TECHNICAL
RESEARCH
REPORT

SYSTEMS
RESEARCH
CENTER

*Supported by the
National Science Foundation
Engineering Research Center
Program (NSFD CD 8803012),
Industry and the University*

Interpolation Approximations for Symmetric Fork-Join Queues

by S. Varma and A.M. Makowski

TR 92-122r1

INTERPOLATION APPROXIMATIONS FOR SYMMETRIC FORK-JOIN QUEUES

Subir Varma[a,*] and Armand M. Makowski[b,†]

[a] IBM Corporation, A37/503, P.O. Box 12195, Research Triangle Park, NC 27709, U.S.A.

[b] Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, MD 20742, U.S.A.

ABSTRACT

In this paper we propose a family of heuristic approximations for the expected response time of K–dimensional symmetric Fork–Join systems in statistical equilibrium with general inter–arrival and service time distributions. To do this, we rely on the light traffic interpolation technique popularized by Reiman and Simon. Our starting point is a formula for the heavy traffic limit of two–dimensional Fork–Join queues that was obtained by the authors in [17,19]. By observing a fortuitous agreement between the light traffic derivative and the heavy traffic limit for this system under Markovian assumptions, we are able to obtain an approximation to the heavy traffic limit for K–dimensional systems with general inter–arrival and service distributions. By combining this heavy traffic limit with light traffic limits, we generate interpolation approximations for the Fork–Join queue, which agree extremely well with simulation results.

I. INTRODUCTION

Although synchronization constraints are inherent to the operation of many computer, communication and production systems, their impact on system performance is far from being well understood. This may be partially attributed to a penury of models which meaningfully incorporate the synchronization constraints of interest, and which are nevertheless tractable. On the other hand, many interesting applications involving synchronization are concerned with problems of resource sharing, and can be adequately described in terms of queueing models which have traditionally provided quantitative insights into system performance. Unfortunately, the incorporation of synchronization constraints into a queueing model often destroys important properties, such as the product form and insensitivity properties, which have fueled the development of the various performance methodologies for data networks and earlier computer systems.

[*] The work of this author was supported by a fellowship from the Systems Research Center, University of Maryland, through the NSF Engineering Research Centers Program under Grant CDR-88-03012.

[†] The work of this author was supported partially through NSF Grants ECS-83-51836 and CDR-88-03012, and partially through a grant from AT&T Bell Laboratories.

This state of affairs is already well apparent for the so–called Fork–Join queue, which is perhaps one of the simplest queueing systems with a synchronization constraint. The symmetric Fork–Join queue, which has been proposed as a queueing model for parallel processing [10], is composed of K (≥ 2) identical servers operating in *parallel*, each with an infinite waiting room. Jobs that arrive to the system are assumed to consist of exactly K tasks, the service requirements of the tasks being independent and identically distributed (i.i.d.). Upon arrival, a job is instantaneously decomposed into its K constituent tasks, with the k^{th} task routed to the k^{th} queue where it is served in FCFS order. As soon as a task completes service, it is put into a synchronization buffer, and a job leaves the system only when all of its constituent tasks have completed service.

Under most model assumptions, exact analysis of the K–dimensional Fork–Join queue appears quite difficult, if not intractable. This is due to the fact that although each server with its attending buffer can be interpreted as a single server queue, these K $GI/GI/1$ systems are usually ***not*** independent as they are coupled through a common arrival stream. Notwithstanding this difficulty, results are available in some special cases when $K = 2$: For Poisson arrivals and exponentially distributed service times, Flatto and Hahn [6] determined the stationary joint distribution of the number of customers in the queues. Additional results were derived in [1] by Baccelli for Poisson arrivals and more general service time distributions. However, in both these references, the results were obtained in a form which does not lend itself in a straighforward manner to parametric performance studies. To remedy this shortcoming, Nelson and Tantawi set out in [10] to provide an approximate analysis of the expected job response time in steady state when all K servers are identical, under the Markovian assumptions of [6]. Their "scaling approximation" was obtained by cleverly combining light traffic information with extensive simulation work to obtain the numerical value for some of the constants which appeared in the postulated approximations.

In this paper we revisit the problem of designing approximations to the expected job response time in *symmetric* Fork–Join queues, i.e., Fork–Join queues where all the servers are identical. Our objective is to derive approximations which all flow from the same "paradigm" without recourse to experimental work as in [10], and which hold under circumstances far more general than the Markovian setup. We achieve this by applying the ***light traffic interpolation*** technique popularized by Reiman and Simon [12–14]. Its basic idea is as follows: While analytical results are typically very hard to come by for the Fork–Join queue, it is, however, possible to obtain ***asymptotic*** results in light and heavy traffic regimes. The light traffic interpolation approximation arises then by suitably interpolating light and heavy traffic regimes. Light traffic refers to the situation where the system is lightly loaded, i.e., the arrival rate λ to the system is very small, in which case it is reasonable to seek a Taylor expansion of the performance measure (as a function of λ) near $\lambda = 0$. This, of course, passes by the computation of derivatives with respect to λ at $\lambda = 0$, and general methods have been proposed in the literature for doing just that [12–14]. On the other hand, heavy traffic deals with the situation where the system operates near capacity; the corresponding behavior of the performance measures is often studied by means of diffusion limits associated with a rescaled version of the "process" of interest [9,20]. When the heavy traffic limit is not available in closed form, we resort to *heuristics* to generate an appropriate estimate for it as we explain below.

With this technique, we have been able to obtain simple yet good approximations to the expected job response time in symmetric Fork–Join queues under a wide set of assumptions on the statistics of inter-arrival and service times, and this for an arbitrary number of servers. The proposed light traffic interpolation technique provides an economical way to obtain "ballpark" figures for the quantities of interest. As an encouraging sign of the potential of this approach, the interpolation approximations have agreed extremely well with simulation results, even in moderate traffic regime.

Of course the approach is not restricted only to Fork–Join queues. It has been applied with success in other situations which occur in a wide variety of important applications: Earlier examples without synchronization constraints are contained in the references [4,5,7,8,15]. More recently, Varma [18] has used interpolation approximations to analyze the performance of a time–stamp ordering algorithm for distributed databases; additional queueing systems with synchronization constraints, including both resequencing and Fork–Join constructs, are discussed in [17]. Tedijanto has applied the technique to polling systems in [16].

We now close this introduction with a brief survey of the paper's organization and contents: The model is described in Section II. We summarize in Section III the needed elements of light and heavy traffic theories, and how they can be combined to yield the light traffic interpolations. Central to heavy traffic is the parameter β defined by (3.20). The heavy traffic limit for K–dimensional Fork–Join queues can be obtained easily for $\beta = 0$ (deterministic arrivals) and $\beta = 1$ (deterministic services). For $\beta = \frac{1}{2}$, when $K = 2$, we were able to find closed form expressions for the heavy traffic limit by solving an appropriate second order elliptic PDE with oblique boundary conditions [17,19]. As the case of Poisson arrivals and exponential services corresponds to $\beta = \frac{1}{2}$, we are thus able to produce a light traffic interpolation for two server systems in the Markovian case. Unfortunately, in order to obtain formulae for the heavy traffic limit when there are more than two queues, even if $\beta = \frac{1}{2}$, it is necessary to solve more complicated PDEs in K dimensions, a most difficult task to say the least. Instead of solving PDEs, we present here a number of heuristic approximations for the heavy traffic limit, which exploit the exact heavy traffic result obtained for $K = 2$ in [17,19]. This is done for the Markovian case in Section IV where we propose an estimate to the heavy traffic limit by extending to arbitrary K the range of validity of a fortuitous relationship between heavy traffic limit and light traffic derivative observed for $K = 1$ and $K = 2$. In Section V, we take the postulated relationship one step further to generate the heavy traffic estimate for the case $\beta = \frac{1}{2}$ without Markovian assumptions. We can now attack the problem of designing interpolations for β in the full range $[0, 1]$. We do this in Section VI as follows: Knowing the heavy traffic limit for $\beta = 0$, $\frac{1}{2}$ and 1, we carry out a quadratic interpolation to extend it to the entire range $[0, 1]$. This approximation to the heavy traffic limit is combined with light traffic limits to obtain interpolation approximations for a number of different choices of the arrival and service distributions, and in each case good agreement with simulation results is observed. For the sake of completeness, we have relegated to several appendices some intermediary, and often tedious, calculations.

II. THE MODEL

We now present the queueing model of interest in this paper, together with the notation and some of the basic assumptions enforced throughout: For any probability distribution function F on $I\!R$, we denote by $m(F)$ and $\mathrm{var}(F)$ its mean and variance, respectively, whenever these quantities exist. A positive integer K is given and held fixed hereafter. We start with the square–integrable $I\!R_+$–valued rvs $\{\tau_{n+1},\ n = 0, 1, \ldots\}$ and $\{\sigma_n^k,\ k = 1, \ldots, K;\ n = 0, 1, \ldots\}$, which are all defined on some underlying probability triple $(\Omega, \mathcal{F}, \mathbf{P})$.

These quantities can be given the following interpretation in the context of a K-dimensional Fork–Join queue: Such a queueing system is composed of K identical servers working in parallel. Each one of these servers has its own buffer of infinite capacity and operates according to the FIFO discipline. Jobs arrive in the system at time epochs $\{A_n,\ n = 0, 1, \ldots\}$ defined by

$$A_0 \equiv 0, \quad A_n \equiv \sum_{m=0}^{n-1} \tau_{m+1}. \qquad n = 1, 2, \ldots (2.1)$$

In other words, the interarrival time between the n^{th} and the $(n+1)^{rst}$ jobs is given by τ_{n+1} with the convention that the 0^{th} job arrives at time $t = 0$. The n^{th} job consists of K tasks; the execution of the k^{th} task from the n^{th} job requires σ_n^k units of time, $k = 1, \ldots, K$. Upon arrival into the Fork–Join system, the n^{th} job is instantaneously decomposed into its K constituent tasks and the k^{th} task is routed to the k^{th} queue where it is served in FCFS order, requesting service for σ_n^k units of time. As soon as a task completes service, it is put into a synchronization buffer, and a job leaves the system when all of its constituent tasks have completed service. The 0^{th} job finds an initial load already awaiting service in the various buffer areas, with the rv W^k representing the amount of time required by the k^{th} server to clear this initial load from its buffer.

