Transactions. Georgian Technical University. AUTOMATED CONTROL SYSTEMS - No 1, 2006

L. Essafi, G. Bolch

Approximation of the Variance of Waiting Time in a Two-QueueTime Dependent Priority System

University of Erlangen-Nuremberg, Institute of Computer Science Martensstrasse 3,
D-91058 Erlangen, Germany lassaad@essafi.de, bolch@informatik.uni-erlangen.de

ABSTRACT

In this paper we derive an approximation of the variance of waiting time in a two-queue time dependent priority system. The derivation method is based on a transformation of the time dependent priorities system onto a static priorities system, with partial class switching. The derivation technique, proof, numerical and simulation results are presented and discussed.

KEYWORDS: Time Dependent Priorities, Variance.

1. INTRODUCTION

In a variety of application areas, different customer classes are defined, for which different grades of service are to be provided (e.g. different waiting times). To achieve this objective, scheduling strategies beyond the simple FIFO strategy are required. Known scheduling strategies include weighted fair queueing, priority queueing and weighted round robin.

One key performance measure considered in the analysis of queueing systems is the waiting time measure. In addition to the mean waiting time, it is in many application cases advantageous to have more insight in the variance or the standard deviation of the waiting time. In a call center for example, it is often not sufficient to only consider the mean waiting times (in the context referred to as ASA - average speed to answer). More insight is required on how far the waiting times actually deviate from an average.

In this paper we present a new method, how to derive the variance of waiting time in a two-queue system using time dependent priorities.

This paper is structured as follows: we first give a brief overview of time dependent riorities. We then present the derivation technique, show numerical examples, simulation results, and conclude.

2. TIME DEPENDENT PRIORITIES

The priority queueing disciplines can be generally classified into static and time dependent priorities. In a static priority system, the priority of a customer is constant during its whole sojourn time in the system. In many cases it is advantageous for a customer priority to increase with time. Such systems are more flexible but need more expense for the administration.

We assume in the following text a queuing model with R classes of customers, where arriving customers belong to a priority class r (r = 1, 2, . . . ,R). The interarrival and services times in all classes are assumed as to be exponential.

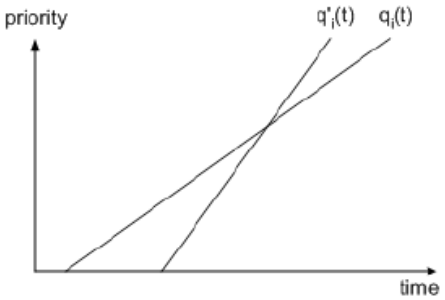


Figure 1: Priority functions with slopes b and b0

Each priority class is assigned a parameter br, which can be interpreted according to the priority function

$$q_r(t) = (t - t_0)b_r \qquad (1)$$

as the increasing rate (slope) of the priority in the class r. A customer enters the system at time t0 and then increases its priority at the rate br (see Figure 1). The priority of a higher class customer increases faster that the priority of a lower class customer, $0 \ b_1 \le b_2 \le \ldots \le b_R$.

Variants of this priority function, where exponents are assigned to the time component and/or slope components have also been introduced in the literature, e.g. in [5, 1].

3. DERIVATION OF THE VARIANCE OF WAITING TIME

3.1. Review of Related Research

The knowledge of the waiting time distribution or higher (central) moments of the waiting time function enables the system designer to perform more appropriate analysis of the queueing system. The waiting time distribution has been extensively studied for single-class systems, however few results are available for multi-class systems [4]. In [4] an approximation formula is given for the waiting time distribution under several queueing policies, including static priorities and weighted fair queueing.

Several works study the behavior of waiting time in static priorities systems with multiple classes. Laplace transforms of the waiting time are provided e.g. in [5, Eqn. 3.32]. This equation can be differentiated and evaluated numerically to obtain higher moments of the mean waiting time. [8] addresses the waiting time distribution functions for a more general class of static priorities using preemption and preemption distances.

