# About Virtual Waiting Time in a Multiprocessor System

Vladimir Sahakyan
Institute for Informatics and Automation Problems of NAS RA
Yerevan, Armenia
email: vladimir.sahakyan@sci.am

Artur Vardanyan
Institute for Informatics and Automation Problems of NAS RA
Yerevan, Armenia
email: artur.vardanyan@iiap.sci.am

***Abstract*—In multiprocessor queueing systems, the waiting time to start servicing tasks is crucial to optimize system performance. However, traditional measures of latency may not accurately reflect the actual task experience due to differences in service requirements and resource availability.**

**The concept of virtual waiting time has become a valuable metric that takes into account the amount of time tasks would wait if they arrived at a given time.**

**The article analyzes the virtual waiting time in a multiprocessor queuing system with limited queue capacity. Taking into account the distributions of arrivals and servicing of tasks, equations are derived that describe the behavior of the virtual waiting time.**

**Based on the analysis, expressions for estimating the expected virtual waiting time are obtained. The conclusions will make it possible to optimize task scheduling and resource allocation in multiprocessor systems, increasing system performance.**

***Keywords*— Multiprocessor system, Queueing system, Scheduling, Virtual waiting time.**

## I. Introduction

This article discusses a single-threaded multiprocessing queuing system, which is denoted by $M|M|m|n$. Jobs will be serviced in the order they arrive in the system, i.e., FIFO discipline is used[1], [2]. The system consists of $m$ processors (cores, cluster nodes, etc.,) dedicated to task servicing, with a queue capable of accommodating up to $n$ tasks awaiting service [3].

The incoming stream of tasks follows a distribution function represented by:

$$A(x) = 1 - e^{-ax},$$

where $a > 0$. Each task requires $\nu$ $(1 \leq \nu \leq m)$ processors for a duration of $\beta$ $(\beta > 0)$. Hence, every task is characterized by two random parameters: $(\nu, \beta)$. Here, $\nu$ denotes the number of computing resources required for task servicing, while $\beta$ represents the maximum time needed to complete the task.

The task parameters are subject to the following distributions:

$$P(\nu = k) = p_k,$$

where $k = 1, 2, ..., m$ and we will obviously have the following relation between $p_k$:

$$\sum_{k=1}^{m} p_k = 1.$$

The distribution function of $\beta$ is given by:

$$P(\beta < x) = B(x),$$

where $B(x)$ follows an exponential distribution:

$$B(x) = e^{-bx}, \text{ where } b > 0.$$

In a steady-state queuing system, the system's condition at a random moment can be characterized by the probabilities of different states denoted by $P_{i,j}$. Here, $P_{i,j}$ represents the probability that there are $i$ tasks being serviced $(1 \leq i \leq m)$, while $j$ tasks are waiting in the queue $(1 \leq j \leq n)$. In order to determine the values of $P_{i,j}$, a system of linear equations is proposed in [4] to account for a more comprehensive scenario.

This paper aims to analyze the virtual waiting time in the aforementioned queueing system by leveraging the steady-state probabilities derived earlier.

## II. Main Notations

This section introduces the main notations and concepts used in the subsequent analysis.

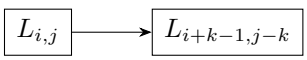
Consider a vector of independent identically distributed random variables $(\beta_1, \beta_2, ..., \beta_i)$, each following the distribution function $B(x)$.

Let us define the random variable:

$$\beta_{i,1} = min(\beta_1, \beta_2, ..., \beta_i).$$

Next, the following events are considered for $1 \leq i \leq m$, $1 \leq j \leq n$, and $0 \leq k \leq \min(j, m - i + 1)$:

$S_{i,j}(k)$ represents the event when, at the nearest end of service, $k$ tasks will be accepted from the queue for servicing, given that there were $i$ tasks being serviced and $j$ tasks in the queue. In other words, if we describe $S_{i,j}(k)$ through the logic of system state changes, it corresponds to the following state transition:




where in accordance with [4], we used the notation $L_{i,j}$ to denote the state of the system when $i$ tasks are being serviced, and $j$ tasks are waiting in the queue.

