Analysis of Computer and Communication Systems with Uncertainties and Variabilities in Workload[*]

Johannes Lüthi and Günter Haring

Institut für Angewandte Informatik und Informationssysteme
Universität Wien
Lenaugasse 2/8, A-1080 Wien, Austria
Email: [luethi|rodo]@ani.univie.ac.at

Abstract

A unified approach to characterize workload parameters used to model computer and communication systems with variabilities and uncertainties in workload is presented. The use of histograms is proposed and approaches to adapt existing analysis techniques to interval-based histogram arguments are discussed. The proposed method is demonstrated by application to existing analysis techniques for single class queueing network models.

1 Introduction

The analysis of computer systems requires effective tools for predicting their performance and analyzing their behavior. Analytic techniques are often used for performance estimation because of their relatively low cost in comparison with simulations and benchmarks. A conventional analytic performance model accepts a set of single valued parameters (such as service demands for different devices) and produces a single point measure for each performance index of interest (such as the mean response time and mean processor utilization). However, the exact value of every parameter for the system may not always be known to the performance analyst leading to uncertainties in the workload characterization (WLC). For example systems are often subject to variabilities in the workload [2]. Different phases in the operation of the system under study may lead to a different set of parameters that characterize each phase.

Appropriate characterization of workload is required to capture any uncertainty or variability associated with it. We propose to characterize the parameters D_k of each aspect in the system that exhibits variability and/or uncertainty by a histogram $H(D_k)$. The histogram consists of a number of intervals and an associated probability of occurrence. Each interval is a range of values: the parameter lies in this given range (uncertainty) with the specified probability of occurrence (variability).

2 Workload characterization

A parameter X may be specified through a histogram $H(X)$ as follows:

$$X_1 = [\underline{x}_1, \overline{x}_1] : p_1; \quad \ldots; \quad X_m = [\underline{x}_m, \overline{x}_m] : p_m.$$

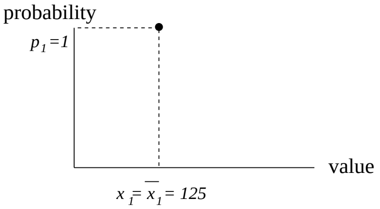
with $\sum_{i=1}^{m} p_i = 1$. Each entry in the definition of X provided above is a two-tuple, an interval $[\underline{x}_i, \overline{x}_i]$ and an associated probability p_i, i.e. with probability p_i the mean value of X lies between $\underline{x}_i$ and $\overline{x}_i$.

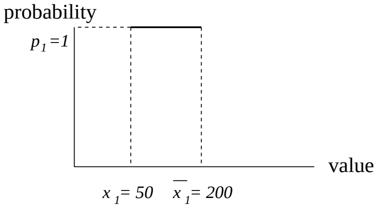
This general model can be used to represent uncertainties and/or variabilities: Uncertainties are characterized by the length of the interval, an interval of width zero ("thin interval" [13]) represents no uncertainty. Variability is described by the distribution of the probabilities p_i, with $m = 1$ obviously corresponding to a workload with no variability. The different types of workload models and the corresponding parameters are summarized in Table 1. In the following, we describe each possible type of model (with the exception of the SV case, which is already well known) and give some examples.

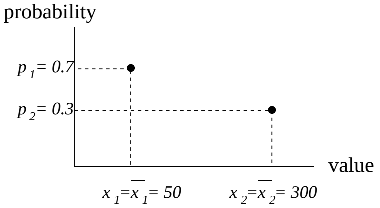
Uncertainties (UN). Associating intervals with parameters of interest is useful when uncertainties are associated with parameter values. The probability of occurrence of any value within an interval can follow any given arbitrary distribution. Consider for example software performance engineering that integrates performance modeling with the various phases of software design and implementation [14]. Uncertainties may be associated with model parameters for various reasons. For example, exact values of system parameters are often unknown in early stages of system design. Although certain uncertainties may be associated with one or more system parameters, the designer may have a good idea about the range of values associated with these parameters from previous experience with similar systems. A single interval with $p_1 = 1$ may be used to describe the range of values associated with each such parameter.

