# Heavy Traffic Limit for a Mobile Phone System Loss Model

Philip J. Fleming and Alexander Stolyar
Motorola, Inc.
Arlington Heights, IL

Burton Simon
Department of Mathematics
University of Colorado at Denver
Denver, CO

## ABSTRACT

An individual cell in a mobile phone system sees two types of requests for service. There are "call setups" which occur when a customer in the cell places or receives a call, and there are "handovers" which occur when a customer with an existing call moves into the cell. Customers that are blocked will "retry" for a while, but will give up if they are not successful after some finite (random) time.

If the two customer types are statistically indistinguishable, the model reduces to an "Erlang-c system with impatient customers". We show that the heavy traffic limit of the stationary distribution for that model is a concatination of two normal densities. In general, our model is two dimensional. However, we show that in heavy traffic it reduces to a one dimensional process, whose stationary distribution is also a concatination of two normal densities. Using the heavy traffic limit we are able to approximate blocking probabilities, dropped call probabilities, average utilization and other performance measures with closed form expressions.

## 1. Introduction.

Consider a single cell in a mobile phone system. A "call setup" is a request for a channel by an idle customer presently in the cell that is either placing or receiveing a call. A "handover" is a request for a channel by an active customer moving into the cell from a neighboring cell. We assume that every request is granted a channel if one is available. If a request is made for a channel when they are all busy the customer is "blocked".

Blocked customers will not give up immediately, nor will they persist indefinitely. When a call setup is blocked it will periodically "retry" (e.g. via the resend button) for some (random) amount of time until it either secures a channel, gives up trying, or leaves the cell. Likewise, a handover that is blocked will retry for some (random) amount of time until it either secures a channel, gets "dropped" by losing communication with the original cell, finishes the call, or leaves the cell. A handover that is dropped will turn into a call setup if the customer wants to continue the conversation.

We propose a model of this system and analyze its stationary distribution in heavy traffic, i.e. we analyze systems with a large number of channels and a traffic load approximately equal to the number of channels.

Even though there are two customer types, the stationary distribution collapses to one dimension in heavy traffic. This allows us to derive closed form approximations for stationary blocking probablities, dropped call rates, system utilization and other performance measures. We are also able to conjecture that the heavy traffic limit of the underlying stochastic process is a certain one dimensional diffusion with piecewise linear drift and constant variance parameters.

There has been considerable work done on blocking systems in heavy traffic starting with Borovkov [1]. Whitt [4] provides approximations similar in spirit to ours, and includes a good survey of published work in the area. Our model differs from previous work in two ways. First, the heavy traffic analysis of a blocking system with "impatient" customers is new. The technically difficult innovation, however, was going from a one to two dimensional state space.

In the next section we describe the model in detail. In Section 3 we analyze the Erlang-c system with impatient customers in heavy traffic. We rove that the stationary density is a concatination of two normal densities, and conjecture the

form of the limiting stochastic process. In Section 4 we analyze the full model in heavy traffic. We prove that in heavy traffic the two dimensional process collapses to one dimension, and that the stationary density has the same form as the Erlang-c system with impatient customers. We conjecture that the stochastic process collapses to the corresponding one dimensional diffusion. In Section 5 we derive approximations for the performance measures based on the heavy traffic limit. The technical details of the proofs of the major theorems can be found in the appendix.

## 2. Model Description.

Our cell has $N$ channels. Call setups (type 1 customers) arrive as a Poisson process with rate $\lambda_1$ and handovers (type 2 customers) arrive as a Poisson process with rate $\lambda_2$. Let $\lambda = \lambda_1 + \lambda_2$. Both type 1 and type 2 customers have service times that are exponentially distributed with mean $1/\mu$. Let $\rho = \lambda/\mu$ and let $\rho_i = \lambda_i/\mu,\ i = 1, 2$.

If a customer requests a channel when all $N$ are busy it will retry until it either secures a channel or disappears. The time until a type $i$ customer disappears (given it does not secure a channel) is exponentially distributed with mean $1/\beta_i$. The interarrival times, service times and maximal waiting times are i.i.d. sequences and are independent of each other.

Let $Q(t)$ be the number of customers in the system at time $t$. If $Q(t) > N$ then there are blocked customers waiting. Let $H_i(t)$ be the number of type $i$ customers waiting, and let $H(t) = H_1(t) + H_2(t) \equiv [Q(t) - N]^+$. Suppose a channel becomes free at time $t$ when $H(t) > 0$. We assume that

(a) one of the waiting customers will seize the free channel immediately, and

(b) the probability that a type $i$ customer will get the channel has the form

$$P(\text{type } i \text{ gets the channel} \mid H_1 = k_1, H_2 = k_2) = \frac{k_i \eta_i}{k_1\eta_1 + k_2\eta_2}, \qquad (2.1)$$

where $\eta_i > 0$.

If the points in time that a blocked type $i$ customers retries is a Poison process with rate $\eta_i$ (until it succeeds or gives up) then assumptions (a) and (b) hold exactly as $\eta_1 + \eta_2 \to \infty$. We therefore interpret $\eta_i$ as the retry rate for type $i$ customers.

