SCHEDULING TO IMPROVE QUEUE JUSTICE

Werner Sandmann
Department of Information Systems and Applied Computer Science
University of Bamberg
Feldkirchenstr. 21
D-96045, Bamberg, Germany
E-mail: werner.sandmann@wiai.uni-bamberg.de

KEYWORDS
Queueing, Justice, Fairness, Scheduling, Discrimination Frequency, Simulation Study

ABSTRACT

Scheduling should manage queues in a satisfactory way. Many technical applications as well as common daily life queueing situations involve humans, meaning that psychological effects and justice is of major importance, where individual perceptions of justice are strongly coupled with a fair and equal treatment of users or customers. Personal impressions of justice are often more important than classical queueing performance measures. Hence, quantifying justice is particularly well suited for evaluating queueing systems and scheduling policies with regard to human attitudes.

We consider the discrimination frequency as a basis for quantifying justice, where being discriminated means to be overtaken or to wait for customers with large service requirements. For a queue to be just, an equal treatment of customers is necessary, i.e. the amount of discriminations should not excessively vary for different customers as it is often the case in commonly used scheduling policies. A new policy, MFD (Most Frequently Discriminated), is introduced and shown to be useful. We provide a comparative simulation study for queueing systems operating under the traditional FCFS and SJF policies and MFD. Our results indicate that MFD significantly improves queue justice without too much worsening mean response times. It even reduces the variance of the response time.

INTRODUCTION

Queueing models are widely used in application areas such as computer and communication systems, operations research, manufacturing, or business processes, amongst others. Especially, when some uncertainty in system behaviour is involved, stochastic models are adequate and have shown to be useful. Moreover, queueing is present almost every day in a variety of different real life situations, in particular situations involving humans such as queueing situations at banks, supermarkets, airport counters and many more. Queues are used to organize the sequence of services offered to customers, and scheduling policies are necessary to manage queues. Thus, queueing and scheduling are intimately related, and many books covering both topics are available, e.g. (Conway et al 1967) and (Kleinrock 1975) to mention only two classical ones. Hence, performance measures should be evaluated for different scheduling policies, and scheduling policies should be compared with regard to their performance.

Queueing and scheduling theory both have extensively dealt with performance measures like response and waiting times, utilization, throughput or other related properties. Probabilities of large backlogs, buffer overflows, customer blocking or losses have been considered, too. All these performance measures have in common that they are clearly defined in the sense that they do not contain subjective components. In contrast to this, justice is highly subjective and in particular quantitative justice has been investigated much less than other quantitative measures, but it received growing attention only very recently, most often in terms of quantitative measures of user-perceived fairness, which is strongly related to and in some terminologies even equivalent to quantifying justice.

QUEUE JUSTICE

In an early work (Larson 1987) introduced the term 'social injustice' meaning violations of the FCFS service discipline in queues. In addition to FCFS violations other factors deteriorate individual perceptions of queueing. According to (Larson 1987) and common intuition a customer would have a better experience entering a queue behind several other customers each of whom requires a relatively small time of service, rather than behind one single customer who requires a large amount of service time. Thus, at least two factors other than the usually taken performance measures as for example response time influence customers' individual justice perceptions: service order with regard to the order of arrivals and service re-

Proceedings 20th European Conference on Modelling and Simulation
Wolfgang Borutzky, Alessandra Orsoni, Richard Zobel © ECMS, 2006
ISBN 0-9553018-0-7 / ISBN 0-9553018-1-5 (CD)

quirements of other customers. Hence, both factors should be the building blocks of quantitative justice.

Among recent work on the topic, where the term 'justice' has mostly disappeared and more or less equivalently substituted by the term 'fairness', slowdown (Wierman and Harchol-Balter 2003), order fairness (Avi-Itzhak and Levy 2004) and the resource allocation queueing fairness measure (Raz et al 2004) are mainly motivated and dedicated to computer and communication systems applications though the latter also accounts for interactive systems where human users are present. More general psychological studies of human attitudes, queues and fairness have been provided by (Rafaeli et al 2002) and (Rafaeli et al 2005).

In (Sandmann 2005) discrimination frequency fairness has been introduced, which basically counts two types of discriminations called 'overtaking' and 'large jobs', where 'overtaking' means FCFS violations and 'large jobs' means waiting for customers who have larger service requirements than oneself has. Hence, the measure is consistent with (Larson 1987) and moreover, it is corroborated by outcomes of the psychological studies that have been presented in (Rafaeli et al 2002) and (Rafaeli et al 2005). Axiomatic properties of discrimination frequency fairness are proven in (Sandmann 2005), and analytical expressions for the expected discrimination frequency in M/GI/1 queues operating under FCFS (First Come First Served), LCFS (Last Come First Served) and SJF (Shortest Job First) scheduling as well as some simulation results for properties related to the second moment of the discrimination frequency are determined in (Sandmann 2006).

