Design of a Tree-Queue Model for a Large-Scale System

Byungsung PARK†, Jaeyeong YOO†, Nonmembers, and Hagbae KIM†(a), Member

SUMMARY In a large queuing system, the effect of the ratio of the filled data on the queue and waiting time from the head of a queue to the service gate are important factors for process efficiency because they are too large to ignore. However, many research works assumed that the factors can be considered to be negligible according to the queuing theory. Thus, the existing queuing models are not applicable to the design of large-scale systems. Such a system could be used as a product classification center for a home delivery service. In this paper, we propose a tree-queue model for large-scale systems that is more adaptive to efficient processes compared to existing models. We analyze and design a mean waiting time equation related to the ratio of the filled data in the queue. Based on simulations, the proposed model demonstrated improvement in process efficiency, and it is more suitable to realistic system modeling than other compared models for large-scale systems.

key words: large-scale system, tree-queue model, queuing process, waiting time, design

1. Introduction

Recent technological development makes it possible to produce large-scale systems which require a multidisciplinary approach. In the development of these large-scale systems, a queuing model is used to approximate a real queuing system. A queuing model allows for the mathematical analysis of queuing behavior. Additionally, this modeling is of increasing importance in determining the best solution for the design of efficient systems [1], [2].

Various research efforts have focused on system modeling based on queuing theory, networks, and models. Alexei Vazquez presented two queuing models which captured both static and dynamic behaviors for resource management [3]. J. Mac and Gregor Smith described the modeling of manufacturing systems using queuing network models and two other closely related modeling techniques, simulation modeling and generative models, for a buffer allocation problem [4]. Akin brought together useful models and modeling approaches which addressed a wide variety of system design and operational issues [5]. Their research assumes that the effects of the process waiting time and the ratio of data filled in the queue can be considered negligible, according to the queuing theory. However, for a large-scale system, these factors must be considered due to an increase in effect depending on the system size. The system is different from the existing system at such a point. A product classification center is a good example of the character of a large-scale system: various products enter the center through a single entrance and are then divided into the many available exits according to their delivery destination. A large-scale system is designed with consideration of queue overflow, thus ensuring that the system always has sufficient queue capacity. This fact enables us to disregard a saturated queue [1], [6].

In this paper, we propose a tree-queue model that focuses on the ratio of filled data in the queue for realistic modeling of large-scale systems. Our proposed model consists of three layers in order to decrease the bottleneck in a typical queuing process. These layers enable the queue to distribute data traffic flow to each service gate, thus improving the queue waiting time. We analyze the mean waiting time equation, which reflects the process distance, using Little’s equation and the queuing theory. Simulations demonstrate that our proposed model has improved process efficiency compared to those of other models.

The paper is organized as follows: Sect. 2 describes our proposed model for large systems, Sect. 3 analyzes the tree-queue model with respect to the effects of process distance and the ratio of filled data in the queue, Sect. 4 presents the simulation results indicating improvement in process efficiency and a reduction in the mean waiting time, and Sect. 5 provides our research conclusions.

2. Tree-Queue Model

In a queue model, it is important to prevent from occurring bottleneck due to decreasing the waiting time and improving the process efficiency. It is known that complete elimination of a bottleneck is impossible in queuing theory, therefore it is necessary to distribute the data traffic flow to solve the bottleneck issue. We propose a tree-queue model (TQM) that consists of three layers by extending the fork queue shown in Fig. 1. The data must pass two diverging points while it moves from the entrance to the end of the queue. At each diverging point, the data chooses its next branch based on which has the least amount of waiting data. The queue bottleneck is thus decreased because this process allows for the efficient distribution of data to the service gates. If all the service gates are occupied by data, they wait in front of the each service gate, which means that data wait in queue.