When a waiting type 2 customer gives up (is "dropped"), it immediately becomes a type 1 customer with probability $p$. With probability $1 - p$ a waiting type

2 customer disappears after being dropped. All waiting type 1 customers disappear after giving up.

The following quantities are of particular interest. Let

$$B = \lim_{t\to\infty} P(Q(t) \geq N) \tag{2.2}$$

be the limiting blocking probability, let

$$\beta_i^* = \beta_i \lim_{t\to\infty} E(H_i(t)), \quad i = 1, 2 \tag{2.3}$$

be the limiting rate that type $i$ customers are lost due to impatience, and let

$$\rho^* = \lim_{t\to\infty} E(Q(t) - H(t)) \tag{2.4}$$

be the limiting average number of busy channels.

## 3. The Erlang-c System with Impatient Customers in Heavy Traffic.

The Erlang-c blocking system is an M/M/N/$\infty$ queue. A customer is temporarily "blocked" if it has to wait for a server. Typically, an Erlang-c model is used to correct the "flaw" in the Erlang-b model, where customers "disappear" when they are blocked. In many applications blocked customers do not disappear immediately. However, the Erlang-c model overcompensates in many cases. Customers are not infinitely patient. For one reason or another they will give up (and disappear) after some finite (random) time. To model this phenomenon we use an Erlang-c model where customers disappear if they are forced to wait more than an exponentially distributed time.

Let $\lambda$ be the arrival rate, $\mu$ the service rate, and suppose that waiting customers will disappear after an exponentially distributed time with mean $1/\beta$ if they do not get service. Let $\rho = \lambda/\mu$ and let $Q^{(\rho)}(t)$ be the number of customers in the system at time $t$. Clearly, $Q^{(\rho)}(t)$ is a birth-death process. The transition diagram for $Q^{(\rho)}(t)$ is shown in Figure 1.

It follows that the stationary distribution for $Q^{(\rho)}$ is

$$\pi^{(\rho)}(k) = \begin{cases} C_\rho \frac{\rho^k}{k!} & \text{if } k \leq N, \\ C_\rho \frac{\rho^N}{N!} \frac{\lambda^{k-N}}{\prod_{i=1}^{k-N}(N\mu + i\beta)} & \text{if } k > N, \end{cases} \tag{3.1}$$

where

$$C_\rho = \left( \sum_{i=0}^{N} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \sum_{i=1}^{\infty} \frac{\lambda^i}{\prod_{j=1}^{i}(N\mu + j\beta)} \right)^{-1}. \tag{3.2}$$

We now consider a sequence of systems indexed by $\rho = 1, 2, \ldots$, where the number of channels, $N^{(\rho)}$, increases with $\rho$. Without loss of generality we take $\mu = 1$ and as a consequence, $\rho = \lambda$, for the remainder of the paper.

THEOREM 3.1. *Suppose the number of servers, $N^{(\rho)}$ increases with $\rho$ so that*

$$\lim_{\rho\to\infty} \frac{N^{(\rho)} - \rho}{\sqrt{\rho}} \to n \tag{3.3}$$

*where $-\infty < n < \infty$. Then the stationary distribution $F^{(\rho)}$ of the scaled process*

$$q^{(\rho)} = \frac{Q^{(\rho)} - \rho}{\sqrt{\rho}} \tag{3.4}$$

*converges weakly to a distribution with density*

$$\pi(x) \equiv \begin{cases} \frac{c_1}{\sqrt{2\pi}} e^{-x^2/2} & \text{if } x \le n, \\ \frac{c_2}{b\sqrt{2\pi}} e^{-(x-a)^2/2b^2} & \text{if } x > n, \end{cases} \tag{3.5}$$

*and*

$$\lim_{\rho\to\infty} \sqrt{\rho}\pi^{(\rho)}(\lceil \rho + x\sqrt{\rho} \rceil) = \pi(x) \tag{3.6}$$

*where*

$$b^2 = \frac{\mu}{\beta}, \tag{3.7}$$

$$a = (1 - b^2)n, \tag{3.8}$$

$$c_1 = \left(\Phi(n) + e^{-na/2} b(1 - \Phi(nb))\right)^{-1} \tag{3.9}$$

$$c_2 = c_1 b e^{-na/2}, \tag{3.10}$$

*and $\Phi(x)$ is the CDF for the standard Normal.*

PROOF: We assume that

$$N \equiv N^{(\rho)} = \lceil \rho + n\sqrt{\rho}\ \rceil.$$

The state space of the process $Q^{(\rho)}(t)$ is $S = \{0, 1, 2, \ldots\}$. For each $\rho$ and corresponding $N^{(\rho)}$ we will consider two truncated state spaces

$$S_-^{(\rho)} = \{0, 1, 2, \ldots, N - 1\}$$

and

$$S_+^{(\rho)} = \{N, N+1, N+2, \ldots\}.$$

Consider the scaled process, $q^{(\rho)}(t)$ given by (3.4). The state space for $q^{(\rho)}(t)$ is therefore

$$s^{(\rho)} = \{k/\sqrt{\rho}, \quad k = \ldots, -2, -1, 0, 1, 2, \ldots\}. \tag{3.11}$$