Considering variance and standard deviation is necessary and most important to account for an equal treatment of customers. Small values of variances, standard deviations and related properties are necessary to provide some predictability on the system behaviour. If one expects some amount of discrimination and one experiences approximately the expected amount, one will not be too angry about that, but if one is unlucky enough to receive an unexpected large amount of discrimination, a deep impression of injustice is established. The range of discrimination frequencies is indicated by maximum values. Note that we distinguish the terms fairness and justice such that fairness only accounts for expectations whereas justice highlights an equal treatment of customers measured by properties related to second moments and maximum values.

All the mentioned fairness measures, including discrimination frequency fairness, have been studied for systems operating under several common already existing scheduling policies yielding comparisons of these policies. But scheduling policies are designed to reach specific goals, typically to optimize some target performance measure. For example, SJF is the non-preemptive policy that minimizes the expected response time, but on the negative side it yields a large increase of response time variance compared to many other policies. Until today, no attempt has been made to design a new scheduling policy that accounts for justice according to the chosen fairness measure as the main target, and in the present paper we remedy this lack for the discrimination frequency by considering our MFD (Most Frequently Discriminated) policy.

DISCRIMINATION FREQUENCY

Consider a single server queue, where the successive customer arrival times are denoted by $a_1, a_2, \ldots$, the corresponding service times by $s_1, s_2, \ldots$, and the (except for FCFS typically not successive) corresponding departure times by $d_1, d_2, \ldots$ That is, the i-th arriving customer arrives at time a_i, departs at time d_i and has service time s_i. In stochastic models all these times are random variables according to certain probability distributions and they are usually denoted by upper case letters.

The amount n_i of overtaking the i-th customer suffers from is the number of customers who arrived not earlier and complete service not later. Formally,

$$n_i := |\{j : a_j \geq a_i \wedge d_j \leq d_i\}|.$$

The amount m_i of large jobs the i-th customer suffers from is the number of customers present in the system upon his arrival who have at least as much remaining service time and complete service not later than himself. Formally,

$$m_i := |\{j : d_i \geq d_j > a_i \wedge s'_j(a_i) \geq s_i\}|,$$

where $s'_j(t)$ denotes the remaining service time of the j-th customer at time t. Note, that if a customer is overtaken by another customer with at least the same service requirements, this affects both quantities defined above, i.e. such cases are taken as doubly unfair. This is consistent with personal feelings of human customers.

The discrimination frequency of a customer is the number of discriminations he suffers from, that is the sum of the amount of overtaking and the amount of large jobs, i.e.

$$D(i) := n_i + m_i.$$

For queueing systems with random interarrival and service times according to some probability distributions, in steady state D denotes the discrimination frequency random variable. Expectations and higher moments are used to quantify fairness and justice, or alternatively unfairness and injustice, of the system. Further motivations and discussions, in particular on the suitability of discrimination frequency based fairness can be found in (Sandmann 2005) and (Sandmann 2006).

MOST FREQUENTLY DISCRIMINATED

We first consider the probably most common non-preemptive scheduling policies, FCFS (First Come First Served) and SJF (Shortest Job First) that show contradictory extremal behaviours. Then we compare our MFD policy with FCFS and SJF. We focus on non-preemptive scheduling since we are mainly considering common daily life environments where humans are involved and where service preemptions are typically not present. Obviously, FCFS minimizes the amount of overtaking, since under FCFS there is no overtaking at all. Similarly, SJF minimizes the number of large jobs, since the only situation in which such a wait occurs is that a customer currently in service has larger remaining service time than an arriving customer, implying exactly one large job.

Unfortunately, both policies do not optimize the discrimination frequency. As shown in (Sandmann 2006) there is no clear ranking which policy performs better in terms of the expected discrimination frequency, and the picture is even more diffuse for justice in terms of the second moment or maximum values. The expected discrimination frequency and several higher moment properties depend on the specific queue structure and design, which means for example they depend on distributions of interarrival and service times and in particular on higher moments of these distributions. In some cases the expected discrimination frequency for FCFS is less than for SJF, in other cases it is just the opposite. There are many cases where SJF yields smaller expectations and much higher variances than FCFS, but there are also some (fewer) cases where the opposite holds.