We now define the performance measure of interest for this Fork–Join queue system: We generate the $I\!R_+^K$–valued rvs $\{(W_n^1, \ldots, W_n^K),\ n = 0, 1, \ldots\}$ componentwise by the Lindley recursions

$$W_0^k = W^k, \quad W_{n+1}^k = \left[W_n^k + \sigma_n^k - \tau_{n+1}\right]^+, \quad k = 1, \ldots, K \qquad n = 0, 1, \ldots (2.2)$$

where W_n^k represents the waiting time of the k^{th} task from the n^{th} job. The corresponding response time R_n^k (through the k^{th} channel) is thus

$$R_n^k \equiv W_n^k + \sigma_n^k, \quad k = 1, \ldots, K. \qquad n = 0, 1, \ldots (2.3)$$

The system response time T_n of the n^{th} job is then given by

$$T_n \equiv \max_{1 \leq k \leq K} R_n^k. \qquad n = 0, 1, \ldots (2.4)$$

Throughout this discussion, we enforce the following renewal assumptions **(R1)**–**(R3)**, where

(R1): The sequences $\{\tau_{n+1},\ n = 0, 1, \ldots\}$, and $\{\sigma_n^k,\ k = 1, \ldots, K;\ n = 0, 1, \ldots\}$ are mutually independent;

(R2): The $\mathbb{R}_+$–valued rvs $\{\tau_{n+1},\ n = 0, 1, \ldots\}$ form an i.i.d. sequence with common distribution A; and

(R3): The $\mathbb{R}_+$–valued rvs $\{\sigma_n^k,\ k = 1, \ldots, K,\ n = 0, 1, \ldots\}$ form an *i.i.d.* sequence with common distribution B.

It is well known [2,3] that under the renewal assumptions **(R1)–(R3)**, the K–dimensional Fork–Join queue system reaches statistical equilibrium if and only if $m(B) < m(A)$. This stability condition does not depend on K, the number of processors, and can be rewritten in the form

$$\rho \equiv \frac{m(B)}{m(A)} = \frac{\lambda}{\mu} < 1 \tag{2.5}$$

where we have used the usual notation $\lambda \equiv m(A)^{-1}$ and $\mu \equiv m(B)^{-1}$. We also find it convenient in many places to use the notation σ^2 to denote the variance $\mathrm{var}(B)$ of the service time distribution B.

Under (2.5) the system is termed stable, in which case the sequences of waiting times and response times all have stationary versions [2,3]. In particular, let $T_K(\lambda)$ denote the stationary system response time for the K–dimensional Fork–Join queue in statistical equilibrium when the arrival rate is λ $(0 < \lambda < \mu)$. We are concerned with the evaluation of $\overline{T}_K(\lambda) \equiv \mathbf{E}[T_K(\lambda)]$, the average response time in statistical equilibrium for the K–dimensional Fork–Join queue when the arrival rate is λ. This is a notoriously difficult endeavor [1,6] as should be apparent from (2.2)–(2.4).

III. LIGHT TRAFFIC INTERPOLATIONS – A SUMMARY

In this paper we combine light traffic information with heavy traffic limits to derive approximations to the performance measure $\overline{T}_K(\lambda)$. We now briefly describe the salient features of this approach as used here.

III.1. Light traffic theory: Light traffic theory is concerned with the evaluation of the performance measure for small values of λ, a task best accomplished by considering the Taylor series expansion of $\overline{T}_K(\lambda)$ at $\lambda = 0$. Assuming the existence of the n first derivatives of $\overline{T}_K(\lambda)$ at $\lambda = 0$, we have

$$\overline{T}_K(\lambda) \simeq \overline{T}_K^{(0)}(0) + \lambda \overline{T}_K^{(1)}(0) \ldots + \frac{\lambda^n}{n!} \overline{T}_K^{(n)}(0), \quad \lambda \simeq 0 \tag{3.1}$$

where $\overline{T}_K^{(m)}(0)$ denotes the derivative of order m at $\lambda = 0$, $m = 0, 1, \ldots$; by convention, we set $\overline{T}_K^{(0)}(0) = \overline{T}_K(0)$ and often write $\overline{T}'_K(0)$ for the first derivative $\overline{T}_K^{(1)}(0)$. We call $\overline{T}_K^{(m)}(0)$ the light traffic derivative of order m.

To this date, a fairly large body of literature exists on the light traffic of queueing systems [12–14] with the bulk of the efforts devoted to finding conditions for the existence of light traffic derivatives and to providing methods to evaluate them. Most of the results are concerned with systems with Poisson arrival processes, but can be extended to more general arrival processes which are "driven" by a Poisson process, e.g., phase–type renewal processes, non–stationary Poisson processes and Markov–modulated Poisson processes to name a few [12–14]. Here, as we confine the discussion to light traffic derivatives of order $m = 0, 1$, we follow mostly the viewpoint developed by Reiman and Simon in [12–14] for Poisson arrival processes. The relevant facts of the theory are summarized below for easy reference in the context of the stable K–dimensional Fork–Join queue with Poisson arrivals of rate λ and general service time distribution B. In the interest of brevity, we omit discussing the relevant technical conditions; details are available in [12–14,17]. It should be noted however that all situations considered in this paper satisfy the appropriate conditions.

As we are interested in statistical equilibrium, we can think of the system as having been operated from $t = -\infty$ onward, so that $T_K(\lambda)$ can be interpreted as the response time of a tagged job entering the system at $t = 0$. This tagged job has service times $v_1, \ldots, v_K$, which are i.i.d. rvs with common distribution B and which are independent of all other rvs. Following Reiman and Simon [12–14], we then define

$$\overline{\psi}(\emptyset) \equiv \mathbf{E}[T_K(\lambda) | \text{ No arrivals on } (-\infty, \infty)] \tag{3.2}$$

and

$$\overline{\psi}(\{t\}) \equiv \mathbf{E}\left[T_K(\lambda) | \begin{matrix}\text{There is exactly one arrival at time } t \\ \text{on } (-\infty,\infty)\end{matrix}\right], \quad t \in I\!R. \tag{3.3}$$

In [12–14], it is shown that the quantities (3.2)–(3.3) are independent of λ, and that

$$\overline{T}_K(0) = \overline{\psi}(\emptyset) \tag{3.4}$$

and

$$\overline{T}_K^{(1)}(0) = \int_{-\infty}^{\infty} \left(\overline{\psi}(\{t\}) - \overline{\psi}(\emptyset)\right) dt. \tag{3.5}$$

We now indicate how (3.2)–(3.5) pave the way for an evaluation of the light traffic quantities $\overline{T}_K(0)$ and $\overline{T}_K^{(1)}(0)$.

The definition (3.2) suggests the following scenario for computing $\overline{\psi}(\emptyset)$: The tagged job enters the system given that no other job will enter the system on $(-\infty, \infty)$. Since the tagged job arriving at $t = 0$ does not experience interference from any other job, its queueing delay will be zero, and its response time is simply the maximum of the K i.i.d. rvs $v_1, \ldots, v_k$ with common distribution B. As a result, we have

$$\overline{\psi}(\emptyset) = \mathbf{E}[\max\{v_1, \ldots, v_K\}] \tag{3.6}$$

so that

$$\overline{T}_K(0) = \int_0^{\infty} \left(1 - B(x)^K\right) dx. \tag{3.7}$$

Note that (3.6)–(3.7) also yield $\overline{T}_K(0)$ when the arrival process is *not* Poisson.

To evaluate $\overline{T}_K^{(1)}(0)$, we need to compute $\overline{\psi}(\{t\})$ for all t in $I\!R$. To do so, we introduce the rv $T_K(t, s_1, \ldots, s_K, v_1, \ldots, v_K)$ to represent the response time of the tagged job that enters the system at time $t = 0$ with service times $v_1, \ldots, v_K$, given that another job arrives at time t (in $I\!R$) with service times $s_1, \ldots, s_K$. Under the foregoing assumptions, the rvs $v_1, \ldots, v_K, s_1, \ldots, s_K$ are i.i.d. rvs with common distribution B, and it is clear that

$$\begin{aligned} &T_K(t, s_1, \ldots, s_K, v_1, \ldots, v_K) \\ &= \begin{cases} \max\{v_1, \ldots, v_K\}, & \text{if } t \geq 0 \\ \\ \max\{v_1 + (s_1 + t)^+, \ldots, v_K + (s_K + t)^+\} & \text{if } t < 0, \end{cases} \end{aligned} \tag{3.8}$$

and from the definition (3.3) we readily get

$$\overline{\psi}(\{t\}) = \mathbf{E}\left[T_K(t, s_1, \ldots, s_K, v_1, \ldots, v_K)\right], \quad t \in I\!R. \tag{3.9}$$

Since

$$\overline{\psi}(\{t\}) = \mathbf{E}[\max\{v_1, \ldots, v_K\}] = \overline{\psi}(\emptyset), \quad t \geq 0, \tag{3.10}$$

it is plain that (3.5) now reduces to

$$\overline{T}'_K(0) = \int_{-\infty}^{0} \left(\overline{\psi}(\{t\}) - \overline{\psi}(\emptyset)\right) dt. \tag{3.11}$$

For $t < 0$, we define the rvs $Y_1, \ldots, Y_K$ by

$$Y_k = v_k + (s_k + t)^+, \quad k = 1, \ldots, K. \tag{3.12}$$

The rvs $Y_1, \ldots, Y_K$ are i.i.d. with common distribution F_Y which can be easily derived in terms of B as will be done later in specific instances. With the representation

$$T_K(t, s_1, \ldots, s_K, v_1, \ldots, v_K) = \max\{Y_1, \ldots, Y_K\}, \quad t < 0 \tag{3.13}$$

as a starting point, we use the independence of the rvs $Y_1, \ldots, Y_K$ to obtain

$$\overline{\psi}(\{t\}) = \int_0^\infty \left(1 - F_Y(x)^K\right) dx, \quad t < 0. \tag{3.14}$$

III.2. Heavy traffic theory: Heavy traffic refers to the situation where λ increases to its critical value μ, in which case $\overline{T}_K(\lambda)$ grows unbounded and attention shifts to finding a function $\alpha_K : [0, \mu] \to I\!R_+$ and a non–zero constant C_K such that

$$\lim_{\lambda \uparrow \mu} \frac{\overline{T}_K(\lambda)}{\alpha_K(\lambda)} = C_K. \tag{3.15a}$$

We refer to C_K as the heavy traffic limit. If α_K and C_K are "easily" computable, then (3.15) provides a means to approximate $\overline{T}_K(\lambda)$ in heavy traffic, since then

$$\overline{T}_K(\lambda) \simeq \alpha_K(\lambda) C_K, \quad \lambda \simeq \mu. \tag{3.15b}$$

Heavy traffic limits for the Fork–Join queue have been investigated in [11,17,19]. To describe the results, we consider a family of *stable* Fork–Join queues indexed by a parameter $r = 1, 2, \ldots$; the r^{th} system has interarrival time distribution $A(r)$ and service time distribution B. For technical reasons, we assume that for some $\epsilon > 0$, B has a finite moment of order $2 + \epsilon$ and

$$\sup_{r \geq 1} \int_0^\infty t^{2+\epsilon} dA(r)(t) < \infty. \tag{3.16}$$

This family of stable Fork–Join queues is chosen such that as $r \uparrow \infty$, the systems become increasingly less stable, i.e., $\lim_{r \uparrow \infty} \lambda(r) = \mu$ where $\lambda(r) \equiv m(A(r))^{-1}$, $r = 1, 2, \ldots$. This trend to heavy traffic is achieved under the following conditions

$$\lim_{r \uparrow \infty} \sqrt{r}\,(m(A(r)) - m(B)) = \gamma \tag{3.17}$$

and

$$\lim_{r \uparrow \infty} \mathrm{var}(A(r)) = \sigma_0^2 \tag{3.18}$$

for constants $\gamma > 0$ and $\sigma_0^2 \geq 0$.

