$, $j = \phi(i)$ is another thread such that $f_j = s_i$. There may be more than one thread, i, such that $j = \phi(i)$, which is meant to indicate that the thread j has *fathered* a collection of *children*. The sum of the weights of the collection of all the children of any thread which has children should equal the weight of the father thread, i.e.

$$\sum_i \{p_i(s_i+) : \phi(i) = j\} = p_j(f_j-),$$

in which $t+$ denotes a value infinitessimally larger than t and $t-$ denotes a value infinitessimally smaller than t.

It is also possible that a thread may terminate and not leave behind any children. In this case, the *weight* of the terminating thread will need to be distributed amongst some other threads, in order that the property (1) is preserved.

The function, ϕ and the weights p_i are required for any QSP, even if it is simply a random number stream. In practise the function ϕ does not need to be used explicitly. The reason why we *might* need to make use of this function is that the definition of an event could, in principle, force consideration of a range of values which extends back to times earlier than the start of the present thread. In such a case, we shall need to trace the process back through this point where the cloning occurred. This is where the *prior function* has its role. In many cases the events of greatest interest do not need to make use of the prior function. On the other hand, when we need to prove that a quantum simulation has certain properties or that certain procedures should be used, we shall need to make use of the prior function, because formally speaking, it is an essential feature of the quantum simulation.

2.1 Consistency Property

We want a QSP to be able to substitute for a normal stochastic process, i.e. any use to which a conventional process (or simulation) can be put, there should be a standard way to use a QSP in the same way. The following *consistency property* is required to hold in order that we can substitute a QSP $\left\{\{X_t^{[i]}\}_t : i \in I\right\}$ for a conventional stochastic process $\{X_t\}$.

Definition 2.2 *Let $\mathcal{F}_t$ denote the space of measurable sets defined in terms of the past of a process $\{X_t\}_t$, which we shall denote also by $\sigma\{\{X_t\}_t\}$, let $I(t)$ denote $\{i \in I : s_i \leq t < f_i\}$, and for each $i \in I(t)$ define the stochastic process $\{X_t^{\{i\}}\}$ as the concatenation of the thread i together with the sequence of successive* prior *threads, and let $\mathcal{F}_t^i = \sigma\left\{\{X_t^{\{i\}}\}_t\right\}$, the sigma algebra of events defined in terms of the past of the process $\{X_t^{\{i\}}\}$. All the sets $\mathcal{F}_t^i$ are isomorphic to $\mathcal{F}_t$ and the isomorphism will be denoted by $\Phi_t^{(i)} : \mathcal{F}_t \to \mathcal{F}_t^i$. This isomporphism is uniquely defined by the fact that $\Phi_t^{(i)}(\{\omega : X_{t_1}(\omega) \in A_1, \ldots X_{t_n}(\omega) \in A_n\}) = \{\omega : X_{t_1}^{\{i\}}(\omega) \in A_1, \ldots X_{t_n}^{\{i\}}(\omega) \in A_n\}$. Then, a QSP is* consistent *with the stochastic process $\{X_t\}_t$ if, for all $t \in \mathbb{R}$, $E \in \mathcal{F}_t$,*

$$E\left(\sum_{i \in I_t} p_i(t) \chi_{\left\{X_t^{(i)} \in \Phi_t^{(i)}(E)\right\}}\right) = P\{X_t \in E\}. \qquad (2)$$

Informally, this definition says that a QSP is consistent with a conventional stochastic process so long as the estimates formed by means of weighted averages, using the weights, always have the same mean as the corresponding estimates in the context of the conventional stochastic process. We shall always assume that a QSP is consistent in this manner with some conventional stochastic process. We hereby adopt this property as a *defining characteristic* of a quantum simulation of a process $\{X_t\}$.

2.2 Cloning and Thinning

The procedure of cloning and thinning may be applied to a conventional stochastic process or to a process which is already a QSP. In order to allow for a sequence of cloning and thinning steps, it is important to describe how it is carried out on a process which is already a QSP.

