Disappointments and Delights, Fears and Hopes induced by a few decades in Performance Evaluation

Raymond Marie

Disappointments and Delights, Fears and Hopes induced by a few decades in Performance Evaluation

Raymond A. MARIE

IRISA, Rennes 1 University,UEB, Campus de Beaulieu, 35042 Rennes Cedex, France
marie@irisa.fr

Abstract. The elements of modelling in general and of performance evaluation of discrete event systems (DES) in particular have undergone a tremendous transformation during these last four decades. The aim of this paper is to look back over all this evolution, trying to retain some particular experiences from the past. I will try to classify these elements according to what I have perceived as their positive or negative potentialities. All the views expressed are my own and entirely subjective. Nothing will be proven since there will be no theorems. We first enumerate a list of events or situations which have occurred during these four decades and which I regard as positive. An opposite set of negative arguments will follow. Then, I will enumerate a list of risks that, from my personal perception, represent the dangers for the domain of modelling and of performance evaluation of systems in the field of computers and telecommunications. Finally, some suggestions to preserve the quality of the expertise of the community will be proposed.

1 Introduction

During these last four decades, the elements of modelling in general, and of performance evaluation of discrete event systems (DES) in particular, have gone through a tremendous transformation. The special event motivating this meeting provides an occasion to look back at this evolution, trying to retain some particular experiences from the past. I will try to classify these elements according to what I have perceived as their positive or negative potentialities, inevitably from a personal and subjective standpoint. Nothing will be proven since there will be no theorems presented. In the following section, I will enumerate a list of events or situations that occurred during these four decades and that I consider to have been positive. On the other hand, over the same period, there have been some developments which I consider negative as discussed in Section 3. I will enumerate in Section 4 a list of risks which represent dangers for the domain of modelling and of performance evaluation of systems in the field of computers and telecommunications. Finally, I will suggest in Section 5 some ideas to preserve the quality of the expertise of the community.

2 Delights

In the beginning, our scientific ambition was limited by computing power. We were using our imagination to look for approximations in order to reduce a state space to a few hundred states ; or even less if the model was used as a submodel! Often, we were using all the central memory resource of the computer and that required us to be the single user of the main frame ; this type of privilege was only given during night hours (after midnight !).

We were excited by the novelty of the discipline, combining informatics and telecommunications. From time to time, we were lucky enough to get success stories with simple models. For example, the M/M/1 queue with *processor-sharing* discipline was surprisingly good at representing the congestion phenomena on a main frame. In fact, it was realized later that with the processor-sharing discipline the steady state distribution is invariant with respect to the service time distribution; note also the time sharing policy which was used to execute the list of jobs on a main frame (a rare resource at that time !) has the processor-sharing discipline as asymptotic behavior. This was also the time when a small product form queuing network was able to capture the main factors of a computer room covering many hundreds of square-feet in order to predict the response time with reasonable accuracy.

As already mentioned, we were (almost) all young and this situation gave us the opportunity to take on national and international responsibilities before the average age encountered in other scientific communities. We set-up few, but high quality, international conferences which gave us the opportunity to exchange ideas at a time where the Internet did not exist and where postal mail was taking weeks to arrive from the other side of the world. Throughout the following periods, step by step, use of new concepts (e.g., timed Petri nets, neural networks) also stimulated our research activities.

3 Disappointments

On the other hand, we can observe now that huge amounts of available computing resources increase the trend to solve models through simulation and do not encourage researchers to look for tractable analytical solutions.

Sometimes, we see young researchers "reinventing" new methods that were introduced 20 years before just because they often do not look at publications more than 10 years old, especially if they are working on a new application domain such as telecommunications.

The architecture designers do not always take the performance evaluation people seriously. A pessimist could see there the consequence of an eventual competition. The designers are convinced that they know what they are doing ; as a consequence, they sometimes prefer to increase the number of resource units when the expected performances are not attained rather than looking for other solutions.

From an academic point of view, a general trend is that a certain number of courses, very useful for our scientific field, have been withdrawn from the

standard curriculum of studies in computer sciences. I am thinking of linear algebra, probabilities, stochastic processes such as Markov processes and linear or non-linear optimization. Note that sometimes you find the theory exposed in the particular context induced by the topic of the course (e.g., dynamic programming introduced in the context of graph theory) and taught this way, the theory loses all its generality. This has a negative impact on the learning process of the student.

Actually, the number of international conferences which are organized per year seems to be continuously increasing. Is it because this is good for Science or is it because each researcher wants to write on his curriculum vitae that he has been general chair of some international conference/symposium/workshop ? Therefore the question becomes : when does a lot become too much ?

4 Fears

Let us now try to present a list of different dangers that are, as I see it, threatening the quality of the work of a modeler of DES.