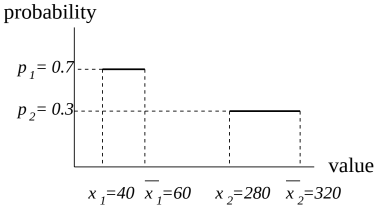
Variabilities (VA, VU). Variabilities in workload can occur in systems that are characterized by different phases of operation. As an example consider a client-server system, where different mean demands at a given device may occur during various time periods of a day. Such a variability may occur explicitly in a point-of-sale system where different amount of work

[*] This work is supported by the Austrian Science Foundation (FWF research grant No. S05303-MAT)

<table>
<tr><th>Condition</th><th>$\forall i: \ \underline{x}_i = \overline{x}_i$</th><th>$\exists \, i: \ \underline{x}_i < \overline{x}_i$</th></tr>
<tr><td>$m = 1$</td><td>

Conventional single value WLC model (SV)</td><td>

WLC model with uncertainties (UN)</td></tr>
<tr><td>$m > 1$</td><td>

WLC model with variabilities
without uncertainties (VA)</td><td>

WLC model with variabilities
and uncertainties (VU)</td></tr>
</table>

Table 1. Types of Models

per transaction occurs during different periods of the day. Variabilities in service demands can also occur implicitly in systems. For example, different service demands have been observed in a database system described in [2]: during periods of time when less memory was available for transaction processing larger number of I/O operations were observed. In such systems neither a conventional single class nor a multiclass queueing network is adequate for the computation of system performance. Using single mean values for the service demands often leads to inaccurate results for such systems. As an alternative, one might think of using a separate model for each time period. The disadvantage of this approach would be twofold. On the one hand, the effort for solving all these models can become significantly high. On the other hand, the solution of the separate models may not provide an overall picture of the performance behavior of the system.

Variation in day to day service demands may also arise in data processing centers. Consider the central data processing system in a bank that supports the computation demands of all the branches in the city. Although the service demands for any device in the system generated by the branches are similar, day to day variations in workload for some of the devices take place. Such a variation in service demand can be captured by a histogram characterizing the service time at each device.

When evaluating models where more than one input parameter is subject to variability (i.e. characterized by a histogram), all possible parameter combinations (variability combinations) and their probability of occurrence have to be determined [8]. Assuming, that there are K input parameters described by a histogram with m_k different intervals for input parameter D_k, the number of combinations is given by $I = \prod_{k=1}^{K} m_k$. These combinations are based on the assumption of stochastic independence of the variabilities of the service demands. In case of dependencies the list of variability combinations has to be specified directly.

3 Adaptation of analysis techniques

3.1 Models with uncertainties

To analyze models with uncertainties by means of existing analysis techniques, we need to extend the mathematical expressions used in the respective analysis techniques to interval arguments.

Interval arithmetic. A direct approach to use intervals as parameters is the use of *interval arithmetic* [12], [13]. This means that we define the usual mathematical operations for intervals. Consider two intervals $A = [\underline{a}, \overline{a}]$, and $B = [\underline{b}, \overline{b}]$. We define $A + B = [\underline{a} + \underline{b}, \overline{a} + \overline{b}]$, $A - B = [\underline{a} - \overline{b}, \overline{a} - \underline{b}]$, $A \cdot B = [\min(\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}), \max(\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b})]$, and if $0 \notin B$, $A/B = A \cdot [1/\overline{b}, 1/\underline{b}]$. Also other elementary functions and operations can be extended to interval arguments. Using this interval arithmetic, a mathematical expression consisting of multiple operations and functions can be evaluated using interval arguments.

Many properties, such as commutativity and associativity of the addition and multiplication operators, also hold for interval arithmetic. However, e.g. the distributive law holds only in the weaker form of *subdistributivity*: $A \cdot (B + C) \subseteq A \cdot B + A \cdot C$. This is only a special case of the so-called *dependency problem* [4]: If an interval argument appears more than once in an arithmetic expression, it is usually treated as a different variable in each occurrence. E.g. $X - X$ is treated as $X - Y$ with $X = [\underline{x}, \overline{x}]$ and $Y = [\underline{x}, \overline{x}]$. Thus, $X - X$ is evaluated to $\{x_1 - x_2 \mid x_1, x_2 \in X\} = [\underline{x} - \overline{x}, \overline{x} - \underline{x}]$, instead of $\{x - x \mid x \in X\} = [0, 0]$.