The analytical evaluation of the delay distributions in transform domains (e.g. Laplace transforms) usually requires complex mathematics and numerical approximations, especially for inversion. Furthermore, the models used for system analysis are often approximate models and exact distribution functions of the arrival or service processes are not known. In such cases, a possible approach is the use a two parameter description (mean and squared coefficient of variation) of the arrival and service processes.

The two-moment approximation was applied in the context of static priorities in [3] to derive the two first moments of waiting time. In [6], an approximation of the waiting time variance in an 2-class M/M/1 static priority system is given.

Other analysis approaches include modelling of the static priority system as a polling system, e.g. in [7] or using simulation.

The aim of our paper is to address the variance of waiting time in two-class time dependent priorities systems, which to the best of our knowledge has not been addressed in previous works.

3.2 Problem Statement

We consider an M/M/1 queueing system with two classes, where the strategy applied in time dependent priorities, as depicted by Figure 2.

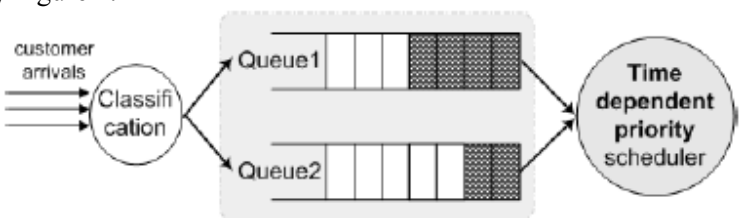


Figure 2: Initial Problem

The inter-arrival times in both classes are exponential, as described by the rates λ_1 and λ_2 respectively. The service times in both classes are also exponential, and described by the rates μ_1 and μ_2 respectively. ρ_i is used to denote the utilization in class i, and ρ for the total system utilization. Furthermore, we assume that the system is in stable condition, i.e. the total system utilization denoted by ρ is less than 1 ($\rho = \lambda_1/\mu_1 + \lambda_2/\mu 2$).

We use the notation W_i^{TDP} (λ_1, λ_2, μ_1, μ_2, b_1/b_2) or shortly Wi to refer to the waiting time of a customer of class i (i = 1 or i = 2), in a two-class time dependent priority system, characterized by the arrival rates λ_1 and λ_2, the service rates μ_1 and μ_2, and ratio of priority slopes b_1/b_2. E[Wi] and VAR[Wi] are used to refer the mean and variance of waiting time in class i.

In a time dependent priority system, the mean waiting times in class 1 and 2 are given by (see for example [5, 2, 1]):

$$E[W_1^{TDP}] = \frac{E[W_0]}{\Big(1-\rho\Big)\Big(1-\rho_2^{TDP}(1-\frac{b_1}{b_2})\Big)} \quad (2)$$

and

$$E[W_2^{TDP}] = \frac{E[W_0]}{(1-\rho)}\Big(1-\frac{\rho_1^{TDP}(1-\frac{b_1}{b_2})}{1-\rho_2^{TDP}(1-\frac{b_1}{b_2})}\Big), \quad (3)$$

where E[W0] represents the mean remaining service time [2]. Our objective is to derive an approximation of the the variance of waiting time in class 1 and 2, denoted by VAR[W_1] and VAR[W_2].

3.3. Problem Transformation

The idea behind our approximative approach is to transform the initial problem using time dependent priorities, depicted by Figure 2, into a problem using static priorities, depicted by Figure 3, where results for the variance of waiting time have already been derived (see section 3.1).

The transformation is based on the idea that for customers waiting for service relatively long in class 1(lower priority), a partial class switching to class 2 is performed, in order to enforce that they get served (also prior to other arriving class 2 customers).

