Dear Editor:

Attached below is a memo that was based on my own observations. The memo illustrates, by examples, some common misconceptions in performance modeling. These misconceptions in engineering community may result in wrong expectations from customers and management. I hope that clarifying the issues raised in the memo will be helpful to the performance community.

Sincerely,

Xiren Cao
Digital Equipment Corporation
Littleton, MA 01460

Some Common Misconceptions About Performance Modeling and Validation

1 Introduction

Queueing networks and Markov processes etc. are widely used in modeling computer systems and communication networks to study their performance and reliability. To solve a real world problem, the model developed has to be validated through measured data. In this paper, we point out that in validating a model, one has to be very clear about one's claims regarding what has been validated; Too "accurate" results do not imply a correct model and usually indicates a validation problem. We discuss some common misconceptions in performance modeling and validation. We illustrate our points through examples. To capture the main concepts, the problems are simplified in these examples.

2 A Queueing Model for Performance

Suppose that we have a device (e.g., a CPU, or an I/O port) to model. Messages arrive in packets to the device, get processed, and then leave the device. The device may be very complicated, but we may consider it simply as a service station providing services. The simplest model for such a service station is, of course, the M/M/1 queue.

For an M/M/1 queue, there are two parameters to be determined: the mean interarrival time τ and the mean service time s. Suppose that we run the device and obtained some

measured data. For the sake of discussion, let's assume that during a 1000 ms period we observed 5000 packets arriving at the device. The device was busy (i.e., there is at least one packet in the device) for 500 ms during this period and was idle at the beginning of the period. From these data, we can get the parameters for the M/M/1 queue model:

$$\tau = \frac{1000}{5000} = 0.2ms, \quad and$$

$$s = \frac{500}{5000} = 0.1ms.$$

Now we have a complete M/M/1 model for our device with parameters specified. Let's analyze the performance of the device. Two main performance measure of the device are: throughput and utilization. Using our measurement and the analytical results from the model, we may get a table that looks like the following:

Performance	Analytical	Measurement	Error
Throughput(pkt/ms)	5	4.98	0.4%
utilization	50%	49.8%	0.4%

The analytical results seem very accurate. However, can we claim that the M/M/1 model has be validated by measurements? The answer is, of course, NO. The reason is: the results listed in the table do not represent any property of an M/M/1 queue. We did nothing to verify the exponential distribution of the service time, nor the Poisson arrival. Thus, using this model to predict the response time may not be accurate. In fact, the analytical results can be obtained by some simple algebraic calculations without the M/M/1 model.

But the table does verify some fundamental principles. They are

1. *Conservation law:* What comes into a device equals what comes out from that device. Namely, the device doesn't "eat" any message.

   If a device obeys the conservation law, then the throughput (output rate) must equal the input rate. Thus, the throughput can be obtained directly by dividing 5000 by 1000. The error may be due to the small processing delay, i.e., at the end of the measurement period some packets may be still in processing and have not left the device yet.

2. *Non-idling principle:* A device is not allowed to be idle when there is any message waiting for service.

   From this principle, the utilization equals the ratio of the busy time over the total observation time.

An M/M/1 queue certainly obeys these two principles, but the verification of these two principles do not justify the M/M/1 model.

The main objective of any queueing model is to capture the contention phenomena. In the M/M/1 case, the response time is the main performance which reflects the resource contention. The response time of a single server queue depends heavily on the distributions of the interarrival time and the service time. For example, the mean response time for an M/M/1 queue is twice as much as that for an M/D/1 queue. Since one can hardly determine the exact distributions in a real system, it is not realistic to expect a response time with less than 5% error.

In sum, if the modeling results look too accurate, such as the throughput and utilization indicated in this example, the results usually can be obtained by simple calculations based on first principles; no queueing model is needed. The "accurate" results do not validate the model, they just verify some fundamental principles (which no one will question). Response time is the main performance measure of contention represented by a queueing model. Similar mistakes may exist in simulation.

3 A Markov Model for Availability

Suppose that we have a single machine and observed that during a 1000 hour period the mean time between failures is 100 hours, and the mean time between recovery is 10 hour.