For convenience, the state space $s^{(\rho)}$ is considered to be infinite from the left. We also break down $s^{(\rho)}$ into two subsets, which are the scaled sets corresponding to $S_-^{(\rho)}$ and $S_+^{(\rho)}$ respectively,

$$s_-^{(\rho)} = \{\sigma \in s^{(\rho)}, \sigma < n\}$$

and

$$s_+^{(\rho)} = \{\sigma \in s^{(\rho)}, \sigma \geq n\}.$$

Let $F^{(\rho)}(y), y \in \Re$, the real numbers, be the stationary distribution function of the scaled process $q^{(\rho)}(t)$, and $p^{(\rho)}(\sigma), \sigma \in s^{(\rho)}$, denote the stationary probability of the state $\sigma$. So

$$p^{(\rho)}(\sigma) \equiv \pi^{(\rho)}(\rho + \sigma\sqrt{\rho}). \tag{3.12}$$

(We will omit upper index $(\rho)$ where it will not cause confusion.)

To prove Theorem 3.1 we need the following results, which are proved in the Appendix.

LEMMA 3.1.1. *The sequence of distributions* $F^{(\rho)}, \quad \rho = 1, 2, \ldots,$ *is relatively compact.*

LEMMA 3.1.2. *For any points* $a_1 < a_2$, $a_1, a_2 \in \Re$, *consider sequences of states* $\sigma_1^{(\rho)} \to a_1$, $\sigma_2^{(\rho)} \to a_2$, *where* $\sigma_1^{(\rho)}, \sigma_2^{(\rho)} \in s^{(\rho)}$. *Then*

$$\lim_{\rho \to \infty} \ln \frac{p^{(\rho)}(\sigma_2^{(\rho)})}{p^{(\rho)}(\sigma_1^{(\rho)})} = -\int_{a_1}^{a_2} g(y)dy$$

*where*

$$g(y) = \begin{cases} y & y < n \\ n + \beta(y - n) & y \geq n. \end{cases}$$

Consider any limiting point (in the sense of weak convergence) $F$ of the sequence of distributions $\{F^{(\rho)}\}$. To simplify notation we will assume that $F^{(\rho)} \to F$. From Lemma 3.1.2 it follows that the distribution $F$ is absolutely continuous with density of the form

$$f(y) = \begin{cases} c_1\phi(y) & y < n \\ c_2\beta\phi(\beta(y-(\beta-1)n/\beta)) & y \geq n, \end{cases}$$

where $c_1$ and $c_2$ are non-negative constants and $\phi(y)$ is the standard normal density. From Lemma 3.1.2 it follows also that $f$ is continuous at $n$. Using the continuity at $n$ along with $\int f(y)dy = 1$ specifies $c_1$ and $c_2$. (It can be easily obtained from Lemma 3.1.2 that the convergence (3.6) is uniform in any finite interval $x \in [a_1, a_2]$.)

For many applications the heavy traffic limit of the stationary distribution of the process is sufficient. However, for estimating transient probabilities it is necessary to use the limit (in the sense of weak convergence) of the sequence of stochastic processes. (The weak convergence of stochstic processes will be denoted by $\Rightarrow$.) If one accepts that $q^{(\rho)}(t) \Rightarrow q(t)$ where $q(t)$ is a diffusion then it is a simple matter to derive the parameters of $q(t)$. We sum this up as

CONJECTURE 3.1. *Under the scaling (3.4), we have*

$$q^{(\rho)}(t) \Rightarrow q(t), \tag{3.13}$$

*where $q(t)$ is a diffusion process on $-\infty < x < \infty$ with drift function*

$$d(x) = \begin{cases} -\mu x & \text{if } x \leq n, \\ -\mu n - \beta(x-n) & \text{if } x > n, \end{cases} \tag{3.14}$$

*and variance function*

$$\sigma^2(x) = 2\mu, \qquad -\infty < x < \infty. \tag{3.15}$$

*The stationary distribution of $q$ is $\pi$ given by (3.5).*

An illustration of $\pi(x)$ is given in Figure 2.

## 4. Heavy Traffic Limit for the Mobile Phone System Model.

Our model for the mobile phone system is an Erlang-c with $N$ servers and two impatient customer types, corresponding to "call setups" and "handover requests". Each type has the same service rate, $\mu$, but their quitting rates, $\beta_1$ and $\beta_2$ may differ. When a type 2 customer quits, it becomes a type 1 customer with probability $p$. As before, we will assume that $\mu = 1$, and therefore $\lambda = \rho$.

Clearly, $Q(t)$ is a simple birth-death process when $Q(t) \leq N$. However, unless $\beta_1 = \beta_2$, knowledge of $H_i(t)$, $i = 1, 2$, (the number of waiting type $i$ customers) is necessary when $Q(t) > N$ for the process to be Markovian. The transition diagram for the process $(Q(t), \vec{H}(t))$ is given in Figure 3.