1. Bad comprehension of the system (this applies to both simulation and analytical approaches) Let us consider the following example ; a processor has to execute two types of job according to a preemptive priority discipline. In order to simplify the example, let us suppose that the service times are exponentially distributed with respective means $1/\mu_1$ and $1/\mu_2$ for class one and class two.
   Let λ_1 and λ_2 denote the respective arrival rates of the two Poisson processes. Let us consider the following numerical values ; $\lambda_1 = 1$, $\mu_1 = 100$, $\lambda_2 = 0.01$ and $\mu_2 = 1$. Then, if the modeling person sees the processor discipline as a preemptive priority repeat different discipline, this person will predict a busy rate of 2 percent for the processor. While, if the real behavior of the system corresponds to a preemptive priority repeat identical discipline, the processor will not be fast enough to allow the treatment of the amount of processing work (since the stability condition, $\lambda_1/\mu_2 < 1$, is not satisfied, the system will blow up). In order to illustrate this example by a figure, let us introduce three complementary notations. Let $\mathbb{E}[T]$ denote the expectation of the total time needed by the server in order to serve a class two customer. Let $\mathbb{E}[B]$ denote the expectation of the service time of a class two customer. Finally let α denote the expectation of the number of preemptions during the service of a class two customer. Figure 1 shows the ratio $\mathbb{E}[T]/\mathbb{E}[B]$ as a function of α. This ratio stays constant if the discipline is a *preemptive priority repeat different* one (because of the memoryless property of the exponential service time distribution). But, if the discipline is a *preemptive priority repeat identical* one, this ratio tends to infinity as soon as α tends to one.
   This second situation may arise if the jobs of the preempted class correspond to executions of different files. Even if a set of execution times corresponding

to the different files of class two (each file having a given execution time) can be seen globally as fitting an exponential distribution, once the execution of a particular file is preempted, its successful execution will need a constant time corresponding to the execution of a constant number of bytes. Therefore the modeling person may mix the two cases corresponding to different situations, especially if the service time is exponentially distributed (keeping in mind the memoryless property of the exponential distribution). Additionally we plotted on the figure the ratio for the special case of the constant service time to show a rare case where a randomness behavior (mixed line) looks better than a non-randomness behavior (dashed line). Note that it can be proven that the two priority disciplines give the same result in the special situation where the service time is constant. The scientific derivation of the functions corresponding to these curves can be found in [1].

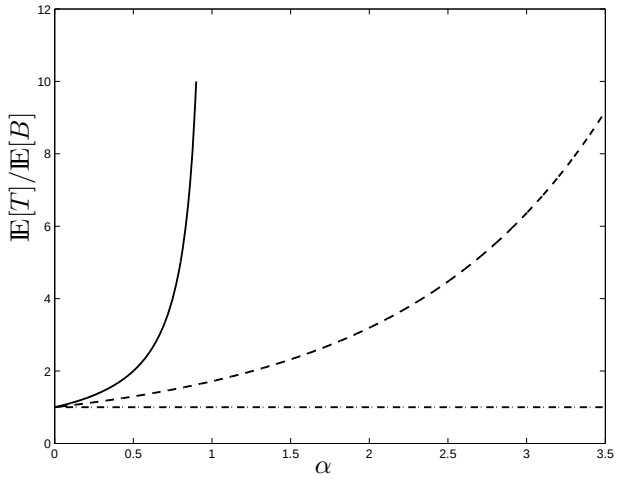

Fig. 1. Ratio $\mathbb{E}[T]/\mathbb{E}[B]$ as a function of α for the *preemptive priority repeat identical* discipline case (solid line), the *preemptive priority repeat different* discipline case (mixed line), and the special case of a constant service time (dashed line).

2. Bad mastery of approximations. There are different categories of approximations ;
   - On the one hand we have approximations at the level of the modeling step.
     For example, we assume that the service time is exponentially distributed while this is not true in reality. This is a classic approximation which is done consciously. The consequence of such an approximation depends generally both on the modeled context and on the performance parameters.

Another frequent approximation that is sometimes done unconsciously corresponds to the fact of considering that two events are independent while they are not. This latter approximation is often dangerous because the influence of non independence is generally underestimated. Some examples taken in the context of dependability are convincing. We can exhibit relative errors on the unavailability of several thousand per cent. We can exhibit cases with an unbounded relative error limit on unavailability when time tends to zero.
In order to illustrate this last assertion, let us consider the case of a complex architecture in which there exist multiple copies of a single element type. We are concerned by the probability that, at the end of a mission time T, the system is not available. We assume that the reliability function of each element of the system is exponentially distributed. If we assume that any breakdown of an element is independent of the other breakdowns, the reliability of the global system can be (more or less easily) exactly determined. Things change if, in order to increase the availability of the system, some extra spares are put on the shelf at the start of the mission by the people in charge of the system. In such a situation, the determination of the unavailability of the system at time T becomes more challenging, even if we disregard the exchange time. In order to give numerical data, let us consider the simplest example of a two element redundancy associated to one spare element illustrated on figure 2. If the spare did not exist, the unavailability of this two element system would be given by $(1 - e^{\lambda T})^2$ where λ denotes the failure rate of one element.