On what basis should we design a just scheduling policy? FCFS schedules based on arrival times, and SJF schedules based on service times. When the discrimination frequency and its second moment related properties are the metrics of main interest, it seems naturally that one should schedule based on discrimination frequencies. Since the goal is an equal treatment of customers, there should be no customers discriminated excessively often. From this reasoning the idea arises that a policy similar to LRU (Least Recently Used) or LFU (Least Frequently Used) which are well known in operating systems theory, should be applied. Hence, transfered to our objective we create MFD (Most Frequently Discriminated), i.e. the customer who has already experienced the most discriminations is served next.

In addition to the scheduling basis, there is at least one significant difference between MFD on the one hand and FCFS and SJF on the other hand concerning the type of policy. Although we still consider a non-preemptive policy, scheduling is now dynamic in the sense that it depends on the service history, and the information a customer initially brings to decide when to serve him changes with his experiences in the queue. The treatment of a customer during his stay in the queue affects the scheduling policy. Since the customer with the largest discrimination frequency is chosen to be served next, MFD should significantly reduce variance related properties of the discrimination frequency and maximum discrimination frequency and thus increase queue justice, as we will indeed demonstrate in the next section.

SIMULATION STUDY

Analytical expressions for the expected discrimination frequency in M/GI/1 queues, including the special case of M/M/1 queues, are given in (Sandmann 2006), but no such expressions are available yet for higher moments of the discrimination frequency as required for evaluating justice. Hence, we performed a simulation study. Here, we provide a comparative simulation study and present representative results for M/M/1, M/Erlang/1, Erlang/M/1, and M/Pareto/1 queueing systems with different values of the server utilization ρ. To be precise we have to describe the involved probability distributions and in particular their parameters.

Models and Parameters

In M/M/1 queueing systems interarrival and service times are exponentially distributed with means λ^{-1} and μ^{-1}, respectively, where the density of an exponential distribution with mean a^{-1} is given by

$$f(x) = a\exp(-ax), \quad x > 0,$$

and the variance equals a^{-2}.

In M/Erlang/1 queueing systems interarrival times are exponentially distributed with mean λ^{-1} and service times are distributed according to an Erlang distribution with k phases and mean μ^{-1}. Similarly, in Erlang/M/1 interarrival times are Erlang distributed with k phases and mean λ^{-1}, and service times are exponentially distributed with mean μ^{-1}. The density of an Erlang distribution with k phases and mean a^{-1} is given by

$$f(x) = \frac{ak(akx)^{k-1}}{(k-1)!}\exp(-akx), \quad x > 0,$$

and the variance equals $1/(ka^2)$. In our study we have chosen the Erlang distribution with 10 phases.

Finally, the density of a Pareto distribution (as present in M/Pareto/1 systems) with parameter a

is given by

$$f(x) = \frac{a}{x^{a+1}}, \quad a > 0, x > 1.$$

The Pareto distribution is heavy-tailed, its expectation $a/(a-1)$ only exists for $a > 1$ and its variance, given by $a/((a-1)(a-2)^2)$, only exists for $a > 2$. In our study we have chosen $a = 2$, which means that the expectation equals 2, whereas the variance does not exist.

As usual, $\rho = E[S]/E[T]$ denotes the server utilization, where S and T are random variables distributed according to the service time distribution and the interarrival time distribution, respectively.

Metrics

As discriminations occur mainly in systems with high utilization and discrimination frequencies do not differ very much for low utilizations we focus on models where the server is on average at least half the time busy, and each model has thus been simulated for $\rho \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. The new policy MFD is compared to FCFS and SJF in terms of the discrimination frequencies' variance $\sigma^2(D)$, the coefficient of variation $c(D) = \sigma(D)/E[D]$, and the maximum observed value of discriminations to a customer $D_{\max}$. Moreover, the impact on response times is studied, in particular the prize that has to be paid in the sense of increased response time compared to SJF and the effects of MFD on the variance of the response time.

Methodology

We applied the classical replication/deletion approach as described for example in (Law and Kelton 2000), i.e independent simulation runs with a sufficiently large warm-up period to determine point estimates and confidence intervals. More specifically, we performed independent runs, where the observation period for each run was of length (=number of served jobs) 10^7 to form 99% confidence intervals with a relative half width less than 1% for M/M/1, M/Erlang/1 and Erlang/M/1. The simulation of M/Pareto/1 queues, meaning a heavy-tailed service time distribution and - as in our particular case - even non-existing variance of the service time distribution is much more demanding. Here, we set the condidence level to 95% and the maximum relative half width to 10%. Even then around 1000 runs were necessary to achieve this accuracy. All simulations have been implemented in C++. Note, that to omit exhausting a random number generator's cycle length we did not use the standard C++ random number generator but the one described in (L'Ecuyer et al 2002).