Suppose we have an existing QSP (or a conventional stochastic process), $\left\{\{X_t^{[i]}\}_{i \in I}\right\}$. *Cloning* of this process is the process of replacing some of the threads by two or more threads. Suppose the thread to be cloned is $\{X_t^{[i]}\}_{t \in [a,b]}$ and the cloning is to take place at time $\tau \in (a,b)$. We shall replace this thread by the two threads $\{X_t^{[i']}\}_{t \in [\tau,b]}$ and $\{X_t^{[i'']}\}_{t \in [\tau,b]}$ in the period of time after τ. The statistical evolution of these processes is chosen in such a way that the processes formed by adjoining $\{X_t^{[i]}\}_{t \in [a,b]}$ up to time τ and $\{X_t^{[i']}\}_{t \in [\tau,b]}$ or $\{X_t^{[i'']}\}_{t \in [\tau,b]}$ afterwards are statistically identical to the process $\{X_t^{[i]}\}_{t \in [a,b]}$. In a simulation, we can ensure that the two threads are statistically identical to the original process simply by using the same software, but with different random numbers.

The *weights* of the processes with indices i' and i'' must add up to the weight of the process with index i and the *prior function* for the cloned process is identical to the prior function for the original process (which would have been empty if the original process was actually a conventional stochastic process), together with the assignments

$$i' \longmapsto i,$$
$$i'' \longmapsto i.$$

The two threads which replace thread i after time τ will, typically, be chosen to evolve independently. As a consequence there is a gain of statistical efficiency from this time on (at the cost of having to undertake more work on these two simulations than previously on only one). The time τ can be chosen arbitrarily, and also the choice of thread to be cloned can be chosen in any way at all. Typically these choices would be made to enhance statistical accuracy in a particular region of the state space explored by the process under study.

Note: If the QSP $\left\{\{X_t^{[i]}\}_t : i \in I\right\}$ is consistent with the stochastic process $\{\{X_t\}_t\}$, then so is any *cloned* QSP $\left\{\{\widetilde{X}_t^{[i]}\}_t : i \in I'\right\}$ obtained from $\left\{\{X_t^{[i]}\}_t : i \in I\right\}$ by one or more *clonings* as defined in the previous paragraphs.

Proof of this consistency result is simple. The proof reduces, by induction, to the case where there is only one cloning. And consistency in the case where there is only one cloning follows directly from the fact that the two threads which replace one, in the interval $[\tau, b]$, are statistically identical to the original thread and therefore any weighted average of observations from these two threads will be consistent with the original simulation so long as the sum of the weights in the average is one.

Thinning is easy to define. It occurs when a certain thread is terminated before the simulation has completed. There is no need to alter the prior function. The weight of the thinned thread must, however be reassigned. If this is not done or if it is done in the wrong way, the consistency property will no longer hold.

The basic rule for thinning is that if one thread is chosen to remain from a group of candidates, the identity of this thread which remains must be selected randomly with a probability in proportion to its weight. The group of *candidates* for thinning may be chosen arbitrarily, and in this way the *state* of the process may be taken into account. For example, threads to be thinned might always be chosen from among those straying into a "boring" part of the state space. Alternatively, if desired, precisely one thread could be chosen to be thinned, from a set of candidates. In this case, the probability of selecting a thread should be in proportion to the sum of the weights of the *other* threads. Other variations along these lines are possible. As always, the constraint which distinguishes acceptable from unacceptable techniques is that the equation (2) must be preserved.

So long as thinning is carried out in this manner, the consistency of the QSP with the underlying model can be preserved. Incidentally, one of the advantages of using this rule for thinning unwanted threads is that they can be eliminated fairly aggressively thereby avoiding spending unnecessary computation time on threads which are of little interest.

3. COMPARISON OF APPROACHES TO RARE EVENT SIMULATION

The relationships between the three types of rare event simulation discussed in this paper are as follows:

Quantum simulation is a *generalisation* of importance splitting *and* of importance sampling. The fact that quantum simulation generalises importance sampling follows from a consideration of the estimation procedures used in each method. In quantum simulation, estimates are formed from a weighted average over results from individual *threads*. In Importance Sampling, the estimate is formed by weighting observed results according to the Radon-Nikodym derivative. The weights in the quantum simulation stand in place of this Radon-Nikodym derivative. The consistency rule for a quantum simulation is the natural generalisation for the case where the observations are made on multiple threads of the rule which is used in Importance Sampling.