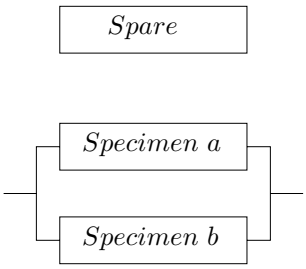

Fig. 2. A minimal system.

Using a Markovian model that disregards the exchange time, it is possible to compute exactly the probability $\mathbb{P}(A)$ (resp. $\mathbb{P}(B)$) that the element a (resp. b) is down at time T by lack of spare. We get $\mathbb{P}(A) = \mathbb{P}(B) = (1 - e^{\lambda T})^2$. Having these probabilities, a naive approach would be to consider that the unavailability of this two element system is $(\mathbb{P}(A))^2$. This is the kind of approach that would be easily adopted in the general case where the different elements of the same type could be at quite

different locations on the reliability diagram of the global system. This approach assumes independence between the different elements would consider a reliability $(1 - \mathbb{P}(A))$ for each of the different elements of the same type. While, for the two element redundant system, it is easy to find that the exact unavailability at time T equals $1 - e^{-\lambda T}[4 - e^{-\lambda T}(3 + 2\lambda T)]$. Figure 3 shows the ratio of the exact unavailability divided by the naive answer (i.e., $\mathbb{P}(A)^2$) as a function of the product λT. On this figure, we can see first that for $\lambda T = 10^{-3}$, the exact unavailability is approximately 600 times larger than the one obtained when we ignore the dependence between the two elements, this dependence being introduced by the existence of the spare element. Secondly, as suggested by the curve, it can be proven that the ratio tends to infinity when λT tends to zero. This means that the more reliable the system is, the more the unavailability is underestimated.

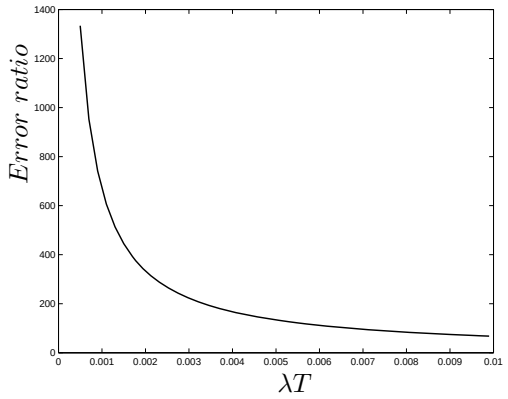

Fig. 3. Unavailability error ratio as a function of λT.

– On the other hand, we have approximations at the level of the resolution step. Whatever we are using - a direct or an iterative approximated method, it is important to know if upper and/or lower error bounds have been exhibited. For iterative approximated methods, it is also important to know if the convergence of the method has been proved ; or if nobody has proved the convergence of the method (but nobody so far has obtained a non convergent case study).

3. Use of numerically unstable algorithms. We all know that the main reason for losing accuracy is the execution of a difference of two smaller and smaller positive numbers. Such a situation arises frequently in our domain. For example we encounter it quite often when we look for the original of the generating function of a probability distribution. This is why it is always profitable to try to find an algorithm adding only positive numbers (in ad-

dition to the use of the product and division operators). In the special but important case of the use of simulation, let us say that in one way, simulating a process on a finite time interval is making an approximation ; and that too many, while simulating, forget to give this approximation by means of the confidence intervals...

4. Inadequate use of Markovian models (when is it dangerous to use Markovian models ?). The main grievance done to the use of Markovian models is that in real systems, the time durations of activities are not really exponentially distributed; although it is known that phase-type distributions can reproduce as close as necessary the non exponential distributions of the real system. However, in general, people expressing such grievances do not master the use of fictitious states and it is true that such a procedure has a significant cost, since it can drastically increase the cardinality of the state space. Otherwise there is the possibility of searching for the solution of a semi-Markovian model, but this is not always an easy task. That is why it is important to know when the steady state performance measures do not depend on the type of distributions of the time duration of activities in the real system. We are in such a favorable situation when activities do not execute simultaneously, but sequentially, according to a stochastic routing using probabilities p_{ij} (activity j starts after finishing activity i with probability p_{ij}). In such a (common) situation, the steady state distribution of the semi-Markovian process modeling the real system with non exponential distributions is the same as the one of the Markovian process obtained by taking the rates equal to the inverses of the mean durations of the activities.
   The danger would be to do this simplification while different activities may be executed concurrently. Because of this competition, for equivalent expectations of the execution times, the behavior will depend on the distributions of the random variables. For example, let us consider a simple triple module redundancy with hot repair facility. If the life-time and the repair-time of an element are constant, the redundant system will be always available (supposing the repair-time shorter than the life-time). While if the life-time and the repair-time of an element are exponentially distributed, there is a strictly positive probability to see the redundant system down.