Simulation Results

Tables 1–4 contain comparisons of the justice under MFD, FCFS and SJF for the different values of the utilization ρ. We let MFD compete with both FCFS and SJF in terms of $\sigma^2(D), c(D)$, and $D_{\max}$. The table entries are the reduction factors for the corresponding metric for MFD compared to FCFS and SJF. That is the metric under FCFS or SJF, respectively, divided by the metric under MFD. Thus, a value greater than one indicates an improvement of justice, and as we can see all the values are greater than one, meaning that MFD improves the queue justice in all cases. The amount of this improvement depends on the utilization and is typically rapidly increasing with increasing ρ. It is also significantly greater for SJF than for FCFS, which is due to the extreme injust behaviour of SJF. As we can conclude from these results, in particular compared to SJF, MFD improves queue justice enormously, and compared to FCFS, too, there is a clear improvement.

		ρ				
		0.5	0.6	0.7	0.8	0.9
MFD	$\sigma^2(D)$	1.35	1.46	1.59	1.73	1.86
vs	$c(D)$	1.14	1.17	1.21	1.25	1.28
FCFS	$D_{\max}$	1.69	1.71	1.71	1.94	1.82
MFD	$\sigma^2(D)$	1.46	1.81	2.53	4.12	8.89
vs	$c(D)$	1.28	1.49	1.86	2.60	4.60
SJF	$D_{\max}$	5.15	9.82	10.38	20.03	36.30

Table 1: Queue Justice for M/M/1 queues

		ρ				
		0.5	0.6	0.7	0.8	0.9
MFD	$\sigma^2(D)$	1.19	1.24	1.29	1.36	1.42
vs	$c(D)$	1.08	1.15	1.18	1.21	1.26
FCFS	$D_{\max}$	1.40	1.38	1.50	1.65	1.64
MFD	$\sigma^2(D)$	2.11	3.02	4.62	8.24	19.58
vs	$c(D)$	1.38	1.65	2.03	2.67	4.16
SJF	$D_{\max}$	5.60	6.69	11.13	18.96	38.31

Table 2: Queue Justice for M/Erlang/1 queues

		ρ				
		0.5	0.6	0.7	0.8	0.9
MFD	$\sigma^2(D)$	1.11	1.19	1.31	1.54	1.75
vs	$c(D)$	1.02	1.08	1.12	1.19	1.25
FCFS	$D_{\max}$	1.38	1.40	2.08	1.85	1.75
MFD	$\sigma^2(D)$	1.11	1.25	1.56	2.58	6.63
vs	$c(D)$	1.05	1.20	1.39	1.94	3.72
SJF	$D_{\max}$	4.63	6.40	8.62	15.12	34.25

Table 3: Queue Justice for Erlang/M/1 queues

		ρ				
		0.5	0.6	0.7	0.8	0.9
MFD	$\sigma^2(D)$	1.40	1.39	1.38	1.38	1.38
vs	$c(D)$	1.18	1.17	1.17	1.16	1.16
FCFS	$D_{\max}$	2.37	2.32	2.24	2.17	2.09
MFD	$\sigma^2(D)$	2.05	2.24	2.49	2.83	3.33
vs	$c(D)$	1.55	1.69	1.92	2.33	3.40
SJF	$D_{\max}$	4.69	5.60	7.23	10.28	20.42

Table 4: Queue Justice for M/Pareto/1 queues

So far we have seen that MFD is successful in improving queue justice. Nevertheless, there must be a prize to pay, since SJF is known to minimize the expected response time (under non-preemptive scheduling policies). Surely, it would be not acceptable, if MFD increases the expected response time too much, and fortunately it does not. Compared to FCFS it even reduces both the expected response time and the variance of the response time. Thus, we do not only get an improvement of our main target metric but also in the most classical one. We omit to present the concrete values for FCFS, because they do not differ very much for the different models and utilizations. Roughly, MFD performs about 15–20% better than FCFS.

More interesting to see are the comparisons with SJF. Table 5 shows in a similar manner as Tables 1–4, but this time the loss of MFD compared to SJF in the expected response time $E[R]$, or in other words the increasing factor for the expected response time. We note that this is less than two in all models, and thus far less than the improvement or gain of justice. Hence, we strongly tend to accept this prize. Moreover, there is another improvement yielded by MFD, namely a reduction of the variance of the response time, as shown in Table 6. Altogether the results of our study show that the only metric, where MFD is worse than SJF is the expected response time, for which SJF is known to be optimal, and MFD performs better than FCFS in terms of all evaluated metrics.

