COOPERATIVE EFFECTS IN
MULTI-SERVER QUEUEING SYSTEMS

GURAMI TSITSIASHVILI,* Russian Academy of Sciences

Abstract
We discuss the cooperative effect of \( n \) single-server queues aggregated into a single queueing system. The performances of two such systems, with and without competition, are compared.

Keywords: Single-server queue; multiserver queue; competition; cooperative effect; performance

2000 Mathematics Subject Classification: Primary 60K25
Secondary 60J27

1. Introduction
This article considers cooperative effects in multiserver queueing systems. These effects are the results of interaction between subsystems aggregated into a single queueing system, and depend on the method of aggregation. This can be with or without competition, and is measured by objective functions such as mean stationary waiting times or queue lengths, the ability to handle customers in an aggregated queueing system, and so on. The cooperative effects depend strongly on the distribution functions (DFs) of known stochastic models, using exponential DFs or DFs from other classes. Important properties of the cooperative effects are jumps in their objective functions at critical points of their parameters. These phenomena may be interpreted as analogues of phase transitions in multiparticle physical systems (see [6]).

2. Cooperative effects in discrete queueing models
In this section we investigate the effects of an aggregation of queueing systems, described by discrete Markov processes. These models look simple, but the investigation of the cooperative effects in them is new, and leads to new results. We aggregate identical single-server queueing systems \( \text{M/M/1/\infty} \) into a multiserver queueing system with a stationary queue length and a waiting time, both of which are much smaller than the analogous characteristics of their components. We investigate quantitatively how the particular aggregation influences the objective functions of the final queueing systems.

Denote by \( X_1, X_2, \ldots \) the sequence of independent queueing systems \( \text{M/M/1/\infty} \), with input and service intensities \( \lambda \) and \( \mu \) respectively, which satisfy the following ergodicity condition:

\[
\rho = \frac{\lambda}{\mu} < 1.
\]
Consider the aggregation of the \( n \) independent single-server queueing systems, \( X_1, \ldots, X_n \), into the multiserver queueing system \( S_n \), organized as \( M/M/n/\infty \) and represented in Figure 1.

Let us write \( \psi = 1 - \rho \), \( C = e^{1/12}/\sqrt{2\pi} \), \( \overline{\psi} = n\psi \), and denote by \( A_n \) and \( M_n \) the mean stationary waiting time and queue length respectively.

**Theorem 1.** (a) If \( \rho \) is a constant and satisfies (1) then

\[
A_n \leq C \exp\left(-n\psi^2/(2-\psi)\right). 
\]

(b) If \( \rho = \rho(n) \), \( 0 < \psi < a < 1 \), then

\[
\frac{1}{\mu\overline{\psi}(n)/(1 + C(\overline{\psi}(n)/\sqrt{n})\exp(\overline{\psi}^2(n)/(2n(1-a))))} \leq A_n \leq \frac{1}{\mu(1-a)\overline{\psi}(n)}. 
\]

**Proof.** (a) It is well known (see, for example, [4]) that

\[
A_n = \frac{\pi_0}{n!} \left( \frac{\lambda n}{\mu} \right)^n \frac{n}{\mu(n - \lambda n/\mu)^2},
\]

where

\[
\pi_0 = \left( \sum_{k=0}^{n} \frac{1}{k!} \left( \frac{\lambda n}{\mu} \right)^k + \sum_{k=n+1}^{\infty} \frac{1}{k!} \left( \frac{\lambda n}{\mu} \right)^k \right)^{-1}.
\]

We now find the upper bound of the quantity \( A_n \). It is clear that

\[
\pi_0^{-1} \geq \sum_{k=0}^{n} \frac{1}{k!} \left( \frac{\lambda n}{\mu} \right)^k + \sum_{k>n}^{\infty} \frac{1}{k!} \left( \frac{\lambda n}{\mu} \right)^k = e^{\lambda n/\mu}.
\]

From (4) and the Stirling formula, we obtain

\[
A_n \leq e^{-n\rho} \rho^n \frac{n^{n-1}}{n!} \frac{\exp(-n\psi(\psi))}{\psi^2 n^{3/2}}, \quad n \geq 1,
\]
where \( \varphi(\psi) = -\psi - \ln(1 - \psi) \). The Taylor formula for \( \psi \in (0, 1) \) gives us
\[
\varphi(\psi) = -\psi + \frac{\psi^2}{2} + \frac{\psi^2}{3} + \cdots \geq \frac{\psi^2}{2} \sum_{k \geq 0} \left( \frac{\psi}{2} \right)^k = \frac{\psi^2}{2 - \psi}.
\]
(6)

Combining the inequalities (6) and (5), we obtain (2).