Suppose, $\rho \to \infty$ in such a way that ratios $\rho_i/\rho = \nu_i$ are held constant, $\nu_1 + \nu_2 = 1$. Although $Q^{(\rho)}(t)$ is not a Markov process, it does have a stationary distribution. And as $\rho \to \infty$ the stationary distributions of the scaled processes $q^{(\rho)}(t)$ converge to a limiting distribution, which is sufficient to compute most performance measures since (as we will show) in the limit $H_1(t)$ and $H_2(t)$ remain in a fixed ratio. We sum this up in the following theorem.

THEOREM 4.1. *Let $\rho \to \infty$, with the ratios $\rho_i/\rho = \nu_i$ fixed, and*

$$\lim_{\rho \to \infty} \frac{N^{(\rho)} - \rho}{\sqrt{\rho}} \to n.$$

*Then the stationary distribution $F^{(\rho)}$ of the scaled process*

$$q^{(\rho)}(t) = \frac{Q^{(\rho)}(t) - \rho}{\sqrt{\rho}}.$$

*converges weakly to a distribution with density $\pi(x)$ specified by (3.5)-(3.10), using*

$$\beta = \alpha_1 \beta_1 + (1 - p)\alpha_2 \beta_2, \tag{4.1}$$

*where*

$$\alpha_1 = \frac{\nu_1 \eta_2}{\nu_1 \eta_2 + \nu_2 \eta_1}, \quad \alpha_2 = \frac{\nu_2 \eta_1}{\nu_1 \eta_2 + \nu_2 \eta_1}. \tag{4.2}$$

*Furthermore, if $G^{(\rho)}$ is the conditional distribution of $(H_1^{(\rho)}, H_2^{(\rho)})$ under the condition $H^{(\rho)} > 0$, then for any $\epsilon > 0$,*

$$\lim_{\rho \to \infty} G^{(\rho)} \left( \{(x, y) : |\frac{x}{y} - \frac{\nu_1 \eta_2}{\nu_2 \eta_1}| < \epsilon\} \right) = 1, \tag{4.3}$$

*i.e. in heavy traffic we have* $\frac{H_1}{H_2} = \frac{\nu_1\eta_2}{\nu_2\eta_1}$ *whenever* $H(t) > 0$.

SKETCH OF THE PROOF: (The complete proof of the theorem and all supplementary results is in the appendix.)

Relative compactness of the sequence of distributions $F^{(\rho)}$ is shown the same way as in the one dimensional case (Lemma 4.1.1).

For any $\rho$, explicit expressions analogous to (3.1) and (3.2) can be written for the distribution $\pi^{(\rho)}$. The difference is that instead of the constant $\beta$ we use $\bar{\beta}_j^{(\rho)}$, where

$$\bar{\beta}_j^{(\rho)} = \beta_1\alpha_j^{(\rho)} + (1-\rho)\beta_2(1-\alpha_j^{(\rho)})$$

and $\alpha_j^{(\rho)}$ is the conditional mean

$$\alpha_j^{(\rho)} = E(\frac{H_1^{(\rho)}}{H^{(\rho)}} \mid H^{(\rho)} = j)$$

We prove (Theorem A.2 in the appendix) that as $\rho \to \infty$ the conditional ratio $H_1^{(\rho)}/H^{(\rho)}$ under the condition that $H^{(\rho)} > 0$, converges to the constant $\alpha_1$ given by (4.2). This allows us to prove that $\alpha_j^{(\rho)}$ as a function of $j$ also converges to the constant $\alpha_1$ (Lemma 4.1.4 in the appendix). The second statement in Theorem 4.1 also follows from Theorem A.2.

As in the previous section, if we accept that $q^{(\rho)}(t) \Rightarrow q(t)$, where $q(t)$ is a diffusion process, then it is a simple matter to derive its drift and variance parameters. We have

CONJECTURE 4.1. *Under the scaling (3.4) we have*

$$q^{(\rho)}(t) \Rightarrow q(t),$$

*where* $q(t)$ *is a one dimensional diffusion process on* $-\infty < x < \infty$ *with drift function* $d(x)$ *given by (3.14) and variance function* $\sigma^2(x)$ *given by (3.15), where* $\beta$ *is given by (4.1). The stationary distribution of* $q$ *is* $\pi$ *given by Theorem 4.1. Furthermore,*

$$\rho^{-\frac{1}{2}}(H_1^{(\rho)}(t), H_2^{(\rho)}(t)) \Rightarrow (h_1(t), h_2(t)),$$

*where*

$$(h_1(t), h_2(t)) = (\alpha_1, \alpha_2)h(t),$$

$\alpha_i$ *is given by (4.2), and*

$$h(t) = [q(t) - n]^+.$$

## 5. Approximations for the Performance Measures.

For the mobile phone system model described in Section 2, let

$$\beta = \alpha_1\beta_1 + \alpha_2(1-p)\beta_2,$$

where $\alpha_i$ is given by (4.2), and let $b^2$, $a$, $c_1$ and $c_2$ be given by (3.7)-(3.10). Let $\pi(x), x \in \Re$ be the probability density specified by (3.5). For large $\rho$ equation (3.6) yields

$$\sqrt{\rho}\pi^{(\rho)}(\lceil\ \rho + x\sqrt{\rho}\ \rceil) \approx \pi(x),$$

or equivalently, for large $t$ and large $\rho$,

$$P(Q^{(\rho)}(t) \in [k_1, k_2]) \approx \int_{\frac{k_1-\rho}{\sqrt{\rho}}}^{\frac{k_2-\rho}{\sqrt{\rho}}} \pi(x)dx.$$