5. Lack of technical background. On the one hand, this risk is highly correlated with the first mentioned danger (bad comprehension of the system) which we will therefore not elaborate futher. On the other hand, it is worthwhile to remark that being technically good is not a sufficient condition for building good models (unfortunately).

6. Lack of scientific background. In my opinion, young people from our domain suffer in particular from a lack of background in applied mathematics and I notice that both simulation and analytical fields are concerned. With respect to the simulation field, is the researcher, involved in a project requiring many months of work, mastering confidence intervals or the special techniques of

importance splitting or of importance sampling ? Such techniques are very important for rare event studies, when the probability of an event equals, for example, 10^{-8} (in high speed telecommunication networks, highly dependable architectures, air-traffic control systems, etc). With respect to the analytical field, does the researcher master Markov regenerative processes ? or stochastic fluid models ? or process algebra ? or timed Petri nets ? or fluid timed Petri nets ? Again, we can mention the case of rare events encountered, for example, with telecommunication networks. Often we are concerned with the dimensioning of a node such that the lost probability in a buffer is lower than 10^{-8}. A realistic queuing model will not have a product form solution, and this will not be reasonable to estimate the probability of this rare event through simulation. A possible solution might be to use the technique of stochastic fluid models, approaching the behavior of the queuing model with a cost independent of the number of customers in the model.

It should be noted that this vision may be biased because of the fact that one enlarges one's expertise in the different mathematical tracks when one spends year after year in the field of modeling.

5 Hopes

Thanks to Research and Development, computational power and data storage have increased the possibilities of performance evaluation studies during the last decades. The development of libraries and of graphical interfaces has increased our productivity. However, if we compare with other industries (space, transport, nuclear), or disciplines (physics, chemistry), we should be more ambitious with respect to our evaluation tools.

There is a place for a large evaluation tool built as a set of cooperative agents including simulation agents and analytical/numerical agents. Following the ideas underlying the notion of "Internet of the future", we could think of a nice virtual machine built on a computer cloud and able to realize all the possible evaluations of performance already done once by one group of an international federation. Of course such a project would need a tremendous effort and the setting-up of standard commissions to define the tasks of tool virtual agents, interface protocols and also to standardize software developments and data structures. There would also be a need for the setting-up of independent teams testing each agent, the possibilities of each method, ranking them with respect to the specifications of the application in case of multiple choice (eg, importance sampling versus importance splitting). Progress in manipulation of UML models and in Model Checking should help the efforts in standardization.

My personal feeling is that the community of the global domain of numerical analysis has done better than we have in the structuring of the efforts. It is true that because of its generality, the task is ambitious (how to standardize the Input/Output in a general way ?) but isn't it worth doing ? Of course, such an action needs a significant financial budget but less than a large program in Physics, and its synergy effect would benefit in the long term.

6 Conclusions

Throughout all this evolution we have seen the development of libraries and of graphical interfaces, increasing our productivity, yet somehow I am pleading for more interfaces and more virtual items. But do the students still understand what the tools are doing below the graphical interfaces ? To do so is necessary to save the level of scientific knowledge ! In fact, from my point of observation, I have come to the conclusion that among the dangers listed in Section 4, the most important one is the lack of scientific background. Discussions on this topic with colleagues from different countries have shown a common agreement on the following facts :

- The attraction of scientific studies is decreasing ; why ? is it because learning theories hurts (isn't it easier to read a book on management than a book on probability theory before going to bed ?) or is it because scientific studies do not maximize the chances of getting rich ?
- The student who comes to CS wants more and more practice of work on keyboard rather than paper and pencil !

In reaction, it is our responsibility not to let the theoretical courses move from compulsory to optional positions in university curricula. But It is also our responsibility to try to make these theoretical courses more like detective stories, i.e., more gripping and more entertaining.

Again this is just a state of my personal vision and such an exercise cannot be fully objective. In addition, there are always exceptions to general observations.

At this point of the story ($E\tau\sigma\iota\ \epsilon\iota\nu\alpha\iota\ \eta\ Z\omega\eta$ [1]), it is now the responsability of the new generation to find a way to keep our young discipline as a field in which reseach remains an attractive way of life.

References

1. R. A. Marie and K. S. Trivedi, *A Note on the Effect of Preemptive Policies on the Stability of a Priority Queue*, Information Processing Letters, Vol. 24, No. 6, pp. 397 – 401, April 1987.

[1] *So ist das Leben*