(b) Since the inequality
\[
\pi_0^{-1} \geq \frac{n^n}{n!} \sum_{k > n} \rho^k = \rho \frac{n^n \rho^n}{n! \psi}
\]
is true, it follows that
\[
A_n \leq \frac{1}{\mu(1 - a) \psi}.
\]
We now find the lower bound of the quantity \( A_n \). Since
\[
\pi_0^{-1} = \sum_{k=0}^{n} \frac{n^k}{k!} \frac{\rho^k}{n!} \sum_{k>n} \rho^k \leq e^n \rho + \frac{n^n (1 - \psi)^n}{n! \psi},
\]
we see from the Stirling formula that
\[
A_n \geq \frac{1}{\mu \overline{\psi}(n)(1 + \beta(n))},
\]
where
\[
\beta(n) = \overline{C} \psi \frac{e^{-n \psi}}{(1 - \psi)^n} \sqrt{n}
\]
\[
\leq \overline{C} \sqrt{n} \exp \left( n \frac{\psi^2}{2(1 - a)} \right)
\]
\[
= \overline{C} \frac{\psi}{\sqrt{n}} \exp \left( \frac{\psi^2}{2n(1 - a)} \right).
\]
Thus, the two-sided bounds of (3) are proved.

**Corollary 1.** Suppose that \( \psi = \psi(n) = 1 - \rho(n) \sim n^{-\alpha} \), then, for \( \alpha < 1 \), \( A_n \to 0 \) as \( n \to \infty \) and, for \( \alpha > 1 \), \( A_n \to \infty \) as \( n \to \infty \).

**Corollary 2.** Suppose that \( \psi = \psi(n) = 1 - \rho(n) \sim n^{-\alpha} \), then, for \( \alpha < \frac{1}{2} \), \( M_n \to 0 \) as \( n \to \infty \) and, for \( \alpha > \frac{1}{2} \), \( M_n \to \infty \) as \( n \to \infty \).

**Proof of Corollary 1 and Corollary 2.** It is well known (see, for example, [4]) that
\[
M_n = \frac{\pi_0^n n^\rho n^{\rho+1}}{n!(1 - \rho)^2}.
\]
Then the statement of Corollary 2 may be obtained in the same way as Corollary 1.
Remark 1. A characteristic feature of many known queueing systems is an increase of the mean stationary waiting time and the mean stationary queue length for limiting values of the load coefficient. In our analysis, if the number \( n \) of the systems M/M/1/\( \infty \), aggregated into the system \( S_n \), tends to infinity, then a new phenomenon occurs, which is analogous to the phase transition in physical systems.

3. Queueing systems with competition between servers

Queueing systems with competition between servers or customers are widely used in modern data transmission and mobile telephone networks, see [1]. But there is no analytical investigation of the influence of competition on the queueing systems’ characteristics. In this section, a mathematical model of a multiserver queueing system with competition between servers is constructed. This system is compared with the classical multiserver queueing system, in terms of their abilities to handle customers, and the distribution tails of their stationary waiting times.

Consider the multiserver queueing system M/G/\( m \)/\( \infty \) with \( m \) servers, input flow \( 0 = t_1 \leq t_2 = t_1 + \xi_1 \leq t_3 = t_2 + \xi_2 \leq \cdots \), and service times \( \eta_1, \eta_2, \ldots \). The random sequences \( \{\eta_1, \eta_2, \ldots\} \) and \( \{\xi_1, \xi_2, \ldots\} \) are independent and consist of independent random variables with DFs \( G(t) \) and \( H(t) \) respectively, on \( [0, \infty) \), and with the tails

\[
P(\eta_k > t) = 1 - G(t) = G(t) \quad \text{and} \quad P(\xi_k > t) = \exp(-\lambda t), \quad k \geq 0,
\]

respectively. Denote this system by \( A_m \), and suppose that it is empty at the moment \( t_1 \) of the first customer’s arrival.

On the basis of the system \( A_m \), define the multiserver queueing system \( B_m \) with competition between servers as follows. Suppose that the first customer arrives in the empty system \( B_m \) at the moment \( t_1 \), and asks all \( m \) servers how long it would take to be served. The customer receives information on its possible service times, \( \eta_1^{(1)}, \ldots, \eta_1^{(m)} \), at each of the \( m \) servers. Then it chooses the server with the minimal service time, i.e. \( \zeta_1 = \min(\eta_1^{(1)}, \ldots, \eta_1^{(m)}) \). During the first customer’s service, all other servers do not work. The second customer arrives into the system \( B_m \) at the moment \( t_2 \) and receives information on its possible service times, \( \eta_2^{(1)}, \ldots, \eta_2^{(m)} \), and so on. The random variables \( \eta_i^{(j)}, \ j = 1, \ldots, m, i \geq 1 \), are independent and have the common DF \( G(t) \). So the system \( B_m \) may be considered as the single-server queueing system M/G/1/\( \infty \), with independent and identically distributed (i.i.d.) service times \( \zeta_1, \zeta_2, \ldots \), such that

\[
P(\zeta_1 > x) = G^m(x), \quad x \geq 0.
\]

