On the Approximation of Higher Moments in Open and Closed Fork/Join Primitives with Limited Buffers

Manfred Mittler, Tarik Ono-Tesfaye, and Alexander K. Schömig

Bayerische Julius-Maximilians-Universität Würzburg
Lehrstuhl für Informatik III
Am Hubland, D – 97074 Würzburg
Tel.: +49-931-8885518, Fax: +49-931-8884601
e-mail: mittler@informatik.uni-wuerzburg.de

Abstract In this paper we present an approximate analysis of fork/join primitives with two parallel servers, limited buffers and non-zero join times where either jobs arrive according to a Poisson process (*open system*) or the number of jobs is limited according to the CONWIP rule (*closed system*). The approximation method is based on the concept of flow-equivalent servers and enables the approximate analysis of higher moments of the cycle time. We extend the concept of flow-equivalent servers and replace the two-processor fork/join primitive by a queuing network consisting of two single servers in tandem where one of them has state-dependent service rates. The comparison of approximate to exact results for both the mean and the variance of response times shows that our method works very accurate for a large range of parameter settings.

1 Introduction

Fork and join queues are used in the analysis of parallel computer systems and assembly/disassembly manufacturing systems. Typical examples include automotive, aerospace, and other metal–working production facilities. Since it is very difficult to deal with synchronizations by means of queuing theory, much attention has been devoted to the approximation and calculation of bounds for throughput and mean response times.

However, most publications consider only fork/join networks with infinite buffer lengths and zero join times. Rao and Suri (1994) present an extensive literature review of relevant papers dealing with assembly/disassembly manufacturing systems. They investigate fabrication/assembly systems and derive algorithms for calculating approximate mean throughput and queue length. The simplifying assumption of exponential service times is made. As the authors underline, so far existing studies concerning similar assembly systems have also used this assumption.

Currently, in industries the capability of meeting due dates has become a crucial factor of competitiveness. Consequently, not only the mean of performance measures has to be considered but also the variance, since the probability of meeting due dates decreases with increasing process variability.

2 Model Description

A *fork-join* queuing network is characterized by a *fork* node where jobs are split into several tasks and a *join* node where the tasks are rejoined and merged into one job. In a manufacturing environment a fork node may be a machine that executes a disassembly operation; the join operation in the queuing model then represents an assembly operation.

In this paper we look at open and closed fork/join networks in which jobs are forked into exactly two tasks that have to be processed separately. Fig. 1 shows the corresponding open fork/join network. Since these networks are the simplest ones that incorporate fork and join operations we refer to these networks as *fork/join primitives*. According to Duda and Czachórski (1987) we indicate the splitting of jobs into two tasks by the prefix *two-processor*.

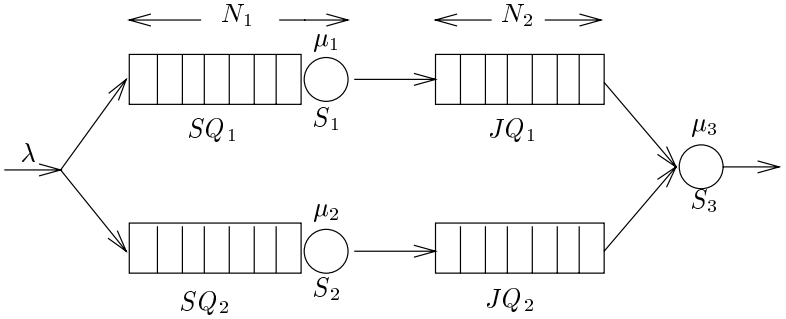

Fig. 1: Open Fork-Join Primitive

We assume that in the open system the customer arrival process is Poisson with rate λ. The *task service station* i, which consists of queue SQ_i and server S_i, can hold N_1 tasks and the join queue JQ_i can hold N_2 tasks (excluding the join server). An arriving job is rejected if either task service station 1 or 2 is fully occupied. Otherwise it is forked into two tasks. Task i, $i = 1, 2$, then waits in queue SQ_i for service in server S_i. We assume that the servers S_1 and S_2 have the same exponential service time distribution with mean μ^{-1} ($\mu_1 = \mu_2 = \mu$). After service completion the task arrives in queue JQ_i. There has to be at least one task waiting in both queues JQ_1 and JQ_2 before they can be joined by server S_3 and the re-joined job can leave the open fork-join system. The time needed for the join operation is also exponentially distributed and has a mean of μ_3^{-1}. The spliting of jobs, however, takes place without time delay. It is furthermore assumed that in the open network servers S_1 and S_2 are blocked according to the *manufacturing blocking*, that is, if queue JQ_i is fully occupied, server S_1 stops service until the number of waiting tasks in JQ_i decreases.

If we apply the CONWIP rule (cf. (Spearman, Woodruff, and Hopp 1990)) to the open fork/join primitive we obtain the corresponding closed network. The number of jobs in the network is constant N such that a job does not leave the system after the join operation but is returned to the fork node, where it is again split into two tasks. Clearly, we assume that the service stations 1 and 2 as well

as the queues JQ_1 and JQ_2 have enough waiting places for all tasks so as to avoid blocking and losses.