In particular, the probability that a customer is blocked is

$$B^{(\rho)} = P(Q^{(\rho)} \geq N^{(\rho)}) \approx \int_n^\infty \pi(x)dx = c_2(1-\Phi(nb)), \tag{5.1}$$

where,

$$n = \frac{N^{(\rho)} - \rho}{\sqrt{\rho}}.$$

Similarly, we can write

$$\begin{aligned} E(Q^{(\rho)}) \approx \rho + \sqrt{\rho}E(q) =& \rho + \sqrt{\rho}\int_{-\infty}^{\infty} x\pi(x)dx \\ =& \rho + \sqrt{\rho}\left(c_2[b\phi(nb) + a(1-\Phi(nb))] - c_1\phi(n)\right), \end{aligned} \tag{5.2}$$

and

$$\begin{aligned} E(H^{(\rho)}) \approx \sqrt{\rho}E(h) =& \sqrt{\rho}\int_n^\infty (x-n)\pi(x)dx \\ =& \sqrt{\rho}c_2 b[\phi(nb) - b(1-\Phi(nb))]. \end{aligned} \tag{5.3}$$

A customer is said to be "dropped" if it leaves the queue before gaining service. Since $E(H_i^{(\rho)}) = \alpha_i E(H^{(\rho)}), i = 1, 2$, the rate type $i$ customers are dropped is

$$\beta_i^* = \beta_i\alpha_i E(H^{(\rho)}), \tag{5.4}$$

which can be approximated by using (5.3), and

$$P(\text{type } i \text{ customer dropped}) = \frac{\beta_i^*}{\lambda_i}, \quad i = 1, 2. \tag{5.5}$$

For type 2 customers it may be better to define a dropped call to be one that is caused by some reason other than the call terminating on its own. In that case we would have

$$\beta_2^* = (\beta_2 - \mu)\alpha_2 E(H^{(\rho)}).$$

From (5.1) we see that in heavy traffic the blocking probability remains $\mathcal{O}(1)$. However, since $\beta_i^* \sim \mathcal{O}(\sqrt{\rho})$ and $\lambda_i \sim \mathcal{O}(\rho)$, from (5.5) we see that the probability that a type $i$ customer is dropped is $\mathcal{O}(\rho^{-\frac{1}{2}})$.

Finally, the system utilization is the expected number of busy servers,

$$\rho^* = E(Q^{(\rho)} - H^{(\rho)}),$$

which can be approximated from (5.2) and (5.3).

## APPENDIX

### A.1. Technical Details for the Proof of Theorem 3.1.

PROOF OF LEMMA 3.1.1: Consider the scaled process $q(t)$ on the truncated state space $s_-^{(\rho)}$. (This means that the non-scaled process is considered on the state space $S_-^{(\rho)}$, corresponding to the loss system with $N-1$ servers.) The stationary distribution $F_-^{(\rho)}$ of such a truncated process is the truncation of the distribution $F^{(\rho)}$ :

$$F_-^{(\rho)}(y) = \begin{cases} F^{(\rho)}(y)/F^{(\rho)}(n-) & \text{if } y < n, \\ 1 & \text{if } y \geq n, \end{cases}$$

so, the distribution $F_-^{(\rho)}$ is an upper bound for the distribution $F^{(\rho)}$, i.e.

$$F^{(\rho)}(y) \leq F_-^{(\rho)}(y), \quad y \in \Re.$$

It is well known (e.g. Whitt [4]) that the distribution $F_-^{(\rho)}$ converges to the distribution

$$F_-(y) = \begin{cases} \Phi(y)/\Phi(n) & \text{if } y < n, \\ 1 & \text{if } y \geq n, \end{cases}$$

(the standard normal distribution $\Phi$ truncated from the right at point n.)

Thus, for any $\epsilon > 0$ there exists sufficiently small $y$ such that

$$\limsup_{\rho\to\infty} F^{(\rho)}(y) \le \lim_{\rho\to\infty} F_{-}^{(\rho)}(y) < \epsilon.$$

Similarly, denote by $F_{+}^{(\rho)}$ the stationary distribution of the process $q(t)$ considered on the state space $s_{+}^{(\rho)}$. This distribution is a truncated (from the left at point $n$) version of the distribution $F^{(\rho)}$. So, it is a lower bound for the distribution $F^{(\rho)}$. The sequence of distributions $F_{+}^{(\rho)}, \rho = 1, 2, \dots$, converges to the normal distribution with mean $(\beta-1)n/\beta$ and variance $1/\beta$, truncated from the left at $n$. Thus,

$$\lim_{y\to\infty} \liminf_{\rho\to\infty} F^{(\rho)}(y) = 1$$

The density of the sequence of distributions $F^{(\rho)}$ has been proven. This implies its relative compactness. The proof is complete.