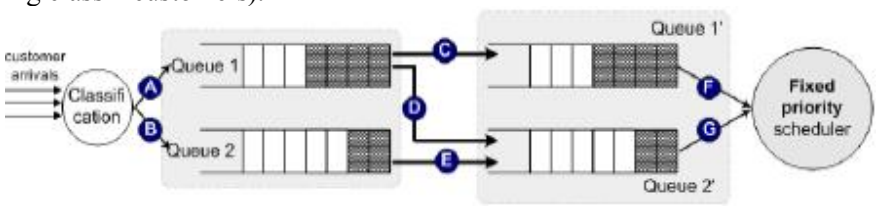


Figure 3: Transformed Problem

If we consider the long term (stationary) behavior of the time dependent priority system, we can generalize by saying, that a portion of class-1 customers, is ”promoted” to class-2. This portion of the traffic, denoted by α, where $0\% \leq \alpha \leq 100\%$, is the class-1 traffic, which is served, even when class-2 customers are in the queue and waiting for service.

It can be expected that for the static priorities case (can be modelled by time dependent priorities by setting $b_1 = 0$ and $b_2 = 1$) that α must be equal to 0%, in order to keep both classes separate and keep the strict differentiation. In the first-come-first-served case and no differentiation (can be modelled by time dependent priorities by setting $b_1 = b_2 = 1$), that α must be equal to 100%, i.e. arrivals are merged in one stream and processed in FIFO order.

Considering the observation above and referring to Figure 2 and Figure 3, we can rephrase the idea behind the transformation as follows:

- Class-2 customers are generally served before class-1 customers, in the static priority scheduling sense. This is depicted in Figure 3 as B-traffic becoming E-traffic.
- A part of class-1 customers have to be served before other class-2 customers, due to their relatively long waiting time. This is depicted by Dtraffic in Figure 3 and is the portion α of the original A-traffic, which undergoes the class switching.
- The rest of the A-traffic (i.e. excluding the D-part), is served last, in the static priority scheduling sense. This is denoted as C-traffic in Figure 3.

The transformed queueing system using static priorities is characterized by:

- customer arrival rate $(1-\alpha)\cdot\lambda_1$ in Queue 1’ being served at service rate μ_1,
- customer arrival rate $(\alpha\cdot\lambda_1)+\lambda_2$ in Queue 2’ being served at service rate μ_1 for D-traffic and μ_2 for E-traffic.

In this transformed queueing system, the mean and variance in Queue 1’ and Queue 2’ can be calculated and used to determine the mean and variance in the original queueing system, using time dependent priorities.

3.4. Determination of α

One central element is to determine the part of the A-traffic, which is ”promoted” and served as a highpriority traffic (i.e. class-2 traffic). In our notation this is referred to as D-traffic and represented as a portion α of the original A-traffic. To determine α, we use two constraints based on the mean waiting times, which are

known for both the original system (using time dependent priorities) and the transformed system (using static priorities).

The constraints can be formulated as:

- The mean waiting time of class-1 customers in the time dependent priority system is the weighted average of the mean waiting times of class 1 and 2 in the static priority system. α "percent" of the traffic has mean waiting time of Queue 2' and (1−α) "percent" has mean waiting time of Queue 1'.
- The mean waiting time of class-1 customers in the time dependent priority system is the same as the mean waiting time of customers in Queue 2' in the static priority system.

The two constraints can be expressed as:

$$E[W_1^{TDP}(\lambda_1, \lambda_2, \mu_1, \mu_2, b_1/b_2)] = (1-\alpha)\ E[W_1^{SP}((1-\alpha)\lambda_1, \alpha\lambda_1 + \lambda_2, \mu_1, \mu_2)] + \alpha\ E[W_2^{SP}((1-\alpha)\lambda_1, \alpha\lambda_1 + \lambda_2, \mu_1, \mu_2)] \quad (4)$$

and

$$E[W_2^{TDP}(\lambda_1, \lambda_2, \mu_1, \mu_2, b_1/b_2)] = E[W_2^{SP}((1-\alpha)\lambda_1, \alpha\lambda_1 + \lambda_2, \mu_1, \mu_2)] \quad (5)$$

whereby the following notation is used:

W_i^{TDP} (λ_1, λ_2, μ_1, μ_2, b_1/b_2) denotes the waiting time of a customer of class i, in a two-class time dependent priority system, with class 1 characterized by arrival rate _1 and service rate μ1, class 2 characterized by arrival rate _2 and service rate μ2 and ratio of priority slopes b1/b2.