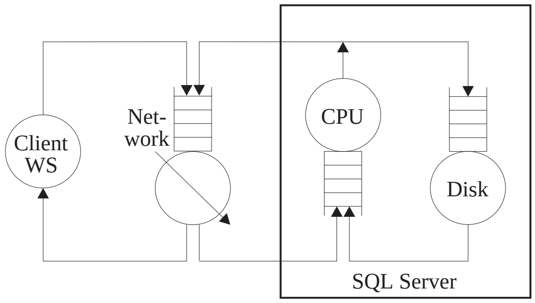


Fig. 1. Closed QNM modeling the computer system of a telemarketing company.

Monotonicity. Of course, monotone expressions can easily be evaluated with interval arguments by just evaluating the expression at the interval endpoints. Thus, if $f(x)$ is a monotone increasing function which can be evaluated as an interval expression, we have $f([\underline{x}, \overline{x}]) = [f(\underline{x}), f(\overline{x})]$. Analogous evaluation is possible for expressions with multiple arguments if the expression is monotone in every input parameter [6].

Global optimization. However, if monotonicity w.r.t. all input parameters can not be guaranteed, other techniques have to be used to obtain sharp interval results. The most general approach would be to employ global optimization algorithms to find the minimum and maximum of the respective expression within the range of input intervals. However, global optimization is usually of very high computational expense. On the other hand, these techniques can be more efficient if convexity or concavity of the analyzed expression can be guaranteed.

Interval splitting. Another technique to obtain sharper interval results is to split the input intervals into subintervals and use those for multiple evaluations of the expression (see e.g. [10] and [11]). This decreases the influence of the dependency problem discussed above. Interval splitting is also used to obtain more accurate results when using histograms to approximate distributions of input parameters (see [9]).

3.2 Models with variabilities

Variability models without uncertainties are described by a list of I parameter vectors $(d_{1,i}, \ldots, d_{K,i})$, $i = 1, \ldots, I$, of single valued service demands with corresponding probabilities of occurrence p_i. These models can be evaluated by simply applying a conventional solution algorithm to each of these parameter sets. This yields e.g. a total response time value $r_i(N)$, $i = 1, \ldots, I$, for every input parameter vector. The respective performance measure for the whole model can be obtained as the weighted sum of these intermediate results, using the probabilities of occurrence as weights: $r(N) = \sum_i p_i r_i(N)$.

For variability models with uncertainties the intermediate results are intervals $R_i(N) = [\underline{r}_i(N), \overline{r}_i(N)]$, $i = 1, \ldots, I$. Again, these intervals are combined to performance measure histograms by weighted summation. Since some of these intervals can be overlapping, the summation technique is required to handle such overlaps. To allow such a weighted summation of intervals to obtain a histogram, the *simplifying assumption* that the respective performance measures are uniformly distributed within each intermediate performance measure interval has to be made. Due to the nonlinearity of solution techniques, this assumption does not hold in general, but the error resulting from that assumption can be arbitrarily decreased by applying interval splitting.