Proof of Lemma 3.1.2: The process $q^{(\rho)}(t)$ is a birth and death process. So,

$$W \equiv \ln \frac{\pi^{(\rho)}(\sigma_2)}{\pi^{(\rho)}(\sigma_1)} = \sum_{\sigma\in G} \ln \frac{\rho}{\mu(\sigma)}$$

where

$$G \equiv s^{(\rho)}[\sigma_1 + 1/\sqrt{\rho}, \sigma_2],$$

$$s^{(\rho)}[x_1, x_2] \equiv \{\sigma \in s^{(\rho)} | \sigma \in [x_1, x_2]\},$$

and

$$\mu(\sigma) = \begin{cases} \rho + \sigma\sqrt{\rho} & \text{if } \rho + \sigma\sqrt{\rho} < N, \\ N + (\rho + \sigma\sqrt{\rho} - N)\beta & \text{if } \rho + \sigma\sqrt{\rho} \ge N. \end{cases}$$

Since $N = \rho + n\sqrt{\rho} + \eta^{(\rho)}, 0 \le \eta^{(\rho)} < 1$, we have

$$\frac{\mu(\sigma)}{\rho} = 1 + g(\sigma)/\sqrt{\rho} + \gamma^{(\rho)}/\rho$$

where $|\gamma^{(\rho)}| \le |1-\beta| < \infty$. Thus,

$$\ln \frac{\rho}{\mu(\sigma)} = -\frac{g(\sigma)}{\sqrt{\rho}} + \mathcal{O}(\frac{1}{\rho}).$$

The function $g$ is bounded on any finite interval, so

$$W = -\sum_{\sigma\in G} g(\sigma)/\sqrt{\rho} + o(1) \to -\int_{a_1}^{a_2} g(y)dy.$$

The proof is complete.

**Technical Details of the Proof of Theorem 4.1.**

PROOF OF THEOREM 4.1:

For notational simplicity we will assume throughout this proof (and all supplementary results) that $p = 0$ and $\eta_1 = \eta_2 = 1$. With obvious modifications the proof holds in the general case.

We will again consider the scaled process $q^{(\rho)}(t)$ given by (3.4) and the scaled state space $s^{(\rho)}$. Although $q^{(\rho)}(t)$ is not Markovian, the process $(q^{(\rho)}, \alpha^{(\rho)})(t)$ is Markov, where $\alpha^{(\rho)}$ is the fraction of type 1 customers among all customers waiting (not being served) at time $t$, i.e.

$$\alpha^{(\rho)} = \begin{cases} \frac{H_1}{H} & \text{if } H > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The condition $\{H > 0\}$ is equivalent to $\{\rho + \sqrt{\rho} q^{(\rho)} > N\}$.

The process $(q^{(\rho)}, \alpha^{(\rho)})(t)$ has a stationary distribution, which induces a stationary distribution of $q^{(\rho)}$. As before, this later distribution is denoted by $F^{(\rho)}$, with $F^{(\rho)}(y),\ y \in \Re$, the corresponding distribution function.

LEMMA 4.1.1. *The sequence of distributions $F^{(\rho)}, \rho = 1, 2, \ldots,$ is relatively compact.*

PROOF: The argument here repeats the proof of the Lemma 3.1.1. The only difference is that for the truncated process defined on $s_+^{(\rho)}$, we have to let $\beta = \min\{\beta_1, \beta_2\}$ to get majorization from above.

Consider any limiting point (in the sense of weak convergence) $F$ of the sequence of distributions $\{F^{(\rho)}\}$ . To simplify notation we will assume that $F^{(\rho)} \to F$.

LEMMA 4.1.2. *Lemma 3.1.2 with $\beta = \alpha_1\beta_1 + \alpha_2\beta_2$ holds for the two-dimensional case.*

The first statement of Theorem 4.1 follows from Lemmas 4.1.1 and 4.1.2. The second statement of Theorem 4.1 follows from Theorem A.2 below.

PROOF OF LEMMA 4.1.2:

To prove Lemma 4.1.2 we need two supplementary results.

LEMMA 4.1.3. *For any fixed interval $[a_1, a_2]$ there exists a constant $A > 0$, such that*

$$\liminf_{\rho\to\infty} \min_{\sigma\in s^{(\rho)}[a_1,a_2]} \sqrt{\rho}p(\sigma) \geq A,$$

*where $p(\sigma)$ is given by (3.12).*

PROOF: The proof consists of the following observations:

(a) As a corollary of Lemma 3.1.2; for any finite interval $[a_1, a_2]$ there exists a constant $C$ such that

$$\max_{\sigma_1,\sigma_2\in s^{(\rho)}[a_1,a_2]} fracp(\sigma_2)p(\sigma_1)$$

and if $a_1 \uparrow a$ and $a_2 \downarrow a$ then $C \downarrow 1$.

(b) As a corollary of (a),

$$F(a_2) - F(a_1-) > 0.$$

(c) The number of elements in the set $s^{(\rho)}[a_1, a_2]$ grows as $(a_2 - a_1)/\sqrt{\rho}$.