Suppose that \( w_n^A \) and \( w_n^B \), \( n \geq 0 \), are the waiting times of the \( n \)th customer in the systems \( A_m \) and \( B_m \) respectively. Denote by

\[
\overline{W}^A_m(x) = \lim_{n \to \infty} P(w_n^A > x) \quad \text{and} \quad \overline{W}^B_m(x) = \lim_{n \to \infty} P(w_n^B > x)
\]

the limiting tail distributions of the waiting times in the systems \( A_m \) and \( B_m \) respectively. Fix the DF \( G(t) \), and define by

\[
\lambda_m^A = \sup \{ \lambda : \lim_{x \to \infty} \overline{W}^A_m(x) = 0 \} \quad \text{and} \quad \lambda_m^B = \sup \{ \lambda : \lim_{x \to \infty} \overline{W}^B_m(x) = 0 \}
\]
Cooperative effects in multi-server queueing systems

the maximal abilities of the systems $A_m$ and $B_m$ respectively to handle customers. Our problem is to discuss the asymptotic analysis for $m \to \infty$ of $K_m = \lambda_m A / \lambda_m B$, and to analyse the asymptotics of the function $W_m^{B}(x), x \to \infty$, when the DF $G(x)$ is regularly varying.

To do this, we introduce some notation and list necessary information. Suppose that $\mathcal{L}$ is the class of slowly varying functions, define

$$\mathcal{R}(a) = \{x^a l(x), l(x) \in \mathcal{L}, -\infty < a < \infty\}.$$ In [3] and [7], the following Karamata theorem is proved: suppose that $l(x) \in \mathcal{L}$, then, for $a < -1$,

$$\int_{x}^{\infty} t^a l(t) \, dt \sim -(a + 1)^{-1} x^{a+1} l(x), \quad x \to \infty.$$ For the system $M/G/1/\infty$ with Poisson input flow (with intensity $\lambda$), if the condition $G \in \mathcal{R}(a)$ and the equality

$$G_{I}(x) = \int_{x}^{\infty} G(y) \, dy$$

hold, then the Embrechts–Veraverbeke formula

$$W(x) \sim \frac{\lambda}{1 - \lambda M \eta_1} G_{I}(x), \quad x \to \infty,$$

(see [2]) holds.

Consider now the Kiefer–Wolfowitz chain, see [5], $(w_{n,1}, w_{n,2}, \ldots, w_{n,m})$, $n \geq 0$, describing the multiserver queueing system $A_m$. Here $w_{n,i}$ is the interval between the moment $t_n$ and the moment when $i$ servers become free of the 1st, \ldots, ($n-1$)th customers of input flow. A necessary and sufficient condition of the Kiefer–Wolfowitz chain ergodicity (and hence the existence of the limit distribution $\lim_{n \to \infty} P(w_{n,1} > t)$) is the inequality

$$M \eta_1 < m \lambda. \quad (7)$$

So the formula

$$\gamma_m^A = \frac{m}{\int_{0}^{\infty} G(t) \, dt}$$

holds, and, according to the equality (8), we obtain

$$K_m = \frac{m \int_{0}^{\infty} G_m(t) \, dt}{\int_{0}^{\infty} G(t) \, dt}.$$ The following theorem is proved in [8].

**Theorem 2.** Suppose that, in the system $A_m$ for some $a > 1$, the DF $G(x) \in \mathcal{R}(-a)$ and the ergodicity condition (7) is true, then the function $W_m^{A}(x) \in \mathcal{R}(-ma + m)$.

**Theorem 3.** Suppose that, for some $a > 0$ and $b > 0$,

$$G(x) \sim ax^b, \quad x \to 0,$$

(9)

then, in the system $B_m$ for $m \to \infty$, the following is true:

$$if \ 0 < b < 1 \ then \ K_m = O(m^{1-1/b}), \quad (10)$$

$$if \ b \geq 1 \ then \ \frac{1}{K_m} = O\left(\frac{1}{m^{1-1/b}}\right), \quad (11)$$
Proof. Suppose that $0 < b < 1$ and $a > 0$ are fixed. Then there exist positive numbers $\alpha, m_0$ such that, for $0 < t < 1/m_0$, the inequality $G(t) \geq \alpha t^b$ holds, and so