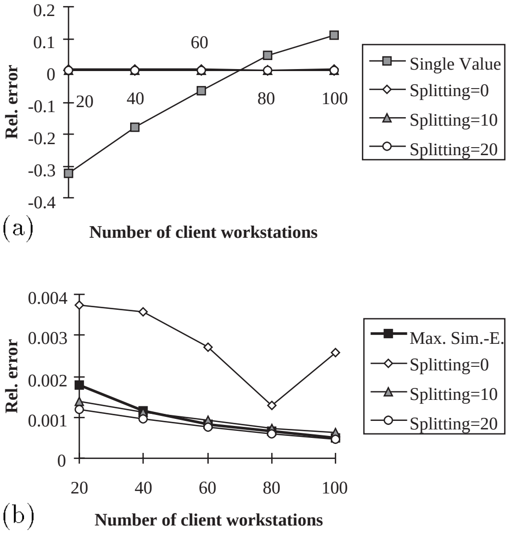


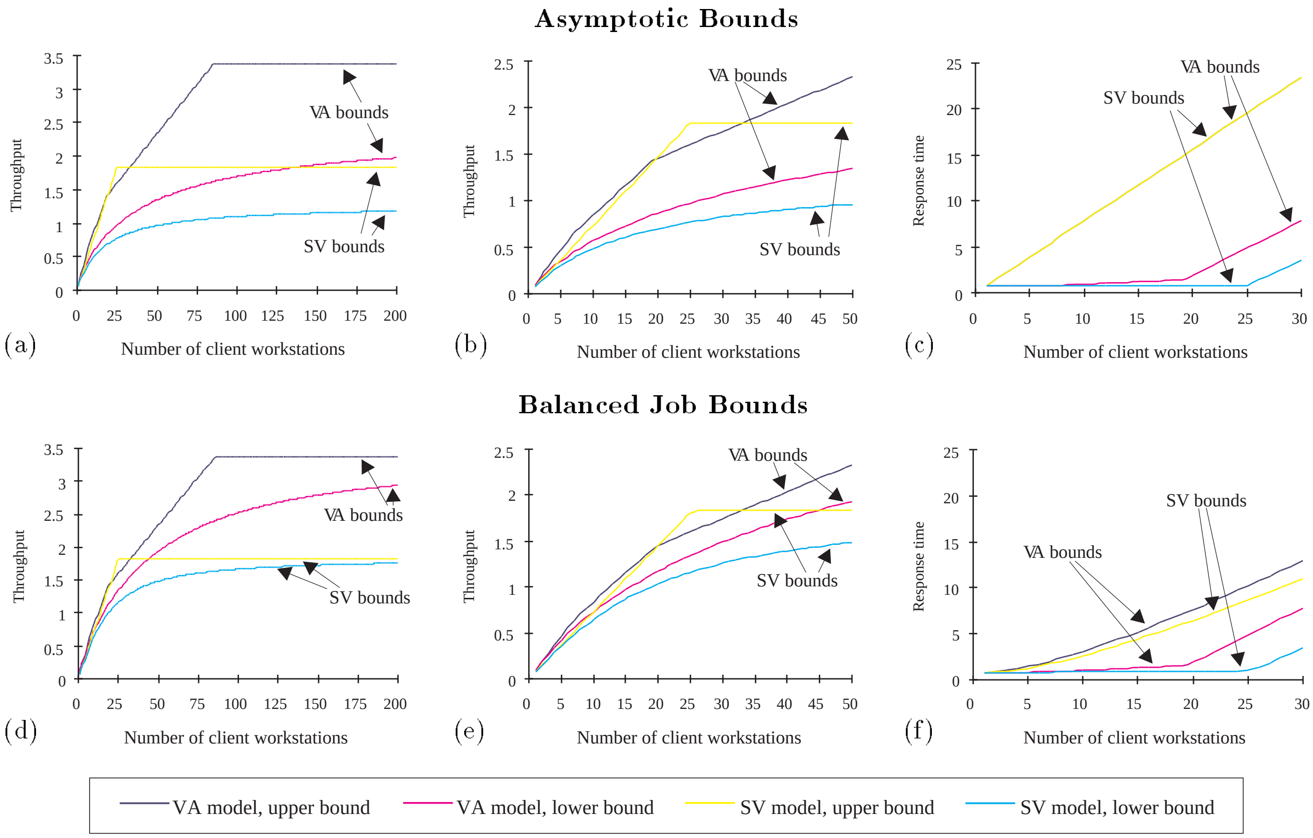
Fig. 2. Relative error resulting from ignoring inherent variabilities and uncertainties for an example QNM presented in [9].

4 Example: adaptation of analysis techniques for single class QNMs

As a demonstration case we have investigated the adaptation of several analysis techniques for single class queueing network models (QNMs):

4.1 Exact solution techniques.

For both, open and closed QNMs, the corresponding solution techniques (i.e. solution formulae for open models and the mean value analysis (MVA) algorithm for closed models – see [5] for example) are monotone w.r.t. all input parameters (see [1], [6], and [15] for details). Thus, the adaptation of these algorithms to interval arguments is possible by performing multiple evaluations of the corresponding conventional analysis techniques using combinations of input interval endpoints as arguments. Results of this investigation are presented in [8] and [9] . High inaccuracies are reported when evaluating models with variabilities and uncertainties using conventional SV analysis techniques instead of considering the variabilities in the WLC. For example, Figure 2a presents the relative error of throughput results for a QNM (depicted in Figure 1) modeling the client-server-based computer system of a telemarketing company, when using aggregated mean values as input parameters for a model with variabilities and uncertainties in workload. The modeled workload is assumed to be subject to