LEMMA 4.1.4. *For any $a_1, a_2$ such that $n < a_1 < a_2 < \infty$, we have*

$$\lim_{\rho\to\infty} E^{(\rho)}[|\alpha - \alpha_1|I\{a_1 \leq q \leq a_2\}] = 0,$$

*where $E^{(\rho)}$ is mathematical expectation with respect to the istribution $F^{(\rho)}$ of the vector $(q, \alpha)$. (We use the same notation for the distribution of the vector $(q, \alpha)$ and the distribution of $q$.)*

(The proof of Lemma 4.1.4 depends on some further results and is postponed.)

Proof of the Lemma 4.1.2. continued:

It is easy to see using statement (a) of the proof of Lemma 4.1.3, that it is sufficient to consider the case $n < a_1 < a_2 < \infty$ (note that $\rho + \sigma\sqrt{\rho} > N$ and $H > 0$ in this case). Our notations here is consistant with the notation in the proof of Lemma 3.1.2. We have

$$W \equiv \ln \frac{p(\sigma_2)}{p(\sigma_1)} = \sum_{\sigma\in G} \ln \frac{\rho}{\bar{\mu}(\sigma)},$$

where

$$\bar{\mu}(\sigma) = N + (\rho + \sigma\sqrt{\rho} - N)\bar{\beta}_\sigma,$$

$$\bar{\beta}_\sigma = \alpha_\sigma \beta_1 + (1-\alpha_\sigma)\beta_2 = (\beta_1 - \beta_2)\alpha_\sigma + \beta_2,$$

and

$$\alpha_\sigma = E^{(\rho)}[\alpha | q = \sigma].$$

Thus,

$$\frac{\bar{\mu}(\sigma)}{\rho} = 1 + [n + \bar{\beta}_\sigma(\sigma - n)]/\sqrt{\rho} + \gamma^{(\rho)}/\rho,$$

where $\bar{\beta}_\sigma \leq \max\{\beta_1, \beta_2\}$ and $|\gamma^{(\rho)}| \leq |1 - \bar{\beta}_\sigma| < \infty$ . The coefficient of $1/\sqrt{\rho}$ is uniformly bounded for $\sigma \in G$. Thus,

$$W = -\sum_{\sigma \in G} [n + \bar{\beta}_\sigma(\sigma - n)]/\sqrt{\rho} + o(1) =$$
$$-\sum_{\sigma \in G} [n + \beta(\sigma - n)]/\sqrt{\rho} - \sum_{\sigma \in G} (\sigma - n)(\beta_1 - \beta_2)(\alpha_\sigma - \nu_1)/\sqrt{\rho} + o(1).$$

Let

$$W_1 = \sum_{\sigma \in G} (\alpha_\sigma - \nu_1)/\sqrt{\rho} = \sum_{\sigma \in G} (\alpha_\sigma - \nu_1) p(\sigma)/(\sqrt{\rho} p(\sigma)).$$

Due to lemma 4.1.3, for large $\rho$,

$$\sqrt{\rho} p(\sigma) \geq A > 0.$$

Thus,

$$|W_1| \leq \frac{1}{A} \sum_{\sigma \in G} |\alpha_\sigma - \nu_1| p(\sigma) \leq \frac{1}{A} E^{(\rho)}[|\alpha - \nu_1| I\{a_1 \leq q \leq a_2\}] \to 0, \quad \text{as } \rho \to \infty$$

Then, following the proof of Lemma 4.1.2 we have

$$\lim_{\rho \to \infty} W = -\int_{a_1}^{a_2} g(y) dy.$$

The proof is complete.

Proof of Lemma 4.1.4:

Lemma 4.1.4 is a corollary of Theorem A.2 (below) via the following argument.

Let $F_+^{(\rho)}(y)$ be the conditional distribution of the process $q^{(\rho)}$ under the condition that $q \geq n$:

$$F_+^{(\rho)}(y) = \begin{cases} \frac{F^{(\rho)}(y) - F^{(\rho)}(n-)}{1 - F^{(\rho)}(n-)} & \text{if } y \geq n, \\ 0 & \text{if } y < n. \end{cases}$$

Then $F_+^{(\rho)}$ is also a stationary distribution of the process $q^{(\rho)}$ restricted to the state space $s_+^{(\rho)}$. In terms of the original (non-scaled) process, truncation means that we will look at the two-dimensional birth-death process

$$\vec{H}^{(\rho)}(t) = (H_1^{(\rho)}, H_2^{(\rho)})(t), \quad t \geq 0,$$

where $H_i$ is the number of $i$-customers waiting in the queue (if any). When the number of waiting customers $H = H_1 + H_2$ drops to 0, the service in the system is interrupted until the next customer arrival. The process

$$\vec{X}^{(\rho)}(t) = (X_1^{(\rho)}, X_2^{(\rho)})(t) \equiv \vec{H}^{(\rho)}(t/\sqrt{\rho}), \quad t \geq 0$$