W_i^{SP} (λ_1, λ_2, μ_1, μ_2) denotes the waiting time of a customer of class i, in a two-class static priority system, with class 1 characterized by arrival rate λ_1 and service rate μ_1, class 2 characterized by arrival rate λ_2 and service rate μ_2.

It can be shown using several mathematical transformations that the portion of the traffic α is given by (refer also to Appendix 1 for the outline of the derivation):

$$\alpha = \frac{\frac{b_1}{b_2}}{1 - \rho(1 - \frac{b_1}{b_2})} \quad (6)$$

Figure 3.4 shows α as function of b_2 and the overall system utilization, where b_1 is equal to 1.

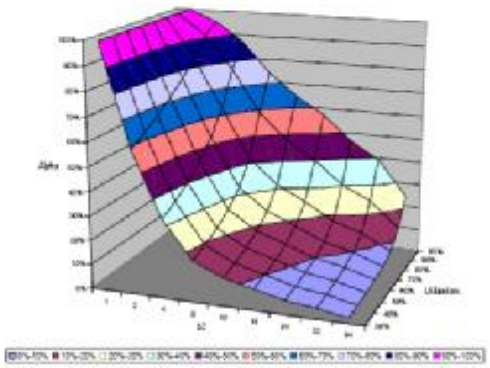


Figure 4: _ as function of utilization and class 2 priority

We consider two limit cases of the formula above.

FIFO Case: To model a FIFO system using time dependent priorities, b_1/b_2 has to be set to 1. In this case α is equal to 1. This means that all class 1 traffic is pushed to the same queue like class 2. In other words, there is no difference between class 1 and class 2 anymore in the equivalent static priority system, i.e. a first-come-first-served mode is applied.

Strict Priority Case: To model a static priority system using time dependent priorities, b_1/b_2 has to be set to 0, in order for the priority of class 1 customers not to increase with time. In this case, α is equal to 0. Which means that no class 1 traffic is "promoted" to class 2 and a static priority order is maintained.

3.5. Waiting Time Variance in Time Dependent Priorities Systems

The variance of waiting time of customers in queue 1 of the time dependent priority system is composed of two parts: first the customers served in queue 1' of the static priority system (C-traffic) and second the customers served in queue 2' (D-traffic). The portions of both parts are $(1 - \alpha)$ and α respectively. For the customers in queue 2 of the time dependent priority system, the variance is equal to the variance of queue 2' in the static priority system.

In the following we use the following two properties of the variance of two random variables X and Y:

$$VAR[aX] = a^2VAR[X], \quad (7)$$

for a constant **a**, and if X and Y are stochastically independent,

$$VAR[X + Y] = VAR[X] + VAR[Y]. \quad (8)$$

For the variance in static priority systems, we discussed in Section 3.1 several methods how it can be derived. In [6], an approximation of the waiting time variance in an M/M/1 2-class static priority system, with arrival rates _1 and _2 respectively and service rate μ in both classes, is erived as:

$$VAR[W_1] \approx \frac{2\lambda\mu^2 - \lambda_2\lambda^2 - \lambda^2\mu}{(\mu - \lambda)^2(\mu - \lambda_2)^3} \quad (9)$$

and

$$VAR[W_2] \approx \frac{2\lambda\mu - \lambda^2}{\mu^2(\mu - \lambda_2)^2} \quad (10)$$

Using the approximations and the variance properties above, he variance of waiting time in class 1 and class 2 in the time dependent priority system can be approximated by:

$$\mathrm{VAR}[W_1] \approx (1-\alpha)^2 \frac{2\lambda\mu^2 - \alpha\lambda_1\lambda^2 - \lambda_2\lambda^2 - \lambda^2\mu}{(\mu-\lambda)^2(\mu-\alpha\lambda_1-\lambda_2)^3} + \alpha^2 \frac{2\lambda\mu-\lambda^2}{\mu^2(\mu-\alpha\lambda_1-\lambda_2)^2} \quad (11)$$

and

$$\mathrm{VAR}[W_2] \approx \frac{2\lambda\mu - \lambda^2}{\mu^2(\mu - \alpha\lambda_1 - \lambda_2)^2} \quad (12)$$

In Table 1, we present three examples, where the mean waiting times in class 1 and class 2 and the corresponding variances in a time dependent priority system are calculated.