Asymptotic Bounds

Balanced Job Bounds

Fig. 3. Comparison of asymptotic (first row) and balanced job (second row) bounds for the throughput and response time of a QNM example presented in [7] (SV and VA models).

three phases each being modeled with parameter uncertainties (i.e. intervals). As it can be seen from this figure, the error from neglecting phase type behavior is up to 30% (depending on the number of client workstations modeled in the QNM). Figure 2b shows the effect of additional interval splitting for analysis with histogram-based input parameters. The curve with the highest error shows results without further splitting of the input intervals, whereas additional interval splitting decreases the error below the expected error of the simulation results used for comparison purposes ('Max.Sim.-E.'). Details of this example can be found in [9].

4.2 Bounds analysis.

The well known techniques of asymptotic and balanced job bounds (see [5] for example), often preferred as a first cut modeling tool for QNMs, have been adapted to handle WLC models with variabilities and uncertainties [7]. For variability models without uncertainties, the respective bound expressions are derived as weighted sum of the conventional asymptotic and balanced job bound expressions. For models with uncertainties, monotonicity of the conventional bound expressions is exploited to derive bound expressions depending only on interval endpoints of the corresponding input intervals. Given a QNM with N customers, K devices, device service demands d_k, and a terminal think time z, consider for example the well-known asymptotic bounds on throughput (see e.g. [5])

$$\frac{N}{N\sum_{k=1}^{K} d_k + z} \le \boldsymbol{x}(\boldsymbol{N}) \le \min\left[\frac{1}{\max_{k=1}^{K} d_k}, \frac{N}{\sum_{k=1}^{K} d_k + z}\right],$$

where $x(N)$ denotes the system throughput. If the workload is described by I variability combinations $(D_{i,1} = [\underline{d}_{i,1}, \overline{d}_{i,1}], \ldots, D_{i,K} = [\underline{d}_{i,K}, \overline{d}_{i,K}], Z_i = [\underline{z}_i, \overline{z}_i])$, $i = 1, \ldots, I$ with probability of occurrence p_i, the corresponding throughput bounds are

$$\sum_{i=1}^{I} \frac{p_i N}{N\sum_{k=1}^{K} \overline{d}_{i,k} + \overline{z}_i} \le \underline{\boldsymbol{x}}(\boldsymbol{N}) \le \overline{\boldsymbol{x}}(\boldsymbol{N}) \le \sum_{i} p_i \min\left[\frac{1}{\max_{k=1}^{K} \underline{d}_{i,k}}, \frac{N}{\sum_{k=1}^{K} \underline{d}_{i,k} + \underline{z}_i}\right].$$

Again, using aggregated single mean values as WLC is reported to produce highly inaccurate and incorrect bounds for performance measures of interest as compared to bounds analysis with explicit consideration of variabilities and uncertainties. E.g. Figure 3 depicts asymptotic (3a-c) as well as balanced job bounds (3d-f) for an example QNM modeled with a workload consisting of three phases. It can be seen that in regions with high load (i.e. high number of client workstations), the bounds produced from single value evaluation of the model do not even overlap with the bounds derived for the workload with consideration of variabilities and uncertainties. Derivation of the complete set of bound expressions and more detailed examples can be found in [7].

4.3 Bottleneck analysis.

Identification of system bottlenecks (BN) and modification analysis is another important technique when analyzing QNMs. Using WLC models with variabilities and uncertainties, the concept of unique primary, secondary, and so forth, bottlenecks is not sufficient. This is because different bottleneck devices at different variability combinations as well as possibly overlapping input parameter intervals hinder a unique bottleneck identification. In [6], bottleneck analysis for conventional WLC is extended to the concept of a *set of potential bottlenecks*. This concept is generalized to the notion of a *Bottleneck Probability Matrix* (BNPM) in [8]. It is shown that using conventional BN analysis for systems with variabilities and uncertainties may identify wrong devices as system bottlenecks. Using BNPMs, eventually several devices are identified as potential bottlenecks and corresponding probabilities are computed.