To show that quantum simulation generalises importance *splitting*, it is necessary to define importance splitting more precisely, which is not feasible in the limited space available in this paper. Suffice it to say that in importance splitting approach, cloning tends to occur at precise state space boundaries, and conditional distributions of state space probabilities are estimated at these boundaries, and then used to make estimates of event probabilities and expectations. However, careful examination of these calculations shows that they reduce to the weighted average defined in (2). More details can be found in (Addie 2001).

Not all quantum simulations can be viewed as arising as an importance sampling type of simulation, or an importance splitting type of simulation. In the case where more than one *thread* is being simulated at once, it is not possible to identify the weights with Radon-Nikodym derivatives, which shows that quantum simulaton is not a type of importance sampling simulation. The fact that both importance sampling and importance splitting can be put into the common framework of quantum simulation also distinguishes it from either of these two techniques.

It is possible, however, to simulate the effect of importance sampling by means of the importance splitting type of approach – this can be done by simply undertaking splitting in proportion to an appropriate Radon-Nikodym derivative. However, when the distorted probability measure differs from the original probability model by more than a certain degree, the process of simulating an importance sampling simulation by an importance splitting simulation becomes impractical.

Important sampling simulations are generally faster than importance splitting simulations. This flows from the fact that analytical information which is known about the model under study is taken into account in an importance sampling simulation. Also, it is often the case, although not essential, that an attempt will be made in an importance sampling simulation to establish the optimal amount of distortion of the original probability model, whereas in importance splitting style simulations an attempt at optimisation in this manner is often not possible, because less analytical information about the model is available.

4. NUMERICAL EXAMPLES

To start with, let us consider a simulation of a queue, the input to which is a series of Gaussian numbers with mean -2 and standard deviation 1. These were simulated using quantum simulation in a form related to importance sampling but which uses multiple simultaneous threads and cloning and thinning to select which threads should be continued (so that, for example, a genetic algorithm can be used to find, adaptively, the optimal level of distortion at any time during a simulation). One thread evolves *naturally*, while all the others evolve in a distorted manner, as in importance sampling. Thinning is used to eliminate threads when their weight becomes insignificant and cloning of the natural thread is used to generate replacement threads. In this way, the entire collection of threads is able to track the dynamics of the natural simulation even though most individual threads follow a path which leads inexorably to more and more unlikely events.

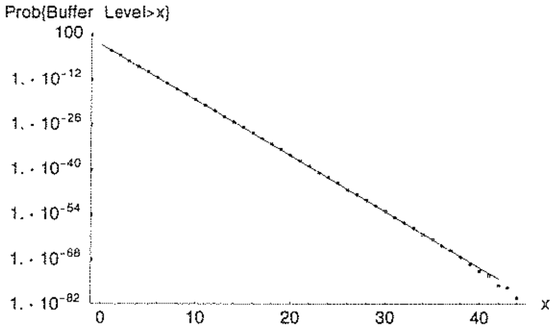

Figure 1: Simulation results and theoretical estimate for the complimentary waiting time distribution in a Gaussian queue (mean -2) – importance sampling case

The results are plotted, together with the expected results from theory (according to (Addie, Mannersalo & Norros 2001)), in Figure 4.. The simulation results and the theoretical results overlap almost perfectly. This simulation included ten simultaneous threads, one of them completely conventional, and the others distorted in the manner of importance sampling.

From large deviations (Addie et al. 2001), we know that the optimal importance sampling simulation will make use of distorted probability measure, γ, in which the input to the queue becomes IID Gaussian with mean 2 (instead of -2) and with unchanged standard deviation. If any other distortion of the input process is used, the results achieved will be less accurate.

The duration of this simulation was 100 cycles of input, buffering, and service, all of which is modelled by the simple equation:

$$B_{t+1} = \mathrm{Max}\,(B_t + X_t, 0)$$

in which $\{B_t\}$ denotes the contents of a buffer at time t, starting with $B_0 = 0$, X_t denotes the *net* input to this buffer, which is an independent and identically distributed sequence of Gaussian random variables with mean -2 and standard deviation 1.