In proposed model shown in Fig. 1, the three layers consist of a root, a parent and a child layer. In the child layer, branches connected to a parent branch are given a number
The process distance in the tree-queue model is described as the head of a queue to the service gate. We designed the tree-or data require some amount of time to walk or move from time. The process distance presents the concept that people recalculate the throughput time. Thus, the throughput

\[
E(\tau_{n_f,n_s}) = \frac{1}{\mu} + \frac{d_0 + d_1 + d_2 + k_1n_f + k_2n_s}{p}
\]  

(2)

Then we obtain the throughput rate.

\[
\mu_{n_f,n_s} = \frac{1}{E(\tau_{n_f,n_s})} = \frac{1}{\mu} + \frac{d_0 + d_1 + d_2 + k_1n_f + k_2n_s}{p}
\]  

(3)

where \( \mu \) is

\[
\hat{\mu} = \frac{1}{\mu} + \frac{d_0 + d_1 + d_2}{p}
\]  

(4)

we can calculate mean throughput rate.

\[
\hat{\mu} = \frac{1}{n} \sum_{n_f,n_s \in S,G} \frac{\hat{\mu}}{1 + \frac{\hat{\mu}}{p}(k_1n_f + k_2n_s)}
\]  

(5)

In Eq. (5), \( n \) is the service gate number and \( S,G \) is the value set of \( n_f \) and \( n_s \) that the service gates have. We apply stationary equation of M/M/8/k [1], [7], the \( k \) is described as follows:

\[
k = d_0 + d_1n_{f\text{cnt}} + d_2n_{f\text{cnt}}n_{s\text{cnt}} + k_1(n_{f\text{cnt}} - 1) + k_2n_f(n_{s\text{cnt}} - 1)
\]  

(6)

where \( k \) is the maximum capacity of queue. In Eq. (6), \( n_{f\text{cnt}} \) is the total number of branch, and \( n_{s\text{cnt}} \) is the total number of service gate. By using Eqs. (5) and (6), TQM’s stationary equations are described as follows:

\[
\begin{align*}
\lambda P_0 &= \hat{\mu}_1 P_1 \\
\lambda P_{n-1} + (n+1)\hat{\mu}_{n+1}P_{n+1} &= (\lambda + n\hat{\mu}_n)P_{n} \\
(1 \leq n \leq s - 1) \\
\lambda P_{n-1} + s\hat{\mu}_{n}P_{n+1} &= (\lambda + s\hat{\mu}_s)P_{n} \\
(s \leq n \leq k)
\end{align*}
\]  

(7)

Then we analytically calculate \( P_0, P_n \) and mean waiting time \( W_q \) by using Eq. (7) and Little’s theorem [2].

\[
P_0 = \left\{ \sum_{j=0}^{s-1} \frac{a^j}{j!} + \frac{a^s}{s!} \left( \frac{1 - \rho^{k-s+1}}{1 - \rho} \right) \right\}^{-1}
\]  

(8)

\[
P_n = \left\{ \frac{a^n}{s^n!\rho^n}P_0 \right\} (1 \leq n \leq s - 1)
\]  

(9)

\[
W_q = \frac{L_q}{\lambda(1 - P_k)}
\]  

(10)

\[
L_q = \sum_{n=s}^{k} (n-s)P_n
\]  

(11)

where, \( a = \frac{\lambda}{\mu} \) and \( \rho = \frac{\lambda}{\mu} \). By controlling parameter such as \( d_0, d_1, d_2, k_1, n_f, k_2, \) and \( n_s \), we obtain \( W_q \) of original and fork queue. Thus our queue model can include characters of another queues.

in order of precedence from the center branch. In the same way, a parent branch connected to a root branch is given a number. Each numbered branch is symmetrically added except for the central branch. In our simulations, a piece of data enters the queue with a probability \( \lambda \), a service rate \( \mu \), and proceeds to the next cell with a probability \( p \).

