A Heavy Traffic Diffusion Limit for a Single Server Queue with Finite Buffer: Two-Sided Reflected Brownian Motion with Drift

by **Christos Zacharias, NYU Stern**
http://moya.bus.miami.edu/~czacharias/

---

1. About

As a PhD student at NYU Stern, I came across the enterprise of developing a heavy traffic diffusion limit for a single server queue with finite waiting room. Without any prior knowledge or systematic training on the subject, I took on the challenge. In this article I lay out a comprehensive, step-by-step construction of such a heavy traffic diffusion limit for the beginner researcher (as I was in 2013). The exposition is based on the same steps that my own comprehension and intuition on the subject led my own journey, when I was exploring this area for the first time and I was trying to make sense out of it and figure out how it works.

2. The Appointment Queue

Consider the random evolution of the appointment backlog of some service system; for example, doctor's appointments with finite buffer. There is a fixed supply of s available time slots per day. There is a random demand of V_k appointment requests for day k, iid with finite mean λ and variance σ^2, $k \geq 1$. Requests for appointments are fully backlogged up to a finite buffer K. Let W_k be the appointment backlog (workload) at the end of day k, $k \geq 0$. Assuming that the appointment system starts empty, the successive workloads can be defined recursively by the Lindley recursion

$$\begin{aligned} W_k &= \min\left\{K, \max\{0, W_{k-1} + V_k - s\}\right\}, \text{ for } k \geq 1, \\ \text{and } W_0 &= 0. \end{aligned} \tag{1}$$

The *maximum* term in (1) is induced by the fact that the workload is never allowed to become negative and at most s customers are scheduled each day. The *minimum* term in (1) restricts the workload in $\{0, 1, ..., K\}$; not more than K customers are allowed to be backlogged.

3. The Diffusion (Brownian) Limit

As in Whitt (2002), we consider a sequence of appointment systems indexed by n, with buffer size K_n, capacity of s_n appointment slots and a fixed input process $\{V_k : k \geq 1\}$. For model n we have

$$W_k^n = \min\left\{K_n, \max\{0, W_{k-1}^n + V_k - s_n\}\right\}, k \geq 1.$$

Let $\{S_k^v : k \geq 0\}$ be the random walk with step size the appointment requests, i.e. $S_k^v = \sum_{i=1}^k V_i$ for $k \geq 1$ and $S_0^v = 0$. For model n, let $S_k^n = \sum_{i=1}^k (V_i - s_n)$ for $k \geq 1$ and $S_0^n = 0$, random walk with steps $V - s_n$. Consider the scaled stochastic processes

$$\mathbf{S}_n^v(t) := \frac{S_{\lfloor nt \rfloor}^v - \lambda nt}{\sqrt{n}},$$
$$\mathbf{S}_n(t) := \frac{S_{\lfloor nt \rfloor}^n}{\sqrt{n}},$$
$$\text{and } \mathbf{W}_n(t) := \frac{W_{\lfloor nt \rfloor}^n}{\sqrt{n}},$$

and note that

$$\mathbf{S}_n(t) = \mathbf{S}_n^v(t) + \sqrt{n}(\lambda - s_n)t.$$

In order to establish a heavy-traffic diffusion limit for the workload process, we assume the following heavy traffic requirements:

Assumption 1. *(a)* $\sqrt{n}(\lambda - s_n) \to \eta \in \mathbb{R}$ *as* $n \to \infty$.
(b) $K_n = \sqrt{n}K$, $K \in \mathbb{R}^+$.
(c) $\sigma^2 < \infty$.

Under Assumption 1(a) and 1(c), and from the *Functional Central Limit Theorem (FCLT)* we get

$$\mathbf{S}_n^v(t) := \frac{S_{\lfloor nt \rfloor}^v - \lambda nt}{\sqrt{n}} \Rightarrow \sigma \mathbf{B}(t),$$
$$\text{and } \mathbf{S}_n(t) = \mathbf{S}_n^v(t) + \sqrt{n}(\lambda - s_n)t \Rightarrow \sigma \mathbf{B}(t) + \eta t,$$

where $\mathbf{B}(t)$ is a standard Brownian motion.

