# On the Busy Cycle Maxima in a Heterogeneous Fork-Join Queue

Pierre M. Fiorini
CF Search Marketing
North Hampton, NH
pierre@cfsearchmarketing.com

## ABSTRACT

This work investigates the distribution of the maximum response time during a busy cycle in a Fork-Join queueing system with two parallel servers having heterogeneous deterministic service times and Poisson arrivals. We formulate a renewal-type Fredholm integral equation of the second kind that explicitly characterizes the cumulative distribution function (CDF) of the maximum response time. An infinite series representation is derived from this integral equation, expressing the probability distribution in terms of recursively defined integrals, each corresponding to a fixed number of customer arrivals during the busy cycle. A computationally efficient numerical scheme, using the infinite series, is developed to accurately approximate this distribution. We validate the accuracy and efficiency of our numerical method through comprehensive simulations, demonstrating strong agreement between analytical and empirical results. Our analysis provides clear insights into the behavior of heterogeneous Fork-Join queue dynamics, facilitating more precise design and performance evaluation of parallel processing and service systems.

## 1. INTRODUCTION

Fork-Join queues are critical for modeling parallel systems in computing, manufacturing, telecommunications, and cloud infrastructure. In these models, incoming jobs are split on arrival for service by multiple servers and are joined before departure. Despite their importance, few analytical results exist for fork–join queues, but various approximations are known. In this work, we analyze a two-node Fork-Join queue with heterogeneous deterministic service times and Poisson arrivals, focusing specifically on the maximum response time during a busy period. Our key contributions are (i) deriving a renewal-type Fredholm integral equation explicitly describing the cumulative distribution function (CDF) for this metric; (ii) providing an infinite-series representation revealing the recursive structure of the solution; and (iii) developing an efficient numerical algorithm to solve this equation. Monte Carlo simulations validate our approach, demonstrating its accuracy and practical utility.

## 2. RELATED WORK

Analytical results for fork–join queues remain limited due to inherent complexity arising from the synchronization of parallel tasks. However, notable analytical achievements include exact steady-state distributions for the two-server exponential service time case - known as the *Flatto-Hahn-Wright* model - obtained through uniformization methods and generating function techniques [6]. Additionally, product-form solutions exist for deterministic arrival processes, simplifying the queue-length analysis significantly [9]. Nelson and Tantawi derived exact analytical expressions for the average response times in two-server exponential fork-join systems, highlighting the intricate interplay between synchronization delays and service distributions [8]. Baccelli and Makowski established rigorous, computable bounds on average response times and higher-order moments in the general M/G/1 scenario, laying foundational analytical tools applicable to broader fork–join systems [4]. Recent analytical approximations and expansions have extended these classical results by addressing the more challenging setting of heterogeneous fork–join queues. For example, Alomari and Menasce introduced harmonic-number-based approximations to efficiently estimate response times in heterogeneous multiclass fork-join networks [1]. Kemper and Mandjes developed accurate response-time approximations specifically tailored to heterogeneous two-server systems [7], while Qiu, Pérez, and Harrison presented analytical approximations leveraging phase-type distributions to capture response-time tails effectively, thereby providing deeper insights into extreme performance metrics in complex fork–join networks [10].

Takács' classical results on cycle maxima in M/G/1 queues established fundamental analytical relationships between the steady-state waiting-time distribution and the distribution of the maximum waiting time observed during a busy period [2, 11]. As detailed in Asmussen [2], these results provide rigorous probabilistic identities based on Markovian arguments linking busy-period maxima explicitly to steady-state quantities. While Takács' results focus specifically on classical single-server queues, they lay important theoretical groundwork by analytically characterizing busy-period maxima through steady-state relationships [2, 11]. In contrast, the approach developed in this paper presents a formulation based on Fredholm integral equations of renewal type to explicitly compute the maximum response-time distribution in Fork-Join queueing systems. This distinction highlights our contribution in extending analytical frameworks towards more complex, parallel queueing models.