https://doi.org/10.51408/csit2023_23

Let $I_{i,j}(k)$ denote the progress indicator of the event $S_{i,j}(k)$. Specifically,

$$I_{i,j}(k) = \begin{cases} 1, & \text{if the event } S_{i,j}(k) \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases}$$

The random variable $\tau_{i,j}$ is also introduced to represent the time until the queue becomes empty of all backlogged tasks when there are $i$ tasks $(1 \leq i \leq m)$ in the system and $j$ tasks $(1 \leq j \leq n)$ waiting in the queue.

Additionally, we define $\omega_{i,j}$ as the virtual waiting time before the start of service for a task that enters the system when there are $i$ tasks being served and $j$ tasks waiting in the queue.

With the notations and concepts defined above, we are now equipped to formulate equations for the analysis of the virtual waiting time in the queueing system described earlier. These equations will enable us to delve deeper into the characteristics and behavior of the system.

## III. Lemmas

This section derives some supporting probabilities and formulas that play a crucial role in evaluating the virtual waiting time in the presented queueing system.

*Lemma* 1. The probability of the event that $\beta_{i,1} > x$, where $x > 0$ is determined as follows:

$$P(\beta_{i,1} > x) = e^{-ibx}.$$

*Lemma* 2. The mathematical expectation of $\beta_{i,1}$ is equal:

$$\mathbb{E}(\beta_{i,1}) = \frac{1}{ib}.$$

Before to describe the probability of the event $S_{i,j}(k)$, some events are defined:
Let $C_{i,j,k}$ be the event that

$$\sum_{l=1}^{i-1} \nu_l + \sum_{l=1}^{k} \nu_l \leq m < \sum_{l=1}^{i-1} \nu_l + \sum_{l=1}^{k+1} \nu_l,$$

if $0 \leq k < min(m - i + 1, j - 1)$ and

$$\sum_{l=1}^{i-1} \nu_l + \sum_{l=1}^{k} \nu_l \leq m,$$

if $k = j$ and $i + j - 1 \leq m$.
Let $D_i$ be the event that

$$\sum_{l=1}^{i} \nu_l \leq m < \sum_{l=1}^{i+1} \nu_l.$$

For a more detailed formulation, obviously, the probability of the event $S_{i,j}(k)$ can be represented as follows:

*Lemma* 3.

$$P(S_{i,j}(k)) = \begin{cases} 1, & \text{if } k = 0 \\ P_{i,j} P\Big(C_{i,j,k} \Big/ D_i\Big), & \text{if } k \leq i-1 \\ 0, & \text{if } k = i \end{cases},$$

where $1 \leq i < m$ and $1 \leq j \leq n$, the conditional probability $P\Big(C_{i,j,k} \Big/ D_{i,k}\Big)$ signifies the probability of event $C_{i,j,k}$ given that event $D_{i,k}$ has occurred.

The proofs and detailed explanations of these lemmas can be found in our previous publication [4].

## IV. Equations

In this section, the equations are constructed to allow the evaluation of the virtual waiting time in detail. By examining the system's behavior and considering the established notations, the aim is to gain insights into the expected waiting time before the start of service for tasks entering the system.

The equations will be formulated that capture the dynamics of task arrival and service completion characteristics. These equations will enable us to estimate the virtual waiting time based on the system's configuration and parameters.

Thus, to describe $\tau_{i,j}$, a system of equations is composed in terms of random variables:

$$\begin{gathered} \tau_{i,0} = 0, \text{ for } i = 0, 1, ..., m; \\ \tau_{i,j} = \beta_{i,1} + \sum_{k=0}^{K} I_{i-1,j}(k) \tau_{i+k-1,j-k}, \\ i = 1, 2, ..., m; \\ j = 1, 2, ..., n; \end{gathered} \tag{1}$$

where $K = \min(m - i + 1, j)$.