is the process $\vec{H}^{(\rho)}(t)$ being slowed down by a factor of $\sqrt{\rho}$. But, of course, the process $\vec{X}^{(\rho)}(t)$ has the same stationary distribution as $\vec{H}^{(\rho)}(t)$ does. Consider also the scaled process

$$x^{(\rho)}(t) = \frac{1}{\sqrt{\rho}} \vec{X}^{(\rho)}(t).$$

Thus, the relation between the processes $x^{(\rho)}(t)$ and $q^{(\rho)}(t)$ s given by the equation

$$||x^{(\rho)}(t)|| \equiv |x_1^{(\rho)}(t)| + |x_2^{(\rho)}(t)| \equiv q^{(\rho)}(t/\sqrt{\rho}) - n - \eta^{(\rho)}/\sqrt{\rho}$$

where $0 \leq \eta^{(\rho)} < 1$. The transition diagram for the process $\vec{X}^{(\rho)}(t)$ is shown in the Figure A.1. The birth intensities of the process $\vec{X}^{(\rho)}(t)$ are described by the (constant) vector $\sqrt{\rho}(\nu_1, \nu_2)$, and death intensities by the vector

$$M(\vec{X}) = \sqrt{\rho}(\frac{X_i}{X} I\{X > 0\}(1 + \frac{n}{\sqrt{\rho}} + \frac{\gamma^{(\rho)}}{\rho}) + \beta_i \frac{X_i}{\rho}, \quad i = 1, 2),$$

where $|\gamma^{(\rho)}| \leq 1$. If the arbitrary constant $c > 0$ is fixed and $\rho \to \infty$, then

$$M(\vec{X}) = \sqrt{\rho}(\frac{X_i}{X} I\{X > 0\} + \psi_i^{(\rho)}(\vec{X}), \quad i = 1, 2),$$

where both $\psi_1^{(\rho)}, \psi_2^{(\rho)} \to 0$ uniformly in the domain $X/\sqrt{\rho} \leq c$.

Each process $x^{(\rho)}(t)$ is a process with state space being a subset of $\Re_+^2$. So, each $x^{(\rho)}(t)$ can be viewed as a process with the state space $(\Re_+^2, \mathcal{B}(\Re_+^2))$ and its stationary distribution $G^{(\rho)}$ can be viewed as a measure on $(\Re_+^2, \mathcal{B}(\Re_+^2))$.

THEOREM A.1. *Let $T > 0$ be fixed and*

$$x^{(\rho)}(0) \to x_0 \in \Re_+^2$$

*Then the process $x^{(\rho)}(t)$ converges to the deterministic process $x(t)$ defined as the solution of the differential equation*

$$\dot{x}(t) = v(x(t)), \quad 0 \le t \le T,$$

*where the vector $v(x) = (v_1, v_2)(x), x \in \Re_+^2$ is defined as follows:*

$$v(x) = \begin{cases} (\nu_i - \frac{x_i}{||x||}, \quad i = 1, 2) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

*More precisely, for any $\epsilon > 0$,*

$$\lim_{\rho \to \infty} \Pr\{ \sup_{t \in [0,T]} ||x^{(\rho)}(t) - x(t)|| \ge \epsilon\} = 0.$$

Theorem A.1 can be obtained, for example, as a corollary of Theorem 1 in [2].

The sequence of stationary distributions $G^{(\rho)}$, $\rho = 1, 2, \ldots$, is relatively compact due to Lemma 4.1.1.

THEOREM A.2. *Let $G$ be any limiting point (in the sense of weak convergence in $(\Re_+^2, \mathcal{B}(\Re_+^2))$) of the distribution sequence $G^{(\rho)}$, $\rho = 1, 2, \ldots$. Then*

$$G(\{x_1 = \nu_1 ||x||\}) = 1$$

PROOF: It follows from Theorem 8.5.1 in [3] and Theorem A.1 that $G$ must be a stationary distribution of the deterministic process $x(t)$. For any $\epsilon > 0$, the sample path of $x(t)$ starting from any point $x_0$, not on the line $\{x_1 = \nu_1 ||x||\}$, reaches the $\epsilon$-neighborhood of the line $\{x_1 = \nu_1 ||x||\}$ in finite time and then never leaves it. So, any $\epsilon$-neighborhood of that line has $G$-measure 1. This implies that the measure of the line itself is 1. The proof is complete.

**References.**


[1] A.A. Borovkov, "On Limit Laws for Service Processes in Multi-Channel Systems", *Siberian Math Journal*, Vol. 8, (1967), pp 746-763, (English translation).

[2] Ya.A.Kogan, R.Sh.Liptser and A.V.Smorodinskii, "Gaussian Diffusion Approximation of Closed Markov Models of Computer Networks", *Problems of Information Transmission*,Vol. 22, (1986), pp.38-66.

[3] R.Sh.Liptser and A.N.Shirjaev, "The Theory of Martingales", Nauka, Moscow, (1986). (In Russian)

[4] W.Whitt, "Heavy-Traffic Approximations for Service Systems with Blocking", *AT&T Bell Laboratories Technical Journal*, Vol. 63, (1984), pp. 689-708.