$$G(t) \leq \exp(-\alpha t^b).$$

Suppose that $m > m_0$, and denote

$$J_m = \int_0^\infty mG^m(t) \, dt = U_m + V_m,$$

$$U_m = \int_0^{1/m} mG^m(t) \, dt,$$

$$V_m = \int_{1/m}^\infty mG^m(t) \, dt.$$

We estimate the quantity $U_m$, for $m \to \infty$, as

$$U_m \leq \int_0^{1/m} m \exp(-amt^b) \, dt$$

$$= \int_0^{1/m} m \exp(-((am)^{1/b})) \frac{dt}{(am)^{1/b}}$$

$$= \frac{m}{(am)^{1/b}} \int_0^{(am)^{1/b}/m} \exp(-v^b) \, dv$$

$$\sim \frac{m}{(am)^{1/b}} \int_0^\infty \exp(-v^b) \, dv$$

$$= O(m^{1-1/b}).$$

and the quantity $V_m$, for $m \to \infty$, as

$$V_m \leq mG^{m-1} \left( \frac{1}{m} \right) \int_{1/m}^\infty G(t) \, dt$$

$$\leq m \exp(-\alpha m - 1)m^{-b} \int_0^\infty G(t) \, dt$$

$$= O(m^{1-1/b}).$$

Then $J_m = O(m^{1-1/b})$ for $m \to \infty$, and

$$K_m = \frac{J_m}{J_1} = O(m^{1-1/b}), \quad \text{as } m \to \infty.$$

This proves (10).

Suppose that $a > 0$ and $b \geq 1$ are fixed. Then there exist positive numbers $\alpha, m_0$ such that, for $0 < t < 1/m_0$, the inequality $G(t) \leq \alpha t^b$ holds, and so $G(t) \geq 1 - \alpha t^b$. Putting $m > m_0$, we obtain

$$J_m \geq \int_0^{1/m^{1/b}} m \left(1 - \frac{\alpha}{m} \right)^m \, dt \sim m^{1-1/b} \exp(-\alpha), \quad m \to \infty.$$
and
\[ \frac{1}{K_m} = \frac{J_1}{J_m} = O\left(\frac{1}{m^{1-1/b}}\right), \quad m \to \infty. \]
This proves (11).

**Corollary 3.** Theorem 3 shows how the ratio of the abilities to handle customers in the systems \( A_m \) (without competition) and \( B_m \) (with competition) depends on the parameter \( b \).

**Example 1.** Consider the following two DFs \( G(x) \), \( x \geq 0 \), satisfying (9).

(a) The Weibull distribution
\[ \overline{G}(x) = \exp(-ax^b), \quad 0 < b < 1. \]

(b) The Burr distribution
\[ \overline{G}(x) = (1 + cx^b)^{-a/c}, \quad 0 < a, 0 < b \leq 1, 0 < c < ab. \]

**Theorem 4.** Suppose that, in the system \( B_m \), the DF \( G(x) \in \mathcal{R}(-a) \), \( a > 1 \), and the ergodicity condition \( \lambda b_m < 1 \) with \( b_m = \int_0^\infty \overline{G^m}(x) \, dx \) is true. Then
\[ \overline{W}_m^B(x) \sim \frac{\lambda x \overline{G^m}(x)}{(1 - \lambda b_m)(ma - 1)}, \quad x \to \infty, \tag{12} \]
and, consequently, \( \overline{W}_m^B(x) \in \mathcal{R}(-ma + 1) \).

**Proof.** As the DF \( G(t) \in \mathcal{R}(-a) \), \( a > 1 \), it follows that, for some function \( l(t) \in \mathcal{L} \) (and, consequently, \( l^m(t) \in \mathcal{L} \), \( \overline{G^m}(t) = l^m(t)^{-ma} \)). Using the Karamata theorem for \( x \to \infty \), we obtain
\[ \frac{1}{b_m} \int_x^\infty \overline{G^m}(t) \, dt \sim \frac{l^m(x)x^{-ma+1}}{b_m(ma - 1)} = \frac{x \overline{G^m}(x)}{b_m(ma - 1)}. \tag{13} \]
Then, the DF
\[ 1 - \frac{\int_x^\infty \overline{G^m}(t) \, dt}{b_m} \]
is subexponential, and, from the ergodicity condition \( \lambda b_m < 1 \), we obtain the Embrechts–Veraverbeke formula
\[ \overline{W}_m^B(x) \sim \frac{\lambda b_m \int_x^\infty \overline{G^m}(t) \, dt}{b_m(1 - \lambda b_m)}, \quad x \to \infty. \tag{14} \]
Substituting (13) into (14), we obtain (12).

**Remark 2.** Comparison of Theorems 2 and 4 shows that, in the system \( B_m \) with competition, the tail of the stationary waiting-time distribution is lighter than in the system \( A_m \) without competition.

**Acknowledgement**

This work is supported by Russian Foundation for Basic Research, project 03-01-00512.
References