3.6 Validation of the Derived Results

We conducted an extensive simulation study with 944 runs to study and validate the accuracy of the derived approximation. Our analysis revealed a good accuracy of the approximation, mostly within 20% deviation, considering the absolute and relative errors. For illustration purposes, we show sample results for the case where $b_1 = 1$ and $b_2 = 4$ at two different load distributions (LD) between class 1 and class 2 in Figures 5 and 6.

Table 1: Waiting time variance using time dependent priorities

ID	Measure	Ex. 1	Ex. 2	Ex. 3
1	λ_1	1	1	1
2	λ_2	3	3	3
3	μ	6.67	5.71	4.44
4	b_2/b_1	8	16	2
5	α	20%	18%	91%
6	$E[W_1]$	0.371	0.804	3.057
7	$E[W_2]$	0.176	0.276	1.681
8	$VAR[W_1]$	0.386	1.672	9.500
9	$VAR[W_2]$	0.073	0.142	3.454

Figure 5: class1 25%, class2 75%

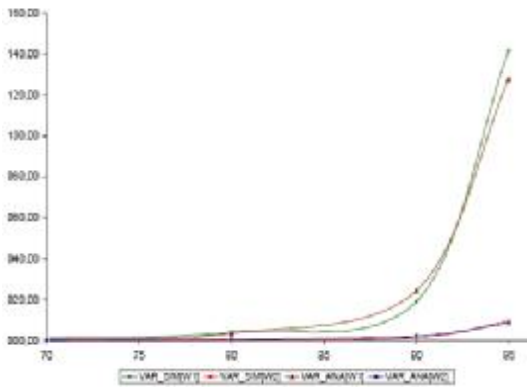
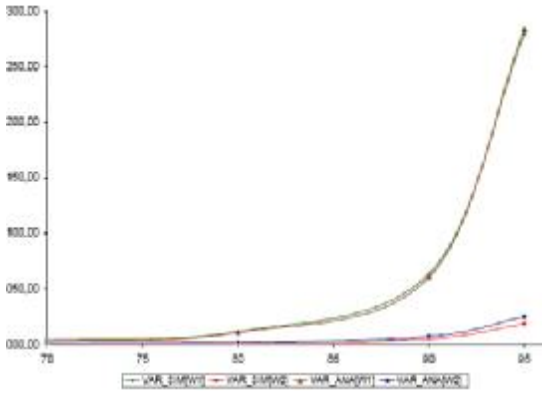


Figure 6: class1 75%, class2 25%



References

[1] Gunter Bolch and Werner Bruchner. Analytische Modelle symmetrischer Mehrprozessoranalagen mit dynamischen Priorit¨aten. Technical report, Universit¨at Erlangen-N¨urnberg, 1982.

[2] Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi. Queueing Networks and Markov Chains : Modeling and Performance Evaluation with Computer Science Applications. John Wiley & Sons, Inc., 1998.

[3] Gabor Horvath. Stochastic Models in Telecommunication Systems. PhD Thesis, Budapest University of Technology and Economics, 2004.

[4] Yuming Jiang, Chen-Khong Tham, and Chi-Chung Ko. An approximation for waiting time tail probabilities in multiclass systems. IEEE Communication Letters, 5(4):175–177, April 2001.

[5] Leonard Kleinrock. Queueing Systems – Volume II. John Wiley & Sons, Inc., 1976.

[6] M. Mittler and N. Gerlich. Reducing the Variance of Sojourn Times in Queueing Networks with Overtaking. Technical Report 73, Universit¨at W¨urzburg, November 1993.

[7] Toshihisa Ozawa. Waiting Time Distribution in a two-queue model with mixed exhaustive and gated type K-limited services. In Proceedings of International Conference on Performance and Management of Complex Communication Networks, pages 231–250, 1997.