Let a_1 denote the up state of the machine and a_0 be the down state. From the above data we can develop a Markov model for the system. The rate of the system leaving state a_1 for state a_0 is 0.01/hr, and that from state a_0 to state a_1 is 0.1/hr.

From the simple Markov model we developed for the availability problem, we can get the analytical results: the machine up rate is 100/110 =0.91 and the machine down rate is 0.09. These results obviously match the measurement data quite well.

Again, can we claim that we have developed and validated a Markov model for the system? The answer is the same: NO!

The analytical result can also be obtained without the Markov model. To obtain the machine available rate for this observed period, no model is needed. The basic feature of a Markov model is the memoryless property, i.e., the machine failure is independent of its history and any other event in the system. The measurement data verifies nothing related to this property. To claim a Markov model is misleading.

One application of a Markov model probably is to answer the following question:

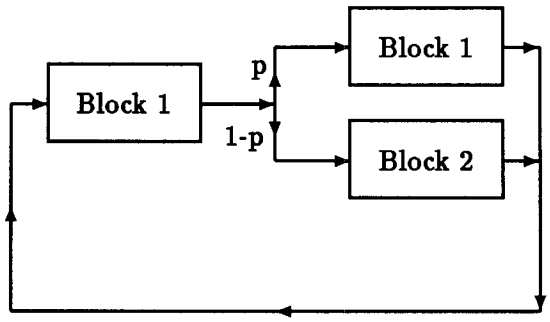

Figure 1: The Structure of a Model

What is the availability of a system consisting of two such identical machines? (The system is considered as available only if both machines are available.)

If the Markov model works for the two machine system, then the analytical results for the availability of the system is 0.91*0.91 =0.82. If the real measurement is close to this data, then at least we can claim that we have tested the independent property between two machines (not the memoryless property).

In sum, a model should tell us some results which are not direct consequences of the measurement that determines the parameters of the model.

4 Parameter Calibration

Figure 1 illustrates the structure of a model, in which the three blocks may be three sub-queueing networks or any other submodels. A task, after completing its work at Block 1, proceeds to Block 2 with probability p and Block 3 with probability $1 - p$. Examples for this branching probability are cache hit/miss ratio, local/remote read or write ratio, etc.

Assume that it is difficult to determine the value of p by measurement. Thus, we have to calibrate the value so that the model can be complete. Suppose that the measured throughput of the system is ρ and we find that, say, $p = 0.18$ yields the same throughput from the model. Then can we claim that we have validated the model of Figure 1 with $p = 0.18$ since it gives us an accurate value for the throughput? The answer is again NO.

What we have verified is in fact the following mathematical statement:

The equation $\rho = f(p)$ has a solution p for the particular ρ.

This is usually the case when the function f is continuous. It has nothing to do with the validation of our model.

5 A Word on Accuracy

We have illustrated, by examples, some misconceptions about modeling and validation. All these misconceptions are "supported" by "accurate" data, yet they are misleading. To clarify these misconceptions may help to set up a right perspective to modeling projects.

Almost all models in computer and communication systems involve stochastic assumptions, and there is no way to get the exact I/O and CPU statistics. Besides, all models are approximate because no model can capture all the details of a complex system. To expect a perfect match between the model and the real measurement is simply not realistic. This should not be a surprise since even measurement data from the same system varies perhaps more than 2-3% from one test to the other.

What is the role of modeling in engineering and management projects has always been a controversial issue. There exist two extreme but closely related views: modeling should give exact data, and modeling is useless. Expecting an unrealistic accuracy for modeling results may lead to the other extreme. The misconceptions discussed here at least partly contribute to the confusion.

A right perspective of modeling is strategically crucial to the performance community. The greatest impact of modeling lies less in providing exact data for real world problems with many uncertainties than in formulating problems, improving the understanding of highly complex issues, and analyzing the effect of different alternatives. We will not continue the discussion of this long-standing problem, and we refer the interested readers to a recent article by Corbett and Van Wassenhove and the references therein.

6 References

1. C. J. Corbett and L. N. Van Wassenhove, "The Natural Drift: What Happened to Operations Research?" *Opns. Res.* Vol.41, 625-640, 1993.