Let $\mathbb{E}(\tau_{i,j})$ represent the mathematical expectation of $\tau_{i,j}$. Considering the independence of the random variable $\tau_{i,j}$ from the system's past events, we can express equation (1) as follows:

$$\mathbb{E}(\tau_{i,j}) = \mathbb{E}(\beta_{i,1}) + \sum_{k=0}^{K} P(S_{i,j}(k)) \mathbb{E}(\tau_{i+k-1,j-k}). \tag{2}$$

Next, it is important to highlight the relationship between the virtual waiting time and the time until the queue becomes empty of all backlogged tasks. Specifically,

$$\omega_{i,j} = \tau_{i,j+1}. \tag{3}$$

Furthermore, it is observed that the mathematical expectations of these two variables are equal:

$$\mathbb{E}(\omega_{i,j}) = \mathbb{E}(\tau_{i,j+1}). \tag{4}$$

This equality allows us to analyze and interpret the expected virtual waiting time in terms of the time until the queue is empty. By leveraging (1) and (2) the equations gained from the analysis of $\tau_{i,j}$ and $\mathbb{E}(\tau_{i,j})$, we can gain the equations that will allow us to evaluate the virtual waiting time in our multiprocessor queueing system.

Therefore, we can present the key findings of our study:

*Theorem.* The expected virtual waiting time can be determined by the following expression:

$$\mathbb{E}(\omega_{i,j}) = \mathbb{E}(\beta_{i,1}) + \sum_{k=0}^{K} P(S_{i,j+1}(k)) \mathbb{E}(\tau_{i+k-1,j-k+1}),$$

where $K = \min(m - i + 1, j)$.

This theorem provides a concise and insightful formulation for computing the expected virtual waiting time in the presented multiprocessor queueing system.

The upcoming publication will delve into further details regarding the development of the computational algorithm, its implementation, and the performance evaluation of our proposed methodology. The publication will provide valuable insights into the practical aspects and real-world implications of our findings.

## V. Conclusion

This study investigated the virtual waiting time in a multiprocessor queueing system with a limited queue capacity. The analysis focused on understanding the behavior and properties of the virtual waiting time, as well as its relationship with the time until the queue becomes empty of backlogged tasks. Based on this analysis, the expression has been obtained to evaluate the expected virtual waiting time.

The findings contribute to the understanding of task scheduling and resource allocation strategies in multiprocessor systems, providing valuable insights for system designers and administrators. By comprehending the virtual waiting time, system performance can be optimized, and task scheduling algorithms can be developed to minimize waiting times and enhance overall system efficiency.

In conclusion, this study sheds light on the virtual waiting time in multiprocessor queueing systems, offering valuable insights and a foundation for future research in this domain. By understanding and optimizing the virtual waiting time, we can enhance the performance and efficiency of multiprocessor systems, leading to improved resource utilization and overall system productivity.

## References


[1] Vladimir Sahakyan, Artur Vardanyan, "The Queue State for Multiprocessor System with Waiting Time Restriction", *Computer Science and Information Technologies 2019, Conference Proceeding*, Yerevan, pp. 116–119, 2019. DOI: https://doi.org/10.1109/CSITechnol.2019.8895093

[2] Vladimir Sahakyan, Artur Vardanyan, "About the possibility of executing tasks with a waiting time restriction in a multiprocessor system", *AIP Conference Proceedings*, vol. 2757, pp. 030003, 2023. DOI: https://doi.org/10.1063/5.0135784

[3] Vladimir Sahakyan, Artur Vardanyan, "The Queue Distribution in Multiprocessor Systems with the Waiting Time Restriction", *Mathematical Problems of Computer Science*, Yerevan, vol. 51, pp. 82–89, 2019. DOI: https://doi.org/10.51408/1963-0035

[4] Vladimir Sahakyan, Artur Vardanyan, "A Computational Approach for Evaluating Steady-State Probabilities of a Multiprocessor Queueing System with a Waiting Time Restriction", *Computer Science and Information Technologies 2023, Conference Proceeding*, Yerevan, 2023.