Figure 2 shows the manner in which the weight of *one* of the distorted threads changes during the simulation. The weight of a thread reduces steadily due to the fact that the Radon-Nikodym derivative tends to decrease steadily. Soon after the weight of a thread becomes the lowest of all the weights of threads in the simulation, the thread is likely to be killed, at which time it appears to have returned to a weight near one, in the plot shown in the figure. The plot then shows the weight of a thread which has been produced by cloning the top thread. This thread therefore starts with weight close to 0.5 and this weight steadily reduces to a value near 10^{-50}, again.

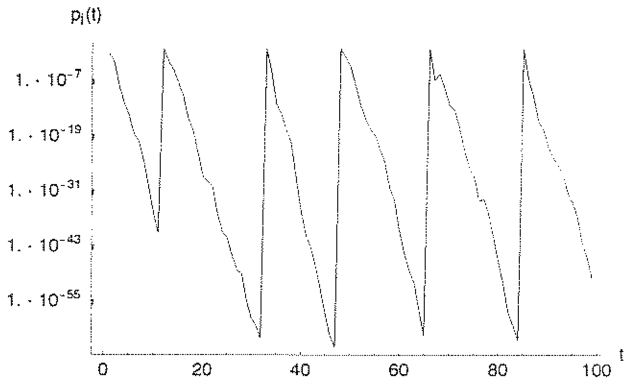

Figure 2: Example of Variation in Weight during a Quantum Simulation of a Gaussian Queue

The elapsed time for the simulation discussed here to complete, as implemented in mathematica, running under Linux, was less than one minute. However, it is meaningless to talk about how much faster this simulation is than a conventional simulation, since a conventional simulation would not be able to achieve the indicated accuracy levels by now even if it was started just after the big bang.

These simulations were repeated using quantum simulation in a form based entirely on cloning and thinning, i.e. no importance sampling or distortion of the input process was adopted. The simulation results and the theoretical results still overlap quite well although not as well as in the importance sampling case. This can be explained by the fact that the distorted distribution of the input to the queue is not *optimal* in this case. The time taken to carry out this simulation was much longer and the accuracy was greatly reduced. Approximately 1000 times as many random numbers had to be generated in order to simulate the same period of time and the simulation time was at least doubled.

The knowledge of a large deviations principal for this particular system is still being used in the present simulation, but in an indirect manner, and in a manner which could be avoided, with some effort. One way to avoid the use of the known large deviations principal would be to *infer* the appropriate large deviations principal from the simulation, dynamically. The thinning phase of a quantum simulation can be used to have this effect, as was touched upon earlier.

In Figure 5., estimates of the complimentary queueing distribution function obtained after 20, 40 and 500 iterations of a quantum simulation of a Gaussian *correlated* queueing system are shown. In this case, the input process had mean -1.7, standard deviation 1.22, and autocovariance $1.49, 0.72, 0.18, -0.2, 0, \ldots$. The theoretically expected stationary complimentary distribution function is also shown, for comparison. Confidence intervals are shown for the estimate obtained at iteration 500. It is apparent that by iteration 500 the estimated distribution is very close to the stationary distribution.

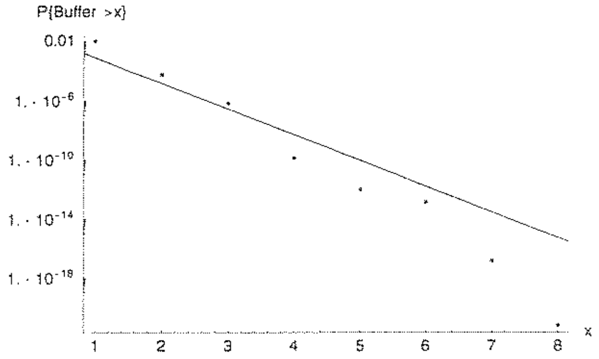

Figure 3: Simulation results and theoretical estimate for the complimentary waiting time distribution in a Gaussian queue (mean -2) – importance splitting case