In [17,19], we showed the existence of the heavy traffic limit (3.15) in the form

$$\lim_{r \uparrow \infty} (\mu - \lambda(r))\overline{T}_K(\lambda(r)) = C_K \tag{3.19}$$

with C_K expressed as the first moment of a fairly complicated functional on a standard K–dimensional Brownian motion. In general no closed–form expression seems available for C_K, except in some special cases which we now briefly describe: We define

$$\beta \equiv \frac{\sigma_0^2}{\sigma_0^2 + \mathrm{var}(B)}. \tag{3.20}$$

For $K = 2$ and $\beta = \frac{1}{2}$, we were able to solve the basic PDE satisfied by the stationary distribution of the diffusion limit of the end–to–end delay [17,19], and showed that

$$\lim_{r \uparrow \infty} (\mu - \lambda(r))\overline{T}_2(\lambda(r)) = \frac{11}{8} \frac{\mathrm{var}(B)}{m(B)^2}, \tag{3.21}$$

so that $C_2 = \frac{11}{8} \frac{\mathrm{var}(B)}{m(B)^2}$. For arbitrary K, the exact value of C_K is not known, even for $\beta = \frac{1}{2}$. Only in the extreme cases $\beta = 0$ and $\beta = 1$ is the heavy traffic limit known; we state the results without proof and refer the reader to [9,17,19] for additional details. Roughly speaking, the value $\beta = 0$ corresponds to the situation with deterministic arrivals ($\sigma_0^2 = 0$), in which case the K–dimensional Fork–Join queue can be interpreted as K i.i.d. $D/GI/1$. It is easily shown that

$$\lim_{r \uparrow \infty} (\mu - \lambda(r))\overline{T}_K(\lambda(r)) = H_K \frac{\mathrm{var}(B)}{2} \tag{3.22}$$

where

$$H_K \equiv \sum_{r=1}^{K} \frac{1}{r}. \qquad K = 1, 2, \ldots (3.23)$$

For $\beta = 1$, the services are deterministic and the K–dimensional Fork–Join queue reduces to a $GI/D/1$ whose heavy traffic limit is then given by

$$\lim_{r \uparrow \infty} (\mu - \lambda(r)) \overline{T}_K(\lambda(r)) = \frac{\sigma_0^2}{2}. \tag{3.24}$$

III.3. Light traffic interpolations: Assume known the light and heavy traffic information as described above, say

$$\overline{T}_K(\lambda) \simeq \overline{T}_K(0) + \lambda \overline{T}_K^{(1)}(0) \ldots + \frac{\lambda^n}{n!} \overline{T}_K^{(n)}(0), \quad \lambda \simeq 0 \tag{3.25}$$

for some $n = 0, 1, \ldots$, and

$$\lim_{r \uparrow \infty} (\mu - \lambda(r)) \overline{T}_K(\lambda(r)) = C_K. \tag{3.26}$$

The basic idea behind the light traffic interpolation is to interpolate

$$t_K(\lambda) \equiv (\mu - \lambda) \overline{T}_K(\lambda), \quad 0 \leq \lambda < \mu \tag{3.27}$$

by a polynomial $\hat{t}_K(\lambda)$ of order $n + 1$, say of the form

$$\hat{t}_K(\lambda) = g_0 + g_1 \lambda + \ldots + g_{n+1} \lambda^{n+1}, \quad \lambda \in I\!R. \tag{3.28}$$

The $n+2$ unknown $g_0, g_1, \ldots, g_{n+1}$ are determined by $n+2$ matching conditions inferred from (3.25)–(3.26), namely the heavy traffic condition

$$\hat{t}_K(\mu^-) = t_K(\mu^-) = C_K \tag{3.29}$$

and the light traffic conditions

$$\hat{t}_K^{(m)}(0) = t_K^{(m)}(0), \quad m = 0, 1, \ldots, n \tag{3.30}$$

with $\hat{t}_K^{(m)}(0)$ and $t_K^{(m)}(0)$ denoting the derivative of order m of $\hat{t}_K(\lambda)$ and $t_K(\lambda)$, respectively, at $\lambda = 0$. Once this interpolation has been performed, we undo the normalization and settle on $\hat{T}_K(\lambda)$ given by

$$\hat{T}_K(\lambda) \equiv \frac{\hat{t}_K(\lambda)}{(\mu - \lambda)}, \quad 0 \leq \lambda < \mu \tag{3.31}$$

as the approximation to $\overline{T}_K(\lambda)$. We refer to $\hat{T}_K(\lambda)$ as the interpolation approximation of order n. Throughout this paper, we consider only either zero order ($n = 0$) or first order ($n = 1$) approximations.

As pointed out in Section III.2, it may not be possible in some cases to obtain a closed form expression for the heavy traffic limit C_K. In such situations, we propose ways to approximate C_K, say by $\hat{C}_K$, and use this approximate value $\hat{C}_K$ instead in the interpolation condition (3.29).

IV. THE MARKOVIAN CASE

In this section we develop approximations for the so–called Markovian case, i.e., Poisson arrivals with rate λ and exponential service times with rate μ. This situation is characterized by

$$A(x) = 1 - e^{-\lambda x}, \quad x \geq 0 \tag{4.1}$$

and

$$B(x) = 1 - e^{-\mu x}, \quad x \geq 0. \tag{4.2}$$

Since $\mathrm{var}(A) = \lambda^{-2}$ and $\mathrm{var}(B) = \mu^{-2}$, in heavy traffic ($\lambda \uparrow \mu$) we find $\sigma_0^2 = \mu^{-2}$ and $\beta = \frac{1}{2}$.

IV.1. The case K = 2 – A first order approximation: We first present the first order approximation as described in Section III.3: The heavy traffic result (3.21) reduces here to

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda)\overline{T}_2(\lambda) = \frac{11}{8}, \tag{4.3}$$

and upon specializing the light traffic result (4.12) to the case $K = 2$, we find

$$\overline{T}_2(0) = \frac{3}{2\mu} \quad \text{and} \quad \overline{T}_2'(0) = \frac{11}{8\mu^2}. \tag{4.4}$$

Using (4.3)–(4.4), with $t_2(\lambda)$ given by (3.27), we readily obtain the relations

$$t_2(0) = \frac{3}{2}, \quad t_2'(0) = -\frac{1}{8\mu}, \quad t_2(\mu^-) = \frac{11}{8}. \tag{4.5}$$

If $\hat{t}_2(\lambda)$ denotes the corresponding quadratic interpolation of $t_2(\lambda)$ over the range $[0, \mu]$, say $\hat{t}_2(\lambda) = g_0 + g_1\lambda + g_2\lambda^2$, $0 \leq \lambda \leq \mu$, then the matching conditions (3.29)–(3.30) easily imply $g_0 = \frac{3}{2}$, $g_1 = -\frac{1}{8\mu}$ and $g_2 = 0$, and we get

$$\hat{t}_2(\lambda) = \frac{3}{2} - \frac{1}{8}\frac{\lambda}{\mu}, \quad 0 \leq \lambda \leq \mu. \tag{4.6}$$

Undoing the normalization, we obtain the first order approximation $\hat{T}_2(\lambda)$ to the average response time in steady state in the form

$$\hat{T}_2(\lambda) = \frac{3}{2(\mu - \lambda)} - \frac{\lambda}{8\mu}\frac{1}{(\mu - \lambda)}, \quad 0 \leq \lambda < \mu. \tag{4.7}$$

Several points are worth noticing at this stage:

1. The first order approximation (4.7) to the average response time of the two–dimensional Fork–Join queue will coincide with its zero order approximation; this is so because $g_2 = 0$.

2. More significant perhaps is the fact that the approximation (4.7) is in fact exact. Indeed, in [10] Nelson and Tantawi derived a closed form expression for the average response time $\overline{T}_2(\lambda)$ of a two–dimensional Fork–Join queue. They showed that

$$\overline{T}_2(\lambda) = \frac{12 - \frac{\lambda}{\mu}}{8(\mu - \lambda)}, \quad 0 \leq \lambda < \mu \tag{4.8}$$

and the equality $\hat{T}_2(\lambda) = \overline{T}_2(\lambda)$ follows by direct inspection of (4.7). This is an encouraging fact and bodes well for the accuracy of the method.

IV.2. The general case $K \geq 2$ – A conjecture: So far we have developed approximations only for two dimensional Fork–Join queues due to the fact that only in the case $K = 2$, were we able solve the basic PDE for the stationary distribution of the diffusion limit of the end–to–end delay [17,19]. Because of the complexity involved, it is unlikely that we shall be able to obtain heavy traffic limits for the case $K > 2$ by solving the corresponding PDEs. Hence, even though light traffic limits are available for $K > 2$, our ignorance of the corresponding heavy traffic limits prevents us from obtaining interpolation approximations in accordance with the program of Section III.