Realistic systems demonstrate that, as the system scale increases, the process distance is of greater importance in the modeling due to the increasing effects on the waiting time. The process distance presents the concept that people or data require some amount of time to walk or move from the head of a queue to the service gate; we designed the tree-queue model in consideration of the process distance. The process distance in the tree-queue model is described as

\[
d_q = d_0 + d_1 + d_2 + k_1n_f + k_2n_s
\]  

(1)

where, \( n_f \) is parent branch number and \( n_s \) is child branch number.

3. Analysis of Tree-Queue Model

To reflect the effects of process distance in a large system, we recalculate the throughput time. Thus, the throughput time \( E(\tau_{n_f,n_s}) \) that has first branch number \( n_f \) and a second branch number \( n_s \) is described using the mean field approximation.
4. Simulations

The analytical modeling was verified using MATLAB simulations for various queue models, such as parallel (M/M/1 × 15), fork (M/M/s), and tree-queue models (M/M/s/k), where s is 15, k is 2340 according to Eq. (6). Based on the simulations, we achieved improvements in process efficiency and a reduction in the mean waiting time by employing the proposed tree model. A comparative analysis of a tree-queue model and previous models was conducted. For the simulations, we used the queue length parameter values of $d_0 = 100$, $d_1 = 100$, $d_2 = 100$, $p = 0.5$, $k_1 = 40$, $k_2 = 30$, $n_{f_{cnt}} = 3$, and $n_{s_{cnt}} = 5$. Service rate, $\mu$ was 0.05 and the ratio of waiting time. We assumed that the distribution of the service time was an exponential distribution. As a result, the mean waiting time $W_q$ of the tree-queue was smaller than that of a parallel queue in $0 \leq \rho \leq 0.7$ according to Fig. 2. The utilization factor $\rho$ the amount of data contained in a queue; this range shows that the queue is moderate. We can observe the crossing of graph: We observe the crossing of the graph: tree-queue and parallel queue. This means that when the utilization $\rho$ less than 0.7, we should form a tree-queue in order to decrease the waiting time $W_q$. However, when the $\rho$ is greater than 0.7, we should form a parallel-queue. As previously mentioned, we are able to disregard a saturated queue for a large-scale system. Therefore, we propose our system as a modeling solution for large-scale systems.

5. Conclusion

In the large-scale systems, such as a product classification center for a home-delivery service, it is important to factor in the waiting time, which considers the ratio of data filled in the queue prior to modeling. In this paper, we proposed a tree-queue model for large-scale system that focuses on waiting time depending on the ratio of data filled in the queue. We analyzed the mean waiting time equation using Little’s equation and the queuing theory and validated the proposed model using comparisons to earlier research on large-scale systems. Experimental results show that the proposed model demonstrated improvements in process efficiency and a reduction in the mean waiting time within the certain threshold value compared to the parallel and fork models.

References

[1] D. Yanagisawa, A. Tomoeda, A. Kimura, and K. Nishinari, “Walking-distance introduced queueing theory,” Lect. Note Comput. Sci., vol.5191, pp.455–462, 2009.
[2] D.G. Kendall, “Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain,” Ann. Math. Statist., vol.24, no.3, pp.338–354, 1953.
[3] A. Vazquez, J.G. Oliveira, Z. Dezso, K. Goh, I. Kondor, and A. Barabasi, “Modeling bursts and heavy tails in human dynamics,” Physical Review E.73, March 2006.
[4] J. Mac and G. Smith, “Design of large scale systems with open finite queueing networks,” Conference on Decision and Control, pp.1088–1091, Dec. 1987.
[5] R.G. Askin and C.R. Standridge, Modeling and analysis of manufacturing systems, Wiley, 1993.
[6] D. Helbing, M. Treiber, and A. Kesting, “Understanding inter arrival and inter departure time statistics from interactions in queueing systems,” Physica A, vol.363, pp.62–72, 2006.
[7] D. Gross, J. Shortle, J. Thompson, and C. Harris, Fundamental of Queueing Theory, John Wiley & Sons, 2008.