The use of Fredholm integral equations in this work provide a mathematically rigorous and explicit characterization of complex renewal-type dependencies inherent in busy pe-

riods of queueing systems. Using them in this paper allows one to systematically describe and compute the cumulative distribution function (CDF) of the maximum response time by capturing the recursive interplay between arrivals, server workloads, and synchronization. Moreover, Fredholm integral formulations enable efficient numerical solutions and facilitate precise analytical approximations [3], essential for analyzing performance extremes in fork-join parallel systems in general.

# 3. MAIN RESULT

In this section, we derive and prove a Fredholm-type renewal integral equation characterizing the cumulative distribution function (CDF) of the maximum response time during a busy period of a two-node Fork-Join queue with heterogeneous deterministic service times and Poisson arrivals.

## 3.1 Preliminaries and Definitions

Consider a two-node Fork-Join queue with jobs arriving according to a Poisson process at rate $\lambda > 0$. Each arriving job splits immediately into two parallel tasks, each processed independently by one of two dedicated servers. The deterministic service times at the two servers are denoted by $d_1 > 0$ and $d_2 > 0$, respectively. The response time of each job is the maximum completion time among its two parallel tasks.

Let $w_1 \geq 0$ and $w_2 \geq 0$ represent the initial workloads of the two servers at the start of the busy period (time $t = 0$). The maximum response time during the busy period, starting with these workloads, is defined as follows:

**Definition 1** (Maximum Response Time)**.** *Given initial workloads $(w_1, w_2)$ at time $t = 0$, define the maximum response time during the busy period as:*

$$R_{\max}(w_1, w_2) \triangleq \sup_{0 \leq t \leq T} R(t),$$

*where $R(t)$ denotes the response time of a job arriving at time $t$, and $T$ is the (random) length of the busy period, ending when both servers become simultaneously idle. The cumulative distribution function (CDF) conditional on $(w_1, w_2)$ is:*

$$F_R(x \mid w_1, w_2) = \Pr\big(R_{\max}(w_1, w_2) \leq x\big), \quad x \geq 0.$$

From this definition, we note two elementary properties immediately:

- If $x < \max(w_1 + d_1, w_2 + d_2)$, then trivially $F_R(x \mid w_1, w_2) = 0$, since initial workloads alone yield a response time exceeding $x$.
- If no arrivals occur during $[0, \max(w_1 + d_1, w_2 + d_2)]$, the busy period ends exactly at $\max(w_1 + d_1, w_2 + d_2)$.

## 3.2 Main Theorem: Integral Equation

We formally state our main result, connecting the distribution $F_R(x \mid w_1, w_2)$ to a Fredholm-Renewal integral equation:

**Theorem 1** (Integral Equation for Heterogeneous Fork-Join Queue)**.** *For the two-node heterogeneous Fork-Join queue described above, the CDF of the maximum response time $F_R(x \mid w_1, w_2)$ during a busy cycle satisfies the integral equation:*

$$F_R(x \mid w_1, w_2) = \begin{cases} 0 & x < r(w_1, w_2) \\ e^{-\lambda r(w_1, w_2)} & x \geq r(w_1, w_2) \\ + \displaystyle\int_0^{r(w_1, w_2)} \lambda e^{-\lambda t} & \\ \quad \times F_R\big(x \mid [w_1 + d_1 - t]^+, [w_2 + d_2 - t]^+\big)\, dt, & \end{cases} \tag{1}$$

*where $r(w_1, w_2) \triangleq \max(w_1 + d_1, w_2 + d_2)$ and $[y]^+ \triangleq \max(y, 0)$.*

Proof. We proceed by conditioning on the occurrence of arrivals within the interval $[0, r(w_1, w_2)]$.