To circumvent this difficulty, we seek to *approximate* the heavy traffic limit for the case $K > 2$. We wish to do so in such a way so that for $K = 2$ the approximation to the heavy traffic limit coincides with the exact expression (4.3). The following observation provides the underpinning of our approach: For the case $K = 1$, the Fork–Join system reduces to a single $M/M/1$ queue, and by direct inspection of classical expressions, we easily verify that

$$\mu^2 \overline{T}'_1(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_1(\lambda) = 1. \tag{4.9}$$

For the case $K = 2$ (with exponential inter–arrival and service times), we also note from (4.3)-(4.4) that

$$\mu^2 \overline{T}'_2(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}(\lambda) = \frac{11}{8}. \tag{4.10}$$

It would be tempting (and very desirable for our purpose) to believe that (4.9)–(4.10) hold more generally for all K–dimensional Fork–Join queues, with Poisson arrivals and exponential service times, namely

$$\mu^2 \overline{T}'_K(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_K(\lambda). \qquad K = 3, 4, \ldots \tag{4.11}$$

Although we were not able to validate this relation, we shall nevertheless use it in generating first order approximations. Therefore, in order to proceed, we *assume* the validity of (4.11) for M/M systems. The significance of doing so is immediate in that it enables us to derive an estimate to the heavy traffic limit since the light traffic derivative can be computed here, as shown in Appendix A. The approximations based on this conjecture agree extremely well with both simulation results and the so–called "scaling approximation" of Nelson and Tantawi [10].

IV.3. The general case $\mathbf{K \geq 2}$ – A first order approximation: We derive here approximations to $\overline{T}_K(\lambda)$ for arbitrary $K \geq 2$, with the help of (4.11). Expressions for the light traffic quantities $\overline{T}_K(0)$ and $\overline{T}'_K(0)$ are derived in Appendix A in accordance with the developments of Section III.2. In particular, we show there that

$$\overline{T}_K(0) = \frac{H_K}{\mu} \quad \text{and} \quad \overline{T}'_K(0) = \frac{V_K}{\mu^2}, \qquad K = 2, 3, \ldots (4.12)$$

where H_K is given by (3.23) and we have used the notation

$$V_K \equiv \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=1}^{r} \binom{r}{m} \frac{(m-1)!}{r^{m+1}}. \qquad K = 2, 3, \ldots (4.13)$$

The values of H_K and V_K have been tabulated for $K = 1, \ldots, 20$ in Appendix E.

To approximate the heavy traffic limit, we now make use of the conjectured equality (4.11), which here takes the form

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_K(\lambda) = V_K. \qquad K = 2, 3, \ldots (4.14)$$

As in Section IV.1., we now combine (4.12)–(4.14) to obtain a first order approximation $\hat{T}_K(\lambda)$ to $\overline{T}_K(\lambda)$; elementary computations show that

$$\hat{T}_K(\lambda) = \left[H_K + (V_K - H_K) \frac{\lambda}{\mu} \right] \frac{1}{\mu - \lambda}, \quad 0 \leq \lambda < \mu. \qquad K = 2, 3, \ldots (4.15)$$

By using both experimental as well as theoretical considerations, Nelson and Tantawi [10] have also derived an approximation $\hat{T}_K^{NT}(\lambda)$ for the average response time of a symmetric K–dimensional Fork–Join queues with exponential service and inter–arrival times. Their "scaling approximation" is given by

$$\hat{T}_K^{NT}(\lambda) = \left[\frac{H_K}{H_2} + \frac{4}{11} (1 - \frac{H_K}{H_2}) \frac{\lambda}{\mu} \right] \overline{T}_2(\lambda), \quad 0 \leq \lambda < \mu \qquad K = 2, 3, \ldots (4.16)$$

with $\overline{T}_2(\lambda)$ given by (4.8). The relative error of their approximation as compared to simulation results was shown to be less than 5% for systems where $K \leq 32$.

We have checked our approximation (4.15) against that of Nelson and Tantawi for $K \leq 15$, and our approximation seems to perform just as well as shown in Section IV.4. The two approximations are closely related as they are both "anchored" at comparable light and heavy traffic regimes. The latter can be seen by taking the heavy traffic limit of $\hat{T}_K^{NT}$, namely

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \hat{T}_K^{NT}(\lambda) = \frac{1}{2} + \frac{7}{12} H_K. \qquad K = 2, 3, \ldots (4.17)$$

Using the table in Appendix E, the reader may check that indeed the right handside of (4.17) agrees quite closely with V_K. The advantages of our approximation over that of Nelson and Tantawi are however two–fold:

1. Nelson and Tantawi resorted to experimental results to obtain the values of the constants in their approximation, while we give exact closed–form expressions for all the constants appearing in our approximation.

2. The approximation (4.16) of Nelson and Tantawi is only valid for Fork–Join queues with exponential inter–arrival and service distributions. On the other hand, as we show in Sections V and VI, our approximation procedure can be extended to cover Fork–Join queues with general inter–arrival and service distributions.

IV.4. Simulation results: In this section, the approximation (4.15) is compared with simulation results for the case when $\mu = 1$ with $K = 2, 5, 10$ and 15. For the cases $K = 5, 10$ and 15, we have also given the Nelson–Tantawi approximation (4.16) for comparison purposes. For the case $K = 2$, the two approximations are identical.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error	$\hat{T}_5^{NT}(\lambda)$
0.1	1.65 ± 0.007	1.65	0.30	2.49 ± 0.008	2.49	0.04	2.48
0.2	1.85 ± 0.011	1.84	0.43	2.75 ± 0.012	2.75	0.07	2.73
0.3	2.09 ± 0.0158	2.09	0.43	3.09 ± 0.017	3.08	0.32	3.06
0.4	2.44 ± 0.024	2.41	1.22	3.55 ± 0.027	3.52	0.84	3.49
0.5	2.91 ± 0.037	2.87	1.37	4.19 ± 0.042	4.14	1.09	4.10
0.6	3.63 ± 0.061	3.56	1.93	5.16 ± 0.074	5.07	1.74	5.01
0.7	4.80 ± 0.109	4.71	1.87	6.80 ± 0.154	6.62	2.64	6.54
0.8	7.16 ± 0.23	7.0	2.23	9.85 ± 0.09	9.72	1.3	9.59
0.9	13.91 ± 0.32	13.87	0.29	19.30 ± 0.43	19.02	1.45	1 8.74

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error	$\hat{T}_{10}^{NT}(\lambda)$
0.1	3.17 ± 0.009	3.17	0.09	3.16
0.2	3.48 ± 0.013	3.48	0.06	3.47
0.3	3.88 ± 0.018	3.86	0.51	3.86
0.4	4.42 ± 0.026	4.39	0.68	4.38
0.5	5.18 ± 0.042	5.12	1.16	5.11
0.6	6.34 ± 0.072	6.22	1.89	6.21
0.7	8.23 ± 0.137	8.05	2.18	8.05
0.8	11.92 ± 0.30	11.71	1.76	11.72
0.8	13.44 ± 0.392	12.83	4.54	13.01
0.9	23.49 ± 0.41	22.68	3.45	22.76

λ	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error	$\hat{T}_{15}^{NT}(\lambda)$
0.1	3.58 ± 0.009	3.58	0.03	3.58
0.2	3.91 ± 0.013	3.91	0.13	3.91
0.3	4.35 ± 0.020	4.34	0.23	4.34
0.4	4.95 ± 0.031	4.90	1.01	4.92
0.5	5.78 ± 0.050	5.70	1.38	5.72
0.6	7.03 ± 0.086	6.88	2.13	6.94
0.7	9.14 ± 0.166	8.87	3.04	8.96
0.8	13.44 ± 0.392	12.83	4.54	13.01
0.9	25.90 ± 0.47	24.73	4.52	25.18

V. THE NON–MARKOVIAN CASE WITH $\beta = \frac{1}{2}$

In Section IV, we were able to obtain approximations for symmetric Fork–Join queues with Poisson arrivals and exponential services by *postulating* (4.11), a relation between the heavy and light traffic regimes. As our next step, we would like to obtain approximations for Fork–Join queues with more general arrival and service characteristics. There are however certain difficulties for carrying out this program: Extending the light traffic theory of Section III to arbitrary arrival patterns may not be possible. Moreover, it is not at all clear that the postulated relation (4.11) is still appropriate in the general case. We plan to circumvent the first difficulty by considering only zero order approximations whenever non–Poissonian streams are involved, in which case (3.6)–(3.7) still hold. The second difficulty is addressed by developing an estimate for the heavy traffic limit for general inter–arrival and service distributions. This will be accomplished in two steps, the first one being taken here in the non–Markovian case with $\beta = \frac{1}{2}$; the second step is discussed in Section VI.

V.1. A zero order approximation: In the Markovian situation, we have $\beta = \frac{1}{2}$, and for $K = 2$, the exact heavy traffic limit is available [17,19] in the form

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_2(\lambda) = \frac{11}{8}. \tag{5.1}$$

For arbitrary K, such a result was not available and we resorted to a conjectured relationship between the heavy and light traffic regimes in order to generate a plausible estimate of the heavy traffic limit. To do so, we extended the domain of validity of the relations (4.9)–(4.10) to all K by postulating

$$\mu^2 \overline{T}'_K(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_K(\lambda) = V_K. \qquad K = 1, 2, \ldots (5.2)$$

Intent on extending this approach, we recall that an exact heavy traffic limit (such as (5.1)) is available for $K = 2$ under the weaker assumption $\beta = \frac{1}{2}$, and takes the general form

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_2(\lambda) = \frac{11}{8} \sigma^2 \mu^2 \tag{5.3}$$

with again no such result being available for $K > 2$. Unfortunately, for non–Poissonian arrivals the equality (5.2) may not even hold for $K = 1$; for instance, in a single server queue with Erlang–2 inter–arrival times and exponential service times, we have

$$\mu^2\overline{T}_1'(0) = 0 \neq \frac{3}{4} = \lim_{\lambda\uparrow\mu}(\mu - \lambda)\overline{T}_1(\lambda). \tag{5.4}$$

However, as we compare the form of the expressions (5.1)–(5.3) when $K = 2$, and recall that $V_2 = \frac{11}{8}$, we would expect that for the case $K > 2$ and $\beta = \frac{1}{2}$, the relation

$$\lim_{\lambda\uparrow\mu}(\mu - \lambda)\overline{T}_K(\lambda) = V_K\sigma^2\mu^2 \qquad K = 3, 4, \ldots (5.5)$$

is either true or at least plausible.