	ρ				
	0.5	0.6	0.7	0.8	0.9
M/M/1	1.11	1.18	1.30	1.51	1.98
M/Erlang/1	1.03	1.06	1.09	1.15	1.28
Erlang/M/1	1.02	1.06	1.13	1.27	1.66
M/Pareto/1	1.09	1.12	1.16	1.20	1.25

Table 5: Loss in $E[R]$ compared to SJF

	ρ				
	0.5	0.6	0.7	0.8	0.9
M/M/1	1.01	1.08	1.24	1.63	2.91
M/Erlang/1	1.41	1.70	2.25	3.36	6.71
Erlang/M/1	1.01	1.03	1.09	1.34	2.41
M/Pareto/1	1.33	1.33	1.30	1.21	1.23

Table 6: Reduction of $\sigma^2(R)$ compared to SJF

CONCLUSION

We investigated the problem of justice in queues and in particular the topic of scheduling to improve justice which is an important task for systems where human customers are involved. In such cases justice is often more important than response time. We have used metrics based on the discrimination frequency to quantify justice, and we have demonstrated by a simulation study that the newly designed scheduling policy MFD outperforms FCFS and SJF with regard to justice. MFD also improves expected response times and the variance of response times compared to FCFS. Compared to SJF, MFD yields larger expected response times but at the same time reduces the response time variance. Thus, MFD should be considered in a variety of situations where expected response times do not play the only dominating role, but where personal satisfaction of customers is driven by the feeling of justice.

Research on quantitative justice and related scheduling policies will be continued, including further investigations of MFD and its properties, where analytical results would be particularly worthy. Comparisons of MFD with other scheduling policies are of interest, too, and additionally considering preemptions may result in a preemptive version of MFD. Furthermore, extensions to multi-server queues or variants like impatient customers seem to be reasonable.

REFERENCES

Avi-Itzhak, B. and Levy, H. 2004. ”Measuring Fairness in Queues”. *Advances in Applied Probability* 36, No.3, 919–936.

Conway, R. W.; W. L. Maxwell; and L. W. Miller. 1967. *Theory of Scheduling.* Addison-Wesley.

Kleinrock, L. 1975. *Queueing Systems. Volume I: Theory.* Wiley & Sons.

Larson, R. C. 1987. ”Perspectives on Queues: Social Justice and the Psychology of Queueing”. *Operations Research* 35, No. 6, 895–905.

Law, A. M. and W. D. Kelton. 2000. *Simulation Modeling and Analysis.* 3rd ed., McGraw Hill.

L'Ecuyer, P.; R. Simard; E. J. Chen; and W. D. Kelton. 2002. "An Object-Oriented Random-Number Package with Many Long Streams and Substreams". *Operations Research* 50, No. 6, 1073–1075.

Rafaeli, A.; G. Barron; and K. Haber. 2002. "The Effects of Queue Structure on Attitudes". *Journal of Service Research* 5, No. 2, 125–139.

Rafaeli, A.; E. Kedmi; D. Vashdi; and G. Barron. 2005. "Queues and Fairness: A Multiple Study Experimental Investigation". Manuscript under Review.

Raz, D.; H. Levy; and B. Avi-Itzhak. 2004. "A Resource Allocation Queueing Fairness Measure". *Performance Evaluation Review* 32, No. 1, 130–141.

Sandmann, W. 2005. "A Discrimination Frequency Based Queueing Fairness Measure with Regard to Job Seniority and Service Requirement". In *Proceedings of the 1st EuroNGI Conference on Next Generation Internet* (Rome, Italy, April 18–20). IEEE Computer Society Press, 106–113.

Sandmann, W. 2006. "Analysis of a Queueing Fairness Measure". In *Proceedings of the 13th GI/ITG Conference on Measurement, Modelling and Evaluation of Computer and Communication Systems* (Nürnberg, Germany, March 27–29). VDE Verlag, 219–231.

Wierman, A. and M. Harchol-Balter. 2003. "Classifying Scheduling Policies with Respect to Unfairness in an M/GI/1". In *Proceedings of the ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems* (San Diego, CA, USA), 238–249.

AUTHOR BIOGRAPHY

WERNER SANDMANN studied Computer Science with Mathematics as a supplementary subject at the University of Bonn, Germany, where he received his diploma degree (Dipl.–Inform.) and his PhD (Dr. rer. nat.) in 1998 and 2004, respectively. From 1998–2003 he was a Research and Teaching Assistant at the Computer Science Department of the University of Bonn. Since 2004 he is an Assistant Professor of Computer Science at the University of Bamberg, Germany. His email address is `werner.sandmann@wiai.uni-bamberg.de`.