**Case 1: No Arrivals.** Let $A_0$ denote the event of no arrivals within $[0, r(w_1, w_2)]$. By the Poisson property:

$$\Pr(A_0) = e^{-\lambda r(w_1, w_2)}.$$

If $A_0$ occurs, the busy period ends exactly at $r(w_1, w_2)$, thus:

$$F_R(x \mid w_1, w_2)|_{A_0} = e^{-\lambda r(w_1, w_2)} \mathbf{1}\{r(w_1, w_2) \leq x\}.$$

**Case 2: At Least One Arrival.** Let event $A_1$ be the complement of $A_0$. Define $T$ as the first arrival time with density:

$$f_T(t) = \lambda e^{-\lambda t}, \quad 0 < t \leq r(w_1, w_2).$$

After an arrival at time $t$, workloads reset to:

$$(w_1', w_2') = \big([w_1 + d_1 - t]^+, [w_2 + d_2 - t]^+\big).$$

By the Markov property, from time $t$ onwards, the distribution of maximum response time matches that starting at $(w_1', w_2')$:

$$\Pr(R_{\max}(w_1, w_2) \leq x, A_1) = \int_0^{r(w_1, w_2)} \lambda e^{-\lambda t} F_R(x \mid w_1', w_2') dt.$$

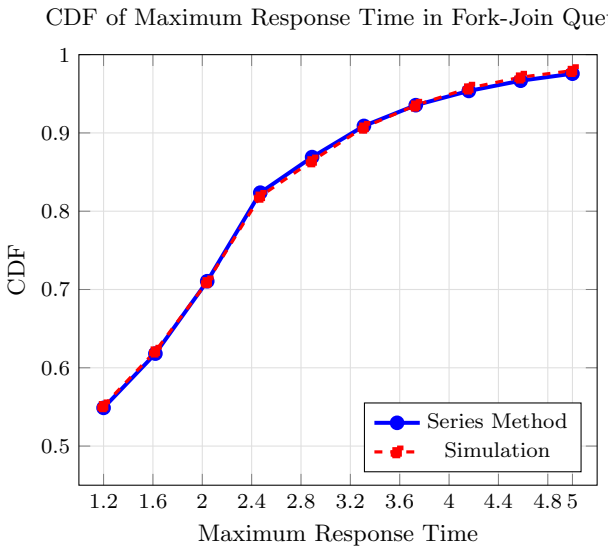
**Combine the cases.** Combining Cases 1 and 2 yields the integral equation (1). Existence and uniqueness follow from standard Fredholm integral equation theory due to the positivity and continuity of the kernel.

This completes the proof.

**Remark 1** (Independent M/D/1 Assumption)**.** *It may seem tempting to simplify the analysis of the heterogeneous M/D/1 Fork-Join queue ($d_1 \neq d_2$) by focusing only on the server with the larger service time. For the first job starting from an empty system ($W_1 = 0, W_2 = 0$), this simplification works since its response time is simply $R^{(1)} = \max(d_1, d_2)$. However, this approach implicitly implies treating the slower server as an isolated M/D/1 queue and thus incorrectly assuming independence. This is problematic because, over time, the workloads $W_1$ and $W_2$ become inherently correlated. The correct evaluation of the response time $R = \max(W_1 + d_1, W_2 + d_2)$ can only be achieved by analyzing the joint workload distribution $(W_1, W_2)$, which accounts for this correlation.*

## 3.3 Infinite Series Representation and Numerical Procedure

We define the infinite series representation for the cumulative distribution function (CDF) $F_R(x \mid w_1, w_2)$ of the maximum response time, obtained by recursively applying the integral equation (1):




| Response Time | Series CDF | Simulation CDF | Abs. Diff. |
|---|---|---|---|
| 1.20 | 0.548812 | 0.551400 | 0.002588 |
| 1.62 | 0.618111 | 0.621200 | 0.003089 |
| 2.04 | 0.710573 | 0.709800 | 0.000773 |
| 2.47 | 0.823691 | 0.819000 | 0.004691 |
| 2.89 | 0.869121 | 0.864200 | 0.004921 |
| 3.31 | 0.908944 | 0.907200 | 0.001744 |
| 3.73 | 0.935459 | 0.935400 | 0.000059 |
| 4.16 | 0.953641 | 0.957200 | 0.003559 |
| 4.58 | 0.966696 | 0.971200 | 0.004504 |
| 5.00 | 0.975671 | 0.979600 | 0.003929 |