As we now show, (5.5) provides the underpining for a zero order approximation to $\overline{T}_K(\lambda)$ for all K when $\beta = \frac{1}{2}$. With $t_K(\lambda)$ still defined by (3.27), we readily see that

$$t_K(0) = \mu\overline{T}_K(0) \quad \text{and} \quad t_K(\mu^-) = V_K\sigma^2\mu^2. \tag{5.6}$$

If $\hat{t}_K(\lambda)$ denotes the corresponding linear interpolation of $t_K(\lambda)$ over the range $[0, \mu]$, say $\hat{t}_K(\lambda) = g_0 + g_1\lambda$, $0 \le \lambda \le \mu$, then (5.6) yields the relations $g_0 = \mu\overline{T}_K(0)$ and $g_0 + g_1\mu = V_K\sigma^2\mu^2$ from which we conclude that

$$\hat{t}_K(\lambda) = \mu\overline{T}_K(0) + (V_K\sigma^2\mu - \overline{T}_K(0))\lambda, \quad 0 \le \lambda \le \mu. \tag{5.7}$$

Undoing the normalization, we obtain the zero order approximation $\hat{T}_K(\lambda)$ to $\overline{T}_K(\lambda)$ in the form

$$\hat{T}_K(\lambda) = \left[\mu\overline{T}_K(0) + (V_K\sigma^2\mu^2 - \mu\overline{T}_K(0))\frac{\lambda}{\mu}\right]\frac{1}{\mu - \lambda}, \quad 0 \le \lambda < \mu \quad K = 2, 3, \ldots (5.8)$$

We illustrate the accuracy of this approximation on a simple example.

V.2. An example – Erlang–2 arrivals and service times: The K–dimensional Fork–Join queue with Erlang–2 inter–arrival and service time distributions is characterized by

$$A(x) = 1 - (1 + 2\lambda x)e^{-2\lambda x}, \quad x \ge 0 \tag{5.9}$$

and

$$B(x) = 1 - (1 + 2\mu x)e^{-2\mu x}, \quad x \ge 0. \tag{5.10}$$

Here, $\text{var}(A) = (2\lambda^2)^{-2}$ and $\text{var}(B) = (2\mu^2)^{-2}$, so that in heavy traffic ($\lambda \uparrow \mu$), we have $\sigma_0^2 = (2\mu^2)^{-2}$ and $\beta = \frac{1}{2}$. An application of (5.5) to this queue yields

$$\lim_{\lambda\uparrow\mu}(\mu - \lambda)\overline{T}_K(\lambda) = \frac{V_K}{2} \qquad K = 2, 3, \ldots (5.11)$$

whereas a formula for $\overline{T}_K(0)$ is derived in Appendix B, namely

$$\overline{T}_K(0) = \frac{F_K}{\mu} \qquad K = 2, 3, \ldots (5.12)$$

where

$$F_K \equiv \frac{1}{\mu} \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} \frac{m!}{2r^{m+1}}. \qquad K = 2, 3, \ldots (5.13)$$

The numbers F_K, $K = 1, \ldots, 20$, are tabulated in Appendix E.

Finally, upon substituting (5.11)–(5.12) into (5.8), we obtain a zero order approximation to the average response time of a K–dimensional Fork–Join queue with two–stage Erlang inter–arrival and service distributions. This approximation takes the form

$$\hat{T}_K(\lambda) = \left[F_K + (\frac{V_K}{2} - F_K)\frac{\lambda}{\mu} \right] \frac{1}{\mu - \lambda}, \quad 0 \le \lambda < \mu \qquad K = 2, 3, \ldots (5.14)$$

and agrees extremely well with simulation results. Below we compare the zero order approximation (5.14) with simulation results for the cases $K = 2, 5, 10$ and 15.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	1.40 ± 0.004	1.45	3.57	1.93 ± 0.004	2.01	4.14
0.2	1.46 ± 0.005	1.54	5.48	2.01 ± 0.005	2.13	5.97
0.3	1.55 ± 0.006	1.67	7.74	2.14 ± 0.007	2.30	7.47
0.4	1.69 ± 0.008	1.83	8.28	2.33 ± 0.010	2.52	8.15
0.5	1.91 ± 0.012	2.06	7.85	2.62 ± 0.015	2.83	8.01
0.6	2.23 ± 0.021	2.41	8.07	3.06 ± 0.024	3.29	7.51
0.7	2.80 ± 0.041	2.98	6.42	3.83 ± 0.046	4.07	6.26
0.8	4.03 ± 0.11	4.12	2.23	5.43 ± 0.11	5.62	3.49
0.9	7.44 ± 0.14	7.56	1.61	10.11 ± 0.15	10.27	1.58

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error
0.1	2.34 ± 0.004	2.43	3.84	2.58 ± 0.003	2.68	3.87
0.2	2.43 ± 0.005	2.58	6.17	2.68 ± 0.004	2.85	6.34
0.3	2.58 ± 0.007	2.78	7.75	2.84 ± 0.006	3.06	7.74
0.4	2.81 ± 0.010	3.04	8.18	3.08 ± 0.009	3.34	8.44
0.5	3.15 ± 0.014	3.41	8.25	3.43 ± 0.014	3.74	9.03
0.6	3.68 ± 0.022	3.96	7.60	33.99 ± 0.025	4.33	8.52
0.7	4.59 ± 0.043	4.87	6.10	4.94 ± 0.052	5.32	7.69
0.8	6.45 ± 0.09	6.70	3.87	6.87 ± 0.12	7.31	6.40
0.9	12.09 ± 0.15	12.19	0.83	12.89 ± 0.47	13.25	2.79

VI. THE GENERAL NON–MARKOVIAN CASE

In the last section we derived a zero order approximation in the non–Markovian case when $\beta = \frac{1}{2}$. As our next step, we would like to obtain approximations for Fork–Join queues possessing more general arrival and service characteristics. However, an equality such as (5.5) may no longer be true nor plausible, and a different approach is required for developing an estimate for the heavy traffic limit in the case of general inter–arrival and service distributions. The main idea behind this approximation is as follows: The heavy traffic limit is easily obtained in the extreme cases $\beta = 0$ (deterministic arrivals) and $\beta = 1$ (deterministic services), while in Section V, we proposed an estimate of the heavy traffic limit for the case $\beta = \frac{1}{2}$. This points the way to a quadratic interpolation (in β over $[0, 1]$) to produce a formula for the heavy traffic limit for general inter–arrival and service distributions.

1. $\beta = 0$ – The K queues are now decoupled from each other since the arrival stream is deterministic (because $\sigma_0 = 0$), and using (3.22), we can write

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = H_K \frac{\sigma^2 \mu^2}{2}. \qquad K = 2, 3, \ldots (6.1)$$

2. $\beta = \frac{1}{2}$ – In this case $\sigma_0 = \sigma$ and according to our conjecture (4.2), we have

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = V_K \sigma^2 \mu^2. \qquad K = 2, 3, \ldots (6.2)$$

3. $\beta = 1$ – In this case $\sigma = 0$ and the system behaves essentially like a $GI/D/1$ queue so that

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = \frac{\sigma_0^2 \mu^2}{2}. \qquad K = 2, 3, \ldots (6.3)$$

Observing the structure of (6.1)–(6.3), we venture the following approximation

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = M_K(\beta) \frac{\sigma^2 + \sigma_0^2}{2} \mu^2, \quad 0 \leq \beta \leq 1 \qquad K = 2, 3, \ldots (6.4)$$

for the heavy traffic limit, where

$$M_K(0) = H_K, \quad M_K(\frac{1}{2}) = V_K, \quad M_K(1) = 1. \qquad (6.5)$$

In the absence of any additional information we may use a quadratic approximation for $M(\beta)$ anchored at (6.5), and this leads to the heavy traffic estimate

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) \qquad K = 2, 3, \ldots (6.6)$$

$$= \left[H_K + (4V_K - 3H_K - 1)\beta + 2(1 + H_K - 2V_K)\beta^2\right] \frac{\sigma^2 + \sigma_0^2}{2} \mu^2, \quad 0 \leq \beta \leq 1.$$

This estimate of the heavy traffic limit can now be used in conjunction with light traffic information to yield an approximation. In the remaining sections of this paper, we validate (6.6) by comparing it with simulations for various choices of inter–arrival and service distributions. We consider the following situations:

1. Erlang–2 arrivals and exponential services;
2. Poisson arrivals and hyper–exponential services;
3. Poisson arrivals and Erlang–2 services; and
4. Hyper–exponential arrivals and exponential services.

VI.2. Example 1 – Erlang–2 arrivals and exponential services: Consider the case when A and B are given by (5.9) and (4.2), respectively. Therefore, $\mathrm{var}(A) = (2\lambda^2)^{-1}$ and $\mathrm{var}(B) = \mu^{-2}$, and in heavy traffic, we find $\sigma_0^2 = (2\mu^2)^{-1}$, whence $\beta = \frac{1}{3}$. Applying (6.6) we obtain

$$\lim_{\lambda\uparrow\mu}(\mu-\lambda)\overline{T}_K(\lambda) = -\frac{1}{12} + \frac{1}{6}H_K + \frac{2}{3}V_K. \qquad K = 2,3,\ldots (6.7)$$

This system has the same light traffic limit $\overline{T}_K(0)$ as the system with Poisson arrivals and exponential services considered in Section IV, namely

$$\overline{T}_K(0) = \frac{H_K}{\mu}. \qquad K = 2,3,\ldots (6.8)$$

Combining (6.7) and (6.8) we obtain the zero order approximation $\hat{T}_K(\lambda)$ in the form

$$\hat{T}_K(\lambda) = \left[H_K + (\frac{2}{3}V_K - \frac{5}{6}H_K - \frac{1}{12})\frac{\lambda}{\mu}\right]\frac{1}{\mu-\lambda}, \quad 0 \le \lambda < \mu. \qquad K = 2,3,\ldots (6.9)$$

This approximation agrees extremely well with simulation results as we show for the case when $\mu = 1$ with $K = 2, 5, 10$ and 15.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	1.55 ± 0.007	1.62	4.51	2.34 ± 0.007	2.45	4.70
0.2	1.65 ± 0.009	1.77	7.27	2.49 ± 0.008	2.67	7.23
0.3	1.81 ± 0.011	1.96	8.28	2.71 ± 0.011	2.94	8.48
0.4	2.05 ± 0.015	2.22	8.29	3.03 ± 0.016	3.31	9.24
0.5	2.39 ± 0.024	2.58	7.95	3.50 ± 0.026	3.82	9.14
0.6	2.91 ± 0.042	3.12	7.21	4.22 ± 0.044	4.59	8.77
0.7	3.81 ± 0.088	4.03	5.77	5.47 ± 0.086	5.87	7.31
0.8	5.65 ± 0.203	5.83	3.18	8.01 ± 0.21	8.43	5.24
0.9	10.96 ± 0.24	11.25	2.64	15.75 ± 0.27	16.12	2.34