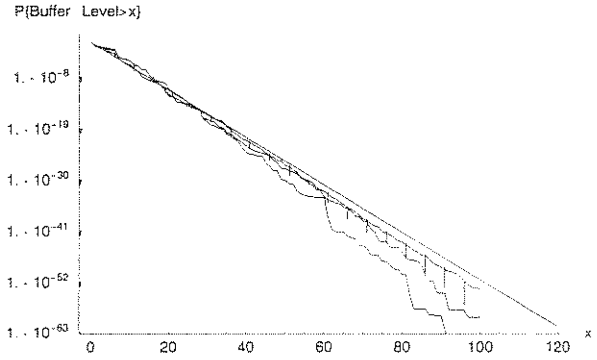

Figure 4: Transient Simulation results and theoretical estimates for the complimentary waiting time distribution in a correlated Gaussian queue, at times 20, 40, and 500 after an empty initial state

5. CONCLUSION

Quantum simulation has been introduced and compared with the well known existing methods of importance sampling and importance splitting and it is shown that both of these methods can be viewed as special cases. A quantum simulation with precisely one thread is equivalent to importance sampling and a quantum simulation in which cloning occurs at certain regions of the state space is equivalent to importance splitting.

In addition, rules for *thinning* have been defined which ensure that the quantum simulation estimators are always unbiased; and it has been demonstrated how importance sampling style simulations can be achieved by means of simulations which use *splitting* and thinning.

Numerical examples have been used to demonstrate the effectiveness of quantum simulation and to demonstrate the relationships between importance sampling importance splitting and quantum simulation. From these examples, and more theoretical considerations, it is clear that in cases where sufficient analytical information is available to carry out importance sampling, it is likely that importance sampling will produce more accurate results in a shorter time. However analytical information of this sort is often not available, and when it is available, the models in question are often sufficiently well understood that simulation is no longer necessary.

Quantum simulation is a general purpose technique which can be used to increase the speed and accuracy of estimation of rare events in any simulation. Quantum simulation is currently being used to model a communications system carrying traffic modelled in a realistic manner which is currently difficult to either analyse or simulate with adequate accuracy.

6. REFERENCES

Addie, R. G., Quantum simulation - rare event simulation by means of cloning and thinning, *in* 'Proceedings of Fourth Operations Research Conference of the Australian Society for Operations Research, QLD Branch', 2001.

Addie, R. G., P. Mannersalo, & I. Norros, 'Most probable paths and performance formulae for buffers with gaussian input traffic', *European Transactions on Telecommunications*, 2001.

Akyamac, A. A., Z. Haraszti, & J. K. Townsend, Efficient rare event simulation using DPR for multidimensional parameter spaces, *in* P. Key & D. Smith, eds, 'Teletraffic Engineering in a Competitive World', Vol. 3B of *Teletraffic Science and Engineering*, 16th International Teletraffic Congress, Elsevier, 1999.

Görg, C. & O. Füss, Simulating rare event details of atm delay time distributions with restart/lre, *in* P. Key & D. Smith, eds, 'Teletraffic Engineering in a Competitive World', Vol. 3B of *Teletraffic Science and Engineering*, 16th International Teletraffic Congress, Elsevier, 1999.

Haraszti, Z. & J. K. Townsend, The theory of direct probability redistribution and its application to rare event simulation, *in* 'Proceedings of IEEE Infocom', 1998.

Iba, Y., Population-based monte carlo algorithms, Technical report, Institute of Statistical Mathematics, Tokyo, 2000.

Lassila, P. E. & J. T. Virtamo, Efficient importance sampling for monte carlo simulation of loss systems, *in* P. Key & D. Smith, eds, 'Teletraffic Engineering in a Competitive World', Vol. 3B of *Teletraffic Science and Engineering*, 16th International Teletraffic Congress, Elsevier, 1999.

Villén-Altamirano, M. & J. Villén-Altamirano, Restart: a method for accelarating rare event simulations, *in* J. W. Cohen & C. D. Pack, eds, '13th International Teletraffic Congress', International Teletraffic Congress, North-Holland, 1991.