5 Conclusions and outlook

Conventional analysis techniques for performance modeling and prediction accept single mean value parameters as input and produce single mean performance measures as output. However, uncertainties in parameter values and variabilities in workloads can make these techniques ineffective. Exact values for all the parameters are often unknown at early stages of system design, but ranges of values that can be taken by these uncertain parameters may be available. Associating ranges or intervals with parameters of interest is appropriate in sensitivity analysis studies as well. Variabilities in workload may give rise to different mean service demands at different devices. For example, different mean device demands may be observed on a system during different periods of the day. Aggregating the workload and using a model characterized by a single mean demand for every device often leads to incorrect results. This research provides a general framework for the analysis of systems with variabilities and/or uncertainties in workload.

As a test case, existing techniques for mean value, bounds, and bottleneck analyses for single class queueing network models have been adapted to handle this type of workload characterization. These techniques are useful in a number of different situations that include the performance evaluation of conventional multiprogrammed systems and client-server systems characterized by variable workloads as well as software performance engineering in which uncertainties are often associated with parameter values in early phases of design. Work is continuing to adapt the methods of bound hierarchies [3] to workloads characterized by uncertainties and variabilities. Application of these techniques to real systems and understanding their effectiveness are important. Current work has focused primarily on queueing networks. Adaptation of other models of computer and communications systems to histogram-based analysis requires investigation.

References

[1] I. Adan and J. Van der Wal. Monotonicity of the Throughput of an Open Network in the Interarrival and Service Times. Memorandum COSOR 87-05, Eindhoven Univ. of Tech., Faculty of Mathematics and Computing Science, Eindhoven, the Netherlands, March 1987.

[2] J. P. Buzen. A Modeler's View of Workload Characterization. In G. Serazzi, ed., *Workload Characterization of Computer Systems and Computer Networks*, pp. 67–72. North-Holland, 1986.

[3] D. L. Eager and K. C. Sevcik. Bound Hierarchies for Multiple-Class Queueing Networks. *J. of the ACM*, 33(1):179–206, January 1986.

[4] E. R. Hansen. *Global Optimization Using Interval Analysis.* Marcel Dekker, Inc., New York, 1992.

[5] E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik. *Quantitative System Performance – Computer System Analysis Using Queueing Network Models.* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[6] J. Lüthi and G. Haring. Mean Value Analysis for Queueing Network Models with Intervals as Input Parameters. Technical Report TR-950101, Univ. Wien, ANIIS, Austria, July 1995.

[7] J. Lüthi, G. Kotsis, S. Majumdar, and G. Haring. Bounds-Based Performance Analysis for Distributed Systems with Variabilities and Uncertainties in Workload. In *Proc. Austrian-Hungarian Workshop on Distributed and Parallel Systems (DAPSYS'96), Miskolc, Hungary, October 2–4, 1996*, October 1996.

[8] J. Lüthi, G. Kotsis, S. Majumdar, and G. Haring. Performance Analysis using Queueing Network Models with Variabilities and Uncertainties in Workload. Technical Report TR-96102, Univ. Wien, ANIIS, Austria, June 1996.

[9] J. Lüthi, S. Majumdar, and G. Haring. Mean Value Analysis for Computer Systems with Variabilities in Workload. In *Proc. IEEE Int. Performance & Dependability Symposium (IPDS'96)*, Urbana-Champaign, IL, USA, September 1996.

[10] S. Majumdar, J. Lüthi, and G. Haring. Histogram-Based Performance Analysis for Computer Systems with Variabilities or Uncertainties in Workload. Research Report SCE-95-22, Carleton Univ., Ottawa, Canada, November 1995.

[11] S. Majumdar and R. Ramadoss. Interval-Based Performance Analysis of Computing Systems. In P. Dowd and E. Gelenbe, eds., *Proc. MASCOTS'95 (Durham, North Carolina, Jan. 18-20, 1995)*, pp. 345–351. IEEE Comp. Soc. Press, January 1995.

[12] R. E. Moore. *Methods and Applications of Interval Analysis.* SIAM, Philadelphia, 1979.

[13] A. Neumaier. *Interval methods for systems of equations.* Cambridge Univ. Press, Cambridge, 1990.

[14] C. U. Smith. *Performance Engineering of Software Systems.* Addison-Wesley, Reading, MA e.a., 1990.

[15] R. Suri. A Concept of Monotonicity and Its Characterization for Closed Queueing Networks. *Op. Res.*, 33:606–624, 1984.