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error
0.1	3.00 ± 0.007	3.13	4.33	3.39 ± 0.007	3.54	4.30
0.2	3.18 ± 0.009	3.39	6.60	3.59 ± 0.009	3.83	6.68
0.3	3.47 ± 0.013	3.73	7.49	3.89 ± 0.011	4.20	7.97
0.4	3.88 ± 0.019	4.17	7.47	4.34 ± 0.017	4.69	8.06
0.5	4.49 ± 0.03	4.79	6.68	4.99 ± 0.027	5.37	7.61
0.6	5.41 ± 0.05	5.73	5.91	5.99 ± 0.045	6.40	6.84
0.7	6.99 ± 0.09	7.29	4.29	7.69 ± 0.008	8.11	5.46
0.8	10.21 ± 0.22	10.40	1.86	11.08 ± 0.19	11.53	4.06
0.9	19.51 ± 0.28	19.74	1.18	21.03 ± 0.85	21.80	3.67

VI.2. Example 2 – Poisson arrivals and hyper–exponential services: Here, we assume that A is given by (4.1) and that B has the form

$$B(x) = 1 - \left(p_1 e^{-\mu_1 x} + p_2 e^{-\mu_2 x}\right), \quad x \geq 0 \tag{6.10}$$

with $0 < p_1 < 1$, $p_2 = 1 - p_1$ and $\mu_1 \neq \mu_2$. To simplify the computations, we also assume $\frac{p_1}{\mu_1} = \frac{p_2}{\mu_2} = \frac{1}{2}$ so that $\mu_1 + \mu_2 = 2$. Simple algebra shows that $m(B) = 1$ and $\text{var}(B) = \frac{2}{\mu_1\mu_2} - 1$. Therefore, $\mu = 1$ and in heavy traffic ($\lambda \uparrow \mu$), we find $\sigma_0^2 = \mu^{-2} = 1$ and $\beta = \frac{\mu_1\mu_2}{2}$.

The light traffic limit $\overline{T}_K(0)$ can be computed via (3.6)–(3.7), and takes the form

$$\overline{T}_K(0) = \sum_{r=1}^{K} \binom{K}{r} (-1)^{r+1} \sum_{m=0}^{r} \binom{r}{m} \frac{p_1^m p_2^{r-m}}{m\mu_1 + (r-m)\mu_2}. \tag{6.11}$$

An expression for the light traffic derivative $\overline{T}'_K(0)$ is obtained in Appendix C, namely

$$\begin{aligned}\overline{T}'_K(0) = &\sum_{r=1}^{K} \binom{K}{r} (-1)^{r+1} \sum_{m_1=0}^{K-r} (-1)^{m_1} \sum_{m_2=0}^{m_1} \binom{m_1}{m_2} p_1^{m_2} p_2^{m_1-m_2} \\ &\times \sum_{k_1=0}^{r} \binom{r}{k_1} (\frac{p_1 p_2}{\mu_2 - \mu_1})^{r-k_1} \sum_{k_2=0}^{k_1} \binom{k_1}{k_2} (p_1^2\mu_1)^{k_2} (p_2^2\mu_2)^{k_1-k_2} \\ &\times \sum_{k_3=0}^{r-k_1} \binom{r-k_1}{k_3} (-1)^{r-k_1-k_3} \sum_{k_4=0}^{r-k_1} \binom{r-k_1}{k_4} \mu_2^{k_4} \mu_1^{r-k_1-k_4} \\ &\times \frac{k_1!}{(\mu_1(m_2+k_2+k_3) + \mu_2(m_1 - m_2 - k_2 + r - k_3))^{k_1+1}} \\ &\times \frac{1}{\mu_1(k_2+k_4) + \mu_2(r - k_2 - k_4)}.\end{aligned} \tag{6.12}$$

We consider the special case when $\mu_1 = 0.1$, $\mu_2 = 1.9$, $p_1 = 0.05$ and $p_2 = 0.95$, in which case $\mu = 1$ and $\beta = 0.095$. The values of $D_K \equiv \overline{T}_K(0)$ and $E_K \equiv \overline{T}'_K(0)$,

$K = 1, \ldots, 20$, are tabulated in Appendix E. Substituting these parameter values into (6.6), we readily conclude that the estimate of the heavy traffic limit is given by

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = 3.85H_K + 1.81V_K - 0.4. \tag{6.13}$$

Combining (6.11)–(6.13), we finally obtain the first order approximation

$$\begin{aligned} \hat{T}_K(\lambda) =& \frac{\mu D_K}{\mu - \lambda} + [\mu E_K - D_K]\frac{\lambda}{\mu - \lambda} \qquad\qquad K = 2, 3, \ldots (6.14) \\ &+ \left[3.85H_K + 1.81V_K - 0.4 - \mu^2 E_K\right] (\frac{\lambda}{\mu})^2 \frac{1}{\mu - \lambda}, \quad 0 \leq \lambda < \mu. \end{aligned}$$

For the cases $K = 2, 5, 10$ and 15, (6.14) becomes

$$\hat{T}_2(\lambda) = \frac{1.702 + 8.118\lambda - 1.94\lambda^2}{1 - \lambda}, \quad 0 \leq \lambda \leq 1 \tag{6.15}$$

$$\hat{T}_5(\lambda) = \frac{3.324 + 18.116\lambda - 9.68\lambda^2}{1 - \lambda}, \quad 0 \leq \lambda \leq 1 \tag{6.16}$$

$$\hat{T}_{10}(\lambda) = \frac{5.447 + 31.013\lambda - 21.61\lambda^2}{1 - \lambda}, \quad 0 \leq \lambda \leq 1 \tag{6.17}$$

and

$$\hat{T}_{15}(\lambda) = \frac{7.227 + 40.573\lambda - 31.12\lambda^2}{1 - \lambda}, \quad 0 \leq \lambda \leq 1. \tag{6.18}$$

These approximations are compared below with simulation results, and as the reader may note, the agreement is quite good.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	2.78 ± 0.021	2.77	0.36	5.59 ± 0.030	5.60	0.17
0.2	4.09 ± 0.042	4.06	0.73	8.22 ± 0.054	8.20	0.24
0.3	5.71 ± 0.067	5.66	0.87	11.32 ± 0.086	11.27	0.35
0.4	7.83 ± 0.105	7.73	1.28	15.10 ± 0.130	15.04	0.39
0.5	10.73 ± 0.172	10.55	1.67	19.93 ± 0.198	19.92	0.05
0.6	14.95 ± 0.289	14.69	1.74	26.70 ± 0.332	26.77	0.26
0.7	21.70 ± 0.503	21.45	1.15	37.40 ± 0.609	37.54	0.37
0.8	34.71 ± 1.50	34.77	0.17	57.94 ± 1.30	58.11	0.29
0.9	71.14 ± 2.59	74.37	4.54	114.03 ± 3.11	117.88	3.37

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error
0.1	9.18 ± 0.038	9.27	0.98	11.99 ± 0.041	12.19	1.67
0.2	13.16 ± 0.066	13.48	2.43	16.79 ± 0.069	17.62	4.94
0.3	17.56 ± 0.099	18.29	4.16	21.92 ± 0.106	23.71	8.17
0.4	22.67 ± 0.147	24.05	6.08	27.76 ± 0.157	30.79	10.91
0.5	28.99 ± 0.219	31.10	7.28	34.94 ± 0.234	39.47	12.96
0.6	37.64 ± 0.34	40.82	8.45	44.72 ± 0.363	50.92	13.86
0.7	50.89 ± 0.59	55.22	8.51	59.97 ± 0.630	67.93	13.27
0.8	79.95 ± 1.19	82.49	3.17	89.09 ± 1.29	98.84	10.94
0.9	146.72 ± 2.95	158.54	8.05	174.26 ± 3.28	185.35	6.36

VI.4. Example 3 – Poisson arrivals and Erlang–2 services: Consider the case when A and B are given by (4.1) and (5.10), respectively. Here we have $\text{var}(A) = \lambda^{-2}$ and $\text{var}(B) = (2\mu)^{-2}$, and in heavy traffic ($\lambda \uparrow \mu$), we find $\sigma_0^2 = \mu^{-2}$ so that $\beta = \frac{2}{3}$. In this case (6.6) becomes

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = \frac{1}{6} - \frac{H_K}{12} + \frac{2}{3}V_K. \qquad K = 2, 3, \ldots (6.19)$$

The light traffic limit $\overline{T}_K(0)$ is the same as that for the system with Erlang–2 arrivals and services, and is therefore given by (5.12), i.e.,

$$\overline{T}_K(0) = \frac{F_K}{\mu} \qquad K = 2, 3, \ldots (6.20)$$

with F_K given by (5.13). Combining (6.19) and (6.20) we obtain the following zero order approximation $\hat{T}_K(\lambda)$ in the form

$$\hat{T}_K(\lambda) = \left[F_K + (\frac{1}{6} - \frac{H_K}{12} + \frac{2}{3}V_K - F_K)\frac{\lambda}{\mu}\right]\frac{1}{\mu - \lambda}, \quad 0 \le \lambda < \mu. \quad K = 2, 3, \ldots (6.21)$$

This approximation is in very good agreement with simulation results, as we now show for the case when $\mu = 1$ with $K = 2, 5, 10$ and 15.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	1.48 ± 0.005	1.48	0.13	2.03 ± 0.005	2.04	0.24
0.2	1.62 ± 0.007	1.61	0.37	2.20 ± 0.007	2.21	0.45
0.3	1.79 ± 0.009	1.78	0.50	2.43 ± 0.011	2.42	0.41
0.4	2.02 ± 0.014	2.01	0.49	2.73 ± 0.015	2.71	0.73
0.5	2.35 ± 0.022	2.33	0.86	3.15 ± 0.022	3.11	1.26
0.6	2.85 ± 0.039	2.81	1.4	3.80 ± 0.041	3.72	2.10
0.7	3.65 ± 0.07	3.61	1.09	4.88 ± 0.079	4.73	3.07
0.8	5.30 ± 0.19	5.20	1.88	6.98 ± 0.06	6.74	3.43
0.9	10.08 ± 0.20	9.98	0.99	13.34 ± 0.23	12.79	4.12

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error
0.1	2.46 ± 0.005	2.46	0.08	2.71 ± 0.005	2.71	0.07
0.2	2.66 ± 0.007	2.66	0.007	2.92 ± 0.008	2.92	0.07
0.3	2.92 ± 0.011	2.91	0.31	3.19 ± 0.012	3.18	0.31
0.4	3.27 ± 0.017	3.23	1.22	3.57 ± 0.021	3.53	1.12
0.5	3.75 ± 0.027	3.70	1.33	4.10 ± 0.037	4.02	1.95
0.6	4.49 ± 0.047	4.39	2.22	4.90 ± 0.062	4.76	2.86
0.7	5.75 ± 0.10	5.54	3.65	6.26 ± 0.114	5.99	4.31
0.8	8.30 ± 0.07	7.85	5.42	9.03 ± 0.29	8.44	6.53
0.9	15.90 ± 0.27	14.77	7.10	17.34 ± 0.30	15.81	8.82