[8] Martin Paterok. Warteschlangensysteme mit R¨uckkopplung und Priorit¨aten. PhD Thesis, Universit ¨at Erlangen-N¨urnberg, 1990.

Appendix 1 - Outline of the Proof of Equation 6

For the proof of the result in Equation (6), the following basic results are used: For a static priority system which two priority classes, the closed formula for the mean waiting times are given by:

$$E[W_1^{SP}] = \frac{E[W_0]}{(1-\rho)(1-\rho_2^{SP})} \qquad (13)$$

and

$$E[W_2^{SP}] = \frac{E[W_0]}{(1-\rho_2^{SP})}; \qquad (14)$$

and for a time dependent priority system by equations (2) and (3). The utilizations quantities in the original queueing system (i.e. using time dependent priorities) are given by

$$\rho_1^{TDP} = \frac{\lambda_1}{\mu_1} \qquad (15)$$

and

$$\rho_2^{TDP} = \frac{\lambda_2}{\mu_2}. \qquad (16)$$

For the transformed queueing system the following utilizations are applicable:

$$\rho_1^{SP} = \frac{(1-\alpha)\lambda_1}{\mu_1} = (1-\alpha)\frac{\lambda_1}{\mu_1} = (1-\alpha)\rho_1^{TDP} \qquad (17)$$

and

$$\rho_2^{SP} = \frac{\alpha\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} = \alpha\rho_1^{TDP} + \rho_2^{TDP}. \qquad (18)$$

Reference is made here also to Equations (3.1) to (3.4) in [5]. We use the constraint equation given in 4 and substitute with Equations (13), (14) and (2) to get:

$$\frac{E[W_0]}{1-\rho}\left(1-\rho_2(1-\frac{b_1}{b_2})\right)^{-1} =$$

$$(1-\alpha)\ \frac{E[W_0]}{1-\rho}(1-\alpha\rho_1-\rho_2)^{-1}$$

$$+\ \alpha\ E[W_0](1-\alpha\rho_1-\rho_2)^{-1}$$

This expression can further be simplified and resolved get _ as stated in equation 6:

$$\alpha = \frac{\frac{b_1}{b_2}}{1-\rho(1-\frac{b_1}{b_2})}$$

It can be shown that for the value of _ derived above, that the two terms given in the second constraint function in Equation (5) are equal.

ლ. ესაფი, გ. ბოლხი

ლოდინის დროის ვარიაციის მიახლოებითი მნიშვნელობის დადგენა ორ-რიგიან დროზე დამოკიდებულ პრიორიტეტის სისტემაში

რეზიუმე

სტატიაში ვადგენთ (ვასკვნით) ორ-რიგიან დროზე დამოკიდებულ პრიორიტეტულ სისტემებში ლოდინის დროის მიახლოებით ცვალებადობას. დასკვნის დადგენის მეთოდი დაფუძნებულია დროზე დამოკიდებული პრიორიტეტული სისტემის გარდაქმნაში სტატიკურ პრიორიტეტულ სისტემად კლასის ნაწილობრივი გადართვით. დასკვნის ტექნიკა, ცდები, რიცხვითი და იმიტაციური შედეგები წარმოდგენილია ნაშრომში და იგი სადისკუსიო საკითხია.

Л. Ессафи, Г. Болх

УСТАНОВЛЕНИЕ ПРИБЛИЖЕННОГО ЗНАЧЕНИЯ ВАРИЯЦИИ ВРЕМЕНИ ОЖИДАНИЯ В ДВУХ-ОЧЕРЕДНОЙ ЗАВИСИМОЙ ОТ ВРЕМЕНИ СИСТЕМЕ ПРИОРИТЕТОВ

Резюме

В работе устанавливаем значение приближеннного изменения времени ожидания в двух-очередной зависящей от времени приоритетных системах. Метод устанавления утверждения основан на преобразование зависящей от времени приоритетной системы в статическую систему с частичным переключением класса. Техника утверждения, опыты, числовые и имитационные значения представлены в работе и являются вопросами дискуссии.