3 Approximate Analysis

The approximate analysis is based on the concept of flow-equivalent servers which has been employed by Duda and Czachórski (1987) to approximate the throughput and the mean cycle time of open fork/join primitives with infinite buffers and zero join times. According to Norton's theorem for queuing networks Duda and Czachórski first "short circuit" the original open two-processor fork/join primitive and show that the throughput of the resulting closed system is the throughput of the join node. The short-circuited fork/join primitive is then replaced by a state-dependent server. Its state-dependent service rate is equal to the throughput of the short-circuited fork/join primitive. We refer the reader to the original literature for further details. Obviously, this approach can be applied to nested fork/join networks. All fork/join primitives are recursively replaced by state-dependent servers until a network containing no fork/join primitives is achieved. Since this approximation method is based on flow or throughput equivalent servers, by Little's law it also yields approximate results for the job mean cycle time in fork/join networks.

We extend the method of Duda and Czachórski to the approximation of open and closed fork/join primitives with limited buffers, manufacturing blocking, and non-zero join times. Due to the lack of space, we only describe our new method informally. For technical details we refer to Mittler, Ono-Tesfaye, and Schömig (1995). We present the idea of the higher moment approximation for the open network only. The approximation of the closed networks is analogous to the approximation of the open network and not discussed explicitly.

If the join times are zero, queues JQ_1 and JQ_2 cannot be occupied at the same time since tasks joined to jobs leave the system immediately without delay. However, if we assume non-zero join times, waiting times in front of the join server may occur. In this case we divide queues JQ_1 and JQ_2 as follows: queue JQ_1^* contains only those tasks processed by processor S_1 which have to wait for their corresponding mates, which are still to be processed by processor S_1 and vice versa. Then, the tasks proceed together to an additional queue where they have to wait to be joined. Obviously, as in the non-zero network, one of the queues JQ_1^* and JQ_2^* is empty at any given time. We therefore replace the fork/join primitive consisting of the original fork part of the network, queues JQ_1^* and JQ_2^* and the additional queue for the mates by a state-dependent server and a second queue in tandem as shown in Fig. 2.

To approximate higher cycle time moments of the closed fork/join primitive, we only have to omit the external arrival process of jobs in Fig. 2 and have to apply the CONWIP rule such that N tasks travel to a closed queuing network consisting of two exponential service stations where the service rates of the first one are state-dependent.

It remains to explain how we calculate higher task cycle time moments in open and closed queuing networks like the one shown in Fig. 2. To the authors'

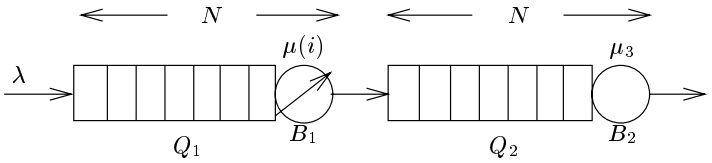

Fig. 2: Substitute Open Network

knowledge there is currently no explicit expression for neither the entire cycle time distribution nor the higher moments of cycle times. The literature deals only with non-state-dependent service rates (cf. Boxma and Daduna (1990)).

We therefore calculate higher moments of cycle times as follows. First we derive the steady-state distribution of the number of jobs in the queues of the networks previously discussed. According to Syski (1986) we are able to derive an ordinary differential equation (*ode*) for the cycle time distribution under the condition that the system is in a particular state at the arrival instant of a particular job. Syski's method employs Kolmogorov's backward equation and the concept of taboo sets. However, from this ode it is very difficult if not impossible to get an explicit expression for the cycle time distribution. Nevertheless, Kühn (1972) obtained a simple recursion for the higher moments of the cycle time distribution. The second step of the analysis of the substitute networks is then the recursive calculation of higher moments according to Kühn. Since this approach yields higher moments conditioned on the state at the arrival instant only, these moments have to be unconditioned using the steady-state distribution of the number of jobs in the queues of the substitute networks. In the following, we refer to this layered method as the Syski-Kühn-Method.

One might suggest to apply the method just described to the original open and closed fork/join primitives. We have already done so since we needed a method to obtain exact cycle time moments. However, due to the large increase of the number of states this method can hardly be applied to networks with nested fork and join operations. We therefore propose to apply our new method to fork/join networks recursively until all fork/join operations have been replaced by tandem networks consisting of two exponential queues where the first one a has state-dependent service rate.

4 Results

As mentioned above we compare the approximate results obtained by using our new approximation method to exact results which were obtained through the application of Syski-Kühn-Method to the original open and closed fork/join primitives. We present results for the cycle time $S(N)$ as a function of either the storage capacity of the open fork/join primitive ($N_1 = N_2 = N$) or of the fixed number of jobs N in the closed fork/join primitive.