At this point we would like to point out that the excellent agreement with simulation for this system, has been obtained without making use of the light traffic derivative. This is contrast to the case for hyper–exponential services discussed in Section VI.3, where we had to make use of the light traffic derivative in order to obtain good approximations. The reason behind this disparity is that, while for the case of hyper–exponential services, the relation (4.11) is clearly not satisfied, for the case of Erlang–2 services this appears to be the case for Erlang–2 services, as we now demonstrate. In Appendix D the light traffic derivative $\overline{T}'_K(0)$ is evaluated as

$$\overline{T}'_K(0) = \frac{G_K}{\mu^2} \qquad K = 2, 3, \ldots \text{(6.22)}$$

with

$$G_K \equiv \sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{2^{r+1}} \sum_{q=0}^{K-r} \binom{K-r}{q} (-1)^q \sum_{m=0}^{q} \binom{q}{m}$$
$$\sum_{n=0}^{r} \binom{r}{n} \frac{1}{3^{r-n}} \frac{(m+3r-n)!}{(r+q)^{m+3r-n+1}} \sum_{p=0}^{n} \binom{n}{p} \frac{p!}{(2r)^{p+1}}, \quad K = 2, 3, \ldots \qquad (6.23)$$

The values of G_K, $K = 1, \ldots, 20$, have been tabulated in Appendix E. Comparing the right handside of (6.19) with the values of G_K, we observe that they indeed match very closely.

VI.5. Example 4 – Hyper–exponential arrivals and exponential services: We close with the case when A is of the form

$$A(x) = 0.05e^{-10x} + 0.95\gamma e^{-\gamma x}, \quad x \geq 0. \qquad (6.24)$$

for some $\gamma > 0$, and B is given by (4.2) with $\mu = 1$. We have

$$m(A) = \lambda^{-1} = 0.95\gamma^{-1} + 0.005 \qquad (6.25)$$

while the variance of A is given by

$$\operatorname{var}(A) = 0.000975 + \frac{0.9975}{\gamma^2} - \frac{0.0095}{\gamma}. \tag{6.26}$$

Since $m(A) \downarrow 1$ is equivalent to $\gamma \to \frac{0.95}{0.995}$, simple calculations show that $\sigma_0^2 = \lim_{\lambda\uparrow 1}$ var $(A) = 1.085$, and therefore $\beta = 0.52$. Substituting these values into (6.6), we obtain

$$\lim_{\lambda\uparrow 1}(1-\lambda)\overline{T}_K(\lambda) = 1.042V_K - 0.021H_K + 0.021. \qquad K = 2,3,\ldots (6.27)$$

The light traffic limit is the same as the one for the system with Poisson arrivals and exponential services, and is thus given by (4.12). Combining this fact with (6.27) we obtain the approximation

$$\hat{T}_K(\lambda) = \frac{H_K + (1.042V_K - 1.021H_K + 0.021)\lambda}{1-\lambda}, \quad 0 \le \lambda \le 1. \qquad K = 2,3,\ldots (6.28)$$

We observe rather good agreement with simulation results displayed below for the case when $K = 2, 5, 10$ and 15.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	1.72 ± 0.003	1.66	3.49	2.57 ± 0.003	2.49	3.11
0.2	1.91 ± 0.004	1.86	2.62	2.84 ± 0.004	2.76	2.82
0.3	2.15 ± 0.005	2.11	1.86	3.16 ± 0.006	3.10	1.89
0.4	2.50 ± 0.008	2.45	2.00	3.64 ± 0.009	3.56	2.19
0.5	3.00 ± 0.022	2.92	2.67	4.31 ± 0.014	4.19	2.78
0.6	3.68 ± 0.019	3.63	1.36	5.23 ± 0.023	5.15	1.53
0.7	4.92 ± 0.038	4.82	2.03	6.94 ± 0.046	6.75	2.74
0.8	7.09 ± 0.084	7.19	1.31	10.01 ± 0.103	9.93	0.80
0.9	14.32 ± 0.331	14.30	0.14	20.26 ± 0.414	19.50	3.75

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error
0.1	3.26 ± 0.003	3.18	2.45	3.99 ± 0.003	3.59	2.44
0.2	3.58 ± 0.004	3.49	2.51	4.02 ± 0.004	3.92	2.49
0.3	3.96 ± 0.006	3.89	1.77	4.44 ± 0.006	4.36	1.80
0.4	4.54 ± 0.009	4.43	2.42	5.07 ± 0.009	4.94	2.56
0.5	5.36 ± 0.015	5.18	3.36	5.97 ± 0.016	5.74	3.85
0.6	6.46 ± 0.025	6.30	2.48	7.18 ± 0.027	6.96	3.06
0.7	8.52 ± 0.049	8.17	4.11	9.43 ± 0.054	8.98	4.77
0.8	12.17 ± 0.11	11.92	2.05	13.44 ± 0.123	13.03	3.05
0.9	24.34 ± 0.44	23.17	0.73	26.99 ± 0.493	25.17	6.74

REFERENCES

[1] F. Baccelli, *Two parallel queues created by arrivals with two demands: The $M/G/2$ symmetric case*, Rapport de Recherche **426**, INRIA–Rocquencourt (France), July 1985.

[2] F. Baccelli, A.M. Makowski and A. Shwartz, "The Fork–Join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds,"*Advances in Applied Probability* **21** (1989), pp. 629–660.

[3] F. Baccelli and A.M. Makowski, "Queueing models for systems with synchronization constraints,"*Proceedings of the IEEE* **77** (1989). Invited paper, Special Issue on Dynamics of Discrete Event Systems, pp. 138–161.

[4] D. Burman and D. Smith, "Approximate analysis of a queueing model with bursty arrivals,"*Bell System Technical Journal* **62** (1983), pp. 1433–1453.

[5] D. Burman and D. Smith, "An asymptotic analysis of a queueing system with Markov–modulated arrivals," *Operations Research* **34** (1986), pp. 105–119.

[6] L. Flatto and S. Hahn, "Two parallel queues created by arrivals with two demands I,"*SIAM Journal in Applied Mathematics* **44** (1984), pp. 1041–1053.

[7] P. Fleming, "An approximate analysis of sojourn times in the $M/G/1$ queue with round–robin service discipline,"*AT&T Bell Labs Technical Journal* **63** (1984), pp. 1521–1535.

[8] P. Fleming and B. Simon, "Interpolation approximations of sojourn time distributions," *Operations Research* **39** (1991), pp. 251–260.

[9] J.M. Harrison, *Brownian Motion and Stochastic Flow Systems*, J. Wiley & Sons, New York (NY) (1985).

[10] R. Nelson and A.N. Tantawi, "Approximate analysis of Fork–Join synchronization in parallel queues,"*IEEE Transactions on Computers* **C–37** (1988), pp. 739–743.

[11] V. Nguyen, *Heavy Traffic Analysis of Processing Networks With Parallel and Sequential Tasks*, Ph.D. Thesis, Stanford University, Palo Alto (CA), 1990.

[12] M.I. Reiman and B. Simon, "An interpolation approximation for queueing systems with Poisson input,"*Operations Research* **36** (1988), pp. 454–469.

[13] M.I. Reiman and B. Simon, "Light traffic limits of sojourn time distributions in Markovian queueing networks,"*Stochastic Models* **4** (1988), pp. 191–233.

[14] M.I. Reiman and B. Simon, "Open queueing systems in light traffic,"*Mathematics of Operations Research* **14** (1989), pp. 26-59.

[15] B. Simon and S. Willie, "Estimation of response time characteristics in priority queueing networks via an interpolation methodology based on simulation and heavy traffic limits,"in *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface*, American Statistical Association (1986), pp. 251–256.

[16] Tedijanto, *Nonexhaustive Policies in Polling Systems and Vacation Models: Qualitative and Approximate Approaches*, Ph.D. Thesis, University of Maryland, College Park (MD), August 1990.

[17] S. Varma, *Heavy and Light Traffic Approximations For Queues With Synchronization Constraints*, Ph.D. Thesis, University of Maryland, College Park MD), August 1990.

[18] S. Varma, "Performance evaluation of the time–stamp ordering algorithm in distributed databases,"*IEEE Transactions on Distributed and Parallel Systems*, forthcoming 1993.

[19] S. Varma and A.M. Makowski, "Heavy traffic limits for Fork–Join queues,"In preparation (1992).

[20] W. Whitt, "Heavy traffic limit theorems for queues: A survey,"in Lecture Notes in Economics and Mathematical Systems **98**, Springer–Verlag, Berlin, pp. 307–350 (1974).