Parameters: $\lambda = 0.5$, $d_1 = 1.0$, $d_2 = 1.2$

**Figure 1: CDF of Maximum Response Time in a Fork-Join Queue with heterogeneous deterministic service times. The comparison shows excellent agreement between the analytical series method and Monte Carlo simulation results, with differences of less than 0.5% when $\lambda = 0.5$, and mean service times $d_1 = 1.0$, $d_2 = 1.2$**

**Definition 2** (Infinite Series Representation)**.** *The CDF of the maximum response time is given by the infinite series:*

$$F_R(x \mid w_1, w_2) = \sum_{k=0}^{\infty} f_k(x \mid w_1, w_2),$$

*where each term $f_k(x \mid w_1, w_2)$ explicitly represents exactly $k$ arrivals during the busy period. These terms are defined recursively as follows:*

- ***Case*** $k = 0$ *(no arrivals occur):*

$$f_0(x \mid w_1, w_2) = e^{-\lambda r(w_1, w_2)} \mathbf{1}\{r(w_1, w_2) \le x\},$$

*where $r(w_1, w_2) = \max(w_1 + d_1, w_2 + d_2)$ denotes the maximum response time if no arrivals occur.*

- ***Case*** $k \ge 1$ *(exactly $k$ arrivals occur):*

$$f_k(x \mid w_1, w_2) = \int_0^{r(w_1, w_2)} \lambda e^{-\lambda t} \times f_{k-1}\big(x \mid [w_1 + d_1 - t]^+, [w_2 + d_2 - t]^+\big)\, dt.$$

The initial term $f_0(x \mid w_1, w_2)$ represents the probability that no arrivals occur during $[0, r(w_1, w_2)]$. Each subsequent term $f_k(x \mid w_1, w_2)$ captures the probability of exactly $k$ arrivals within the busy cycle. The process resets recursively with updated workloads at each arrival, reflecting the integral structure of the series. The numerical procedure computes the CDF of the maximum response time during a busy cycle, $F_R(x|w_1, w_2)$, in this exact manner.

## 4. CONCLUSION

In this paper, we derived and solved a renewal-type Fredholm integral equation characterizing the maximum response-time distribution in a heterogeneous two-server Fork-Join queue with Poisson arrivals. The numerical solution, validated by simulations, demonstrates excellent accuracy. Our approach provides analytical insight into performance extremes, aiding the evaluation and design of parallel systems.

## 5. REFERENCES


[1] F. Alomari and D. A. Menasce, "Efficient response time approximations for multiclass fork and join queues in open and closed queuing networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 6, pp. 1437–1446, 2013.

[2] S. Asmussen, *Applied Probability and Queues*, 2nd ed. Springer-Verlag, New York, 2003.

[3] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 1997.

[4] F. Baccelli and A. Makowski, "Simple computable bounds for the fork-join queue," *INRIA Technical Report RR-0394*, 1985.

[5] O. Boxma, G. Koole, and Z. Liu, "Queueing-theoretic solution methods for models of parallel and distributed systems," *CWI Technical Report BS-R9425*, 1996.

[6] L. Flatto and S. Hahn, "Two parallel queues created by arrivals with two demands I," *SIAM Journal on Applied Mathematics*, vol. 44, no. 5, pp. 1041–1053, 1984.

[7] B. Kemper and M. Mandjes, "Mean sojourn times in two-queue fork-join systems: Bounds and approximations," *OR Spectrum*, vol. 34, no. 3, pp. 723–742, 2011.

[8] R. Nelson and A. N. Tantawi, "Approximate analysis of fork/join synchronization in parallel queues," *IEEE Transactions on Computers*, vol. 37, no. 6, pp. 739–743, 1988.

[9] D. Pinotsi and M. A. Zazanis, "Synchronized queues with deterministic arrivals," *Operations Research Letters*, vol. 33, no. 6, pp. 560–566, 2005.

[10] Z. Qiu, J. F. Pérez, and P. G. Harrison, "Beyond the mean in fork-join queues: Efficient approximation for response-time tails," *Performance Evaluation*, vol. 91, pp. 99–116, 2015.

[11] L. Takács, *Combinatorial Methods in the Theory of Stochastic Processes*. Wiley, New York, 1967.