Fig. 3 shows exact and approximate results for the mean cycle time $E[S(N)]$ and the variance of cycle time $Var[S(N)]$ of the open fork/join primitive for the

parameter setting $\lambda = 2$ and $\mu_1 = \mu_3 = 4$. We can see that our approximate analysis works very accurately although it slightly underestimates both the mean and the variance. The corresponding relative errors are less than 3 % for the

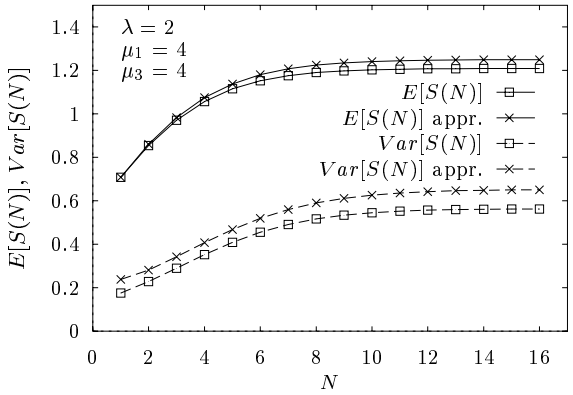

Fig. 3: Mean and Variance of Cycle Times in the Open Fork/Join Network

mean whereas for the variance they range up to 15 % approximately. However, from a operations manager's point of view, relative errors in the range of 15 % are tolerable. We examined this system for higher arrival rates as well. In this case the approximation accuracy increases. For example for the arrival rate of $\lambda = 4$ the relative errors for $E[S(N)]$ fall below 1.5 %, the relative errors for $Var[S(N)]$ below 5 %. Furthermore, we can see that the approximation accuracy also increases with increasing buffer capacity N.

If we apply the CONWIP rule to this system and keep the service rate unchanged we obtain the results depicted in Fig. 4. Obviously, the difference between the approximate and the exact results is rather small. For the mean the

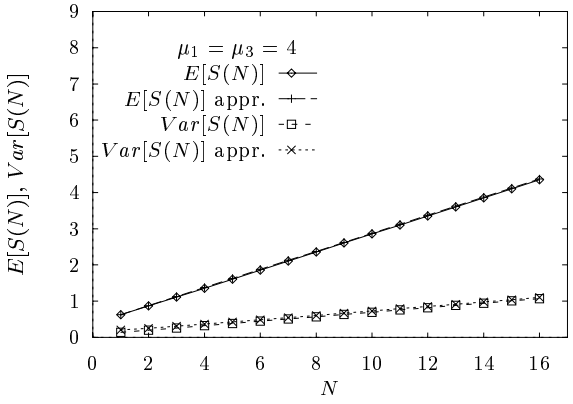

Fig. 4: Mean and Variance of Cycle Times in the Closed Fork/Join Network

relative error is within 1 %; the relative error for the variance of cycle times falls below 10 % when the number of jobs N allowed to enter the system is increased to 16. We further examined this closed system with varying service rates. The results suggest that different service rates can lead to a better approximation accuracy. Above all, if the join service rate μ_3 is small compared to the task service rates, the relative errors decrease to almost zero for $N > 10$. Finally, the approximation accuracy improves with increasing N which here denotes the total number of jobs in system.

5 Conclusion

In this paper we presented a simple approximation method for fork/join primitives with either external arrival of jobs or closed loop inventory control. Since the buffers of our models are limited with assumed that the production is triggered according to manufacturing blocking. In addition to those models investigate in the literature up to now we considered non-zero join times. The comparison of exact to approximate results show that our new method performs very accurately for a large range of parameters settings. It remains to investigate how this method performs for nested fork/join networks.

References

Boxma, O. J. and H. Daduna (1990). Sojourn times in queueing networks. In H. Takagi (Ed.), *Stochastic analysis of computer and communication systems*, pp. 401–450. North-Holland.

Duda, A. and T. Czachórski (1987). Performance evaluation of fork and join synchronization primitives. *Acta Informatica 24*, 525–553.

Kühn, P. J. (1972). *Über die Berechnung der Wartezeiten in Vermittlungs- und Rechnersystemen*. Ph. D. thesis, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, Germany. in German.

Mittler, M., T. Ono-Tesfaye, and A. K. Schömig (1995, August). Higher moment approximation of open and closed fork/join primitives with limited buffers. Forschungsbericht, Preprint-Reihe, Universität Würzburg, Institut für Informatik. In preparation.

Rao, P. C. and R. Suri (1994, Fall). Approximate queueing network models for closed fabrication/assembly systems. part I: Single level systems. *Journal of Production and Operations Management 3*(4), 244–275.

Spearman, M. L., D. L. Woodruff, and W. J. Hopp (1990). CONWIP: a pull alternative to kanban. *International Journal of Production Research 28*(5), 879–894.

Syski, R. (1986). *Introduction to Congestion Theory in Telephone Systems* (2 ed.). Elsevier Science Publishers B. V. Originally published 1960 by Oliver & Boyd, Edinburgh, Scotland.