APPENDICES

APPENDIX A – We now show how to compute $\overline{T}_K(0)$ and the light traffic derivative $\overline{T}'_K(0)$ in the Markovian case. With B given by (4.2), we find by the binomial expansion that

$$B(x)^K = \sum_{r=0}^{K} \binom{K}{r} (-1)^r e^{-r\mu x}, \quad x \geq 0. \tag{A.1}$$

Substituting into (3.7), we readily get

$$\overline{T}_K(0) = \int_0^\infty (1 - F_K(x))\, dx = \frac{1}{\mu} \sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{r} \tag{A.2}$$

and the first equation in (4.12) follows from (A.2) by making use of the identity

$$H_K = \sum_{r=1}^{K} \frac{1}{r} = \sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r+1}}{r}. \qquad K = 1, 2, \ldots \tag{A.3}$$

Fixing $t < 0$, we see that the common distribution F_Y of the rvs B defined through (3.12) is given by

$$F_Y(x) = 1 - (1 + \mu x e^{\mu t}) e^{-\mu x}, \quad x \geq 0 \tag{A.4}$$

and repeated binomial expansions show that

$$\begin{aligned} F_Y(x)^K &= \sum_{r=0}^{K} \binom{K}{r} (-1)^r (1 + \mu x e^{\mu t})^r e^{-r\mu x} \\ &= \sum_{r=0}^{K} \binom{K}{r} (-1)^r e^{-r\mu x} \sum_{m=0}^{r} \binom{r}{m} (\mu x e^{\mu t})^m, \quad x \geq 0. \end{aligned} \tag{A.5}$$

Upon substitution of (A.5) into (3.14), we now find

$$\begin{aligned}\overline{\psi}(\{t\}) &= \int_0^\infty \left(1 - \sum_{r=0}^{K} \binom{K}{r}(-1)^r e^{-r\mu x} \sum_{m=0}^{r} \binom{r}{m}(\mu x e^{\mu t})^m\right) dx \\ &= \int_0^\infty \sum_{r=1}^{K} \binom{K}{r}(-1)^{r-1} e^{-r\mu x} \sum_{m=0}^{r} \binom{r}{m}(\mu x e^{\mu t})^m dx \\ &= \sum_{r=1}^{K} \binom{K}{r}(-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m}(\mu e^{\mu t})^m \int_0^\infty x^m e^{-r\mu x} dx. \qquad (A.6)\end{aligned}$$

Using the well–known identity

$$\int_0^\infty x^m e^{-x} dx = m! \qquad m = 0, 1 \ldots (A.7)$$

for the moments of an exponential distribution, we conclude from (A.6) that

$$\overline{\psi}(\{t\}) = \frac{1}{\mu}\sum_{r=1}^{K} \binom{K}{r}(-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} e^{m\mu t}\frac{m!}{r^{m+1}}, \quad t < 0. \qquad (A.8)$$

Finally, reporting (A.8) in (3.11), we get

$$\begin{aligned}\overline{T}'_K(0) &= \frac{1}{\mu}\int_{-\infty}^{0} \left(\sum_{r=1}^{K} \binom{K}{r}(-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} e^{m\mu t}\frac{m!}{r^{m+1}} - H_K\right) dt \\ &= \frac{1}{\mu}\int_{-\infty}^{0} \sum_{r=1}^{K} \binom{K}{r}(-1)^{r-1} \sum_{m=1}^{r} \binom{r}{m} e^{m\mu t}\frac{m!}{r^{m+1}} dt \\ &= \frac{1}{\mu^2}\sum_{r=1}^{K} \binom{K}{r}(-1)^{r-1} \sum_{m=1}^{r} \binom{r}{m} \frac{(m-1)!}{r^{m+1}}. \qquad (A.9)\end{aligned}$$

APPENDIX B – We derive a formula for $\overline{T}_K(0)$ when the inter–arrival times and the service times are distributed according to a two–stage Erlang distributions. Here, with B given by (5.10), we see by repeated binomial expansion that

$$\begin{aligned}B(x)^K &= [1 - (1 + 2\mu x)e^{-2\mu x}]^K, \quad x \geq 0 \\ &= \sum_{r=0}^{K} \binom{K}{r}(-1)^r (1 + 2\mu x)^r e^{-2r\mu x} \\ &= \sum_{r=0}^{K} \binom{K}{r}(-1)^r \sum_{m=0}^{r} \binom{r}{m}(2\mu x)^m e^{-2r\mu x}. \; K = 2, 3, \ldots (B.1)\end{aligned}$$

Therefore, invoking (3.6)–(3.7) (and the remark that follows), we get

$$\begin{aligned}\overline{T}_K(0) &= \int_0^\infty \left(1 - \sum_{r=0}^{K} \binom{K}{r} (-1)^r \sum_{m=0}^{r} \binom{r}{m} (2\mu x)^m e^{-2r\mu x}\right) dx \\ &= \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} (2\mu)^m \int_0^\infty x^m e^{-2r\mu x} dx \\ &= \frac{1}{\mu} \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} \frac{m!}{(2r)^{m+1}} \qquad K = 2, 3, \ldots (B.2)\end{aligned}$$

upon using (A.7) in the last step, and (5.12)–(5.13) is obtained.

APPENDIX C – We obtain a formula for $\overline{T}'_K(0)$ when the arrivals are Poisson and the service times are hyper–exponential, with B given by (6.10). Following the development of Section III.2, we fix $t < 0$ and observe by simple computations that

$$\begin{aligned}F_Y(x) = 1 &- p_1 e^{-\mu_1 x} - p_2 e^{-\mu_2 x} - p_1^2 \mu_1 x e^{\mu_1 (t-x)} - p_2^2 \mu_2 x e^{\mu_2 (t-x)} \\ &+ \frac{p_1 p_2}{\mu_2 - \mu_1} (e^{-\mu_1 x} - e^{-\mu_2 x})(\mu_2 e^{\mu_1 t} + \mu_1 e^{\mu_2 t}), \quad x \geq 0. \qquad (C.1)\end{aligned}$$

Substituting into (3.11), we obtain after some tedious calculations that

$$\overline{T}'_K(0) = -\int_{-\infty}^{0} \left(\int_0^\infty \sum_{r=1}^{K} \binom{K}{r} U^{K-r} V^r dx\right) dt \qquad (C.2)$$

where for $x \geq 0$ and $t \geq 0$, we have set

$$U = 1 - p_1 e^{-\mu_1 x} - p_2 e^{-\mu_2 x}$$

and

$$V = -p_1^2 \mu_1 x e^{\mu_1 (t-x)} - p_2^2 \mu_2 x e^{\mu_2 (t-x)} - \frac{p_1 p_2}{\mu_2 - \mu_1} (e^{-\mu_1 x} - e^{-\mu_2 x})(\mu_2 e^{\mu_1 t} + \mu_1 e^{\mu_2 t}).$$

More tedious computations show that

$$U^{K-r} = \sum_{m_1=0}^{K-r} \binom{K-r}{m_1} (-1)^{m_1} \sum_{m_2=0}^{m_1} \binom{m_1}{m_2} p_1^{m_2} p_2^{m_1 - m_2} e^{(\mu_1 m_2 + \mu_2 (m_1 - m_2))x} \qquad (C.3)$$

and

$$\begin{aligned}V^r =& (-1)^r \sum_{k_1=0}^{r} \binom{r}{k_1} (\frac{p_1 p_2}{\mu_2 - \mu_1})^{r-k_1} \sum_{k_2=0}^{k_1} \binom{k_1}{k_2} (p_1^2 \mu_1)^{k_2} (p_2^2 \mu_2)^{k_1 - k_2} \\ &\times \sum_{k_3=0}^{r-k_1} \binom{r-k_1}{k_3} (-1)^{r-k_1-k_3} \sum_{k_4=0}^{r-k_1} \binom{r-k_1}{k_4} \mu_2^{k_4} \mu_1^{r-k_1-k_4} \\ &\times x^{k_1} e^{(\mu_1 (k_2+k_3) + \mu_2 (r-k_2-k_3))x} e^{(\mu_1 (k_2+k_4) + \mu_2 (r-k_2-k_4))t} \qquad (C.4)\end{aligned}$$

Substituting (C.3)–(C.4) into (C.2) we readily get the result (6.12).

APPENDIX D – We outline the derivation of a formula for $\overline{T}'_K(0)$ when the arrivals are Poisson and service times are distributed according to a two–stage Erlang, in which case B is given by (5.10). Fixing $t < 0$, we see that the common ditsibution F_Y of the rvs defined by (3.2) is this time given by

$$F_Y(x) = 1 - (1 + 2\mu x)e^{-2\mu x} - \left[(2\mu^2 - 4\mu^3 t)x^2 + \frac{4}{3}\mu^3 x^3\right] e^{2\mu t} e^{-2\mu x}, \quad x \geq 0, \quad (D.1)$$

We can show after some tedious calculations that

$$\overline{T}'_K(0) = -\int_{t=-\infty}^{0} \int_{x=0}^{\infty} \sum_{r=1}^{K} \binom{K}{r} U^{K-r} V^r dxdt \quad (D.2)$$

where we have set

$$U = 1 - (1 + 2\mu x)e^{-2\mu x}, \quad x \geq 0$$

and

$$V = -[(2\mu^2 - 4\mu^3 t)x^2 + \frac{4}{3}\mu^3 x^3]e^{2\mu t} e^{-2\mu x}, \quad x \geq 0,\ t \geq 0.$$

Through some more tedious computations, we can show that

$$U^{K-r} = \sum_{q=0}^{K-r} \binom{K-r}{q} (-1)^q \sum_{m=0}^{q} \binom{q}{m} (2\mu x)^m e^{-2q\mu x} \quad (D.3)$$

and

$$V^r = (-1)^r \sum_{n=0}^{r} \binom{r}{n} (2\mu^2 x^2)^n (\frac{4}{3}\mu^3 x^3)^{r-n} \sum_{p=0}^{n} \binom{n}{p} (-1)^p (2\mu t)^p e^{2r\mu t} e^{-2r\mu x}. \quad (D.4)$$

Combining (D.2)–(D.4), we obtain

$$\begin{aligned}\overline{T}'_K(0) =& \frac{1}{\mu^2} \sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{2^{r+1}} \sum_{q=0}^{K-r} \binom{K-r}{q} (-1)^q \sum_{m=0}^{q} \binom{q}{m} \\ & \sum_{n=0}^{r} \binom{r}{n} \frac{1}{3^{r-n}} \frac{(m+3r-n)!}{(r+q)^{m+3r-n+1}} \sum_{p=0}^{n} \binom{n}{p} \frac{p!}{(2r)^{p+1}} \end{aligned} \quad (C8)$$

which can be written as (6.22)–(6.23).

APPENDIX E – The various constants which were derived during the course of the discussion are tabulated below.

K	H_K	V_K	F_K	G_K	D_K	E_K
2	1.5	1.375	1.375	0.957	1.702	9.82
3	1.833	1.594	1.606	1.072	2.296	13.98
4	2.083	1.745	1.773	1.151	2.830	17.84
5	2.283	1.860	1.904	1.210	3.324	21.44
6	2.449	1.951	2.011	1.258	3.789	24.82
7	2.593	2.027	2.101	1.297	4.230	27.99
8	2.717	2.091	2.180	1.330	4.652	30.98
9	2.829	2.147	2.249	1.359	5.057	33.80
10	2.929	2.195	2.313	1.384	5.447	36.46
11	3.019	2.240	2.367	1.407	5.825	38.98
12	3.103	2.280	2.418	1.427	6.191	41.37
13	3.180	2.316	2.465	1.446	6.546	43.63
14	3.251	2.349	2.508	1.463	6.891	45.77
15	3.318	2.379	2.549	1.474	7.227	47.80
16	3.380	2.408	2.587	1.494	7.554	49.73
17	3.439	2.434	2.622	1.507	7.873	51.57
18	3.495	2.460	2.658	1.520	8.184	53.31
19	3.547	2.478	2.688	1.532	8.488	54.97
20	3.597	2.510	2.734	1.547	8.788	56.55