Now we define two more processes of interest. For system n, let U_k^n be the cumulative number of customers lost (blocked) up to day k and L_k^n be the number of unutilized slots up to day k. The workload at the end of day k satisfies

$$\begin{aligned} W_k^n &= \left(\sum_{i=1}^k V_i - U_k^n \right) - (ks_n - L_k^n) \\ &= \sum_{i=1}^k (V_i - s_n) + L_k^n - U_k^n \\ &= S_k^n + L_k^n - U_k^n. \end{aligned}$$

Then, we define the associated scaled stochastic processes

$$\mathbf{L}_n(t) := \frac{L_{\lfloor nt \rfloor}^n}{\sqrt{n}},$$
$$\text{and } \mathbf{U}_n(t) := \frac{U_{\lfloor nt \rfloor}^n}{\sqrt{n}}.$$

Consequently we note that the triplet $(\mathbf{W}_n, \mathbf{L}_n, \mathbf{U}_n)$ satisfies the following three conditions:

(a) $\mathbf{W}_n(t) = \mathbf{S}_n(t) + \mathbf{L}_n(t) - \mathbf{U}_n(t) \in [0, K]$.

(b) $\mathbf{L}_n(t)$ and $\mathbf{U}_n(t)$ are non-decreasing with $\mathbf{L}_n(0) = \mathbf{U}_n(0) = 0$.

(c) $\mathbf{L}_n(t)$ and $\mathbf{U}_n(t)$ increase only when $\mathbf{W}_n(t) = 0$ and $\mathbf{W}_n(t) = K$ respectively.

The triplet $(\mathbf{W}_n, \mathbf{L}_n, \mathbf{U}_n)$ is said to solve the Skorokhod problem for $\mathbf{S}_n$ on $[0, K]$. Such a triplet exists and it is unique, see Harrison (1985). One explicit solution to the Skorokhod problem is provided by Andersen and Mandjes (2008):

$$\mathbf{W}_n(t) = R(\mathbf{S}_n)(t) := \sup_{s\in[0,t]} \left[(\mathbf{S}_n(t) - \mathbf{S}_n(s)) \wedge \left(\inf_{u\in[s,t]} (K + \mathbf{S}_n(t) - \mathbf{S}_n(u) \right) \right].$$

The mapping R is often referred to as the "two-sided reflection map". The *Continuous Mapping Theorem (CMT)* provides the desired diffusion limit for the workload

$$\mathbf{W}_n(t) = R(\mathbf{S}_n)(t) \Rightarrow R(\sigma \mathbf{B}(t) + \eta t) =: \mathbf{W}(t).$$

4. Concluding Remarks

The limiting workload (backlog) process $\mathbf{W}(t)$ is a two-sided reflected Brownian motion (RBM) with drift η, infinitesimal variance σ^2, with reflective barriers 0 and K. As in Whitt (2004), $\mathbf{W}(t) \Rightarrow \mathbf{W}(\infty)$ as $t \to \infty$ with density

$$f_{\mathbf{W}(\infty)}(x) = \frac{2\eta e^{\frac{2\eta x}{\sigma^2}}}{\sigma^2 (e^{\frac{2\eta K}{\sigma^2}} - 1)} \ , \ 0 \le x \le K$$

when $\eta \neq 0$, and the uniform density on $[0, K]$ when $\eta = 0$.

References

Andersen, L.N., M. Mandjes. 2008. Structural properties of reflected lévy processes. *Report-Probability, networks and algorithms* (14) 1–19.

Harrison, J.M. 1985. *Brownian motion and stochastic flow systems*. Wiley New York.

Whitt, W. 2002. *Stochastic-Process Limits: An introduction to stochastic-process limits and their application to queues*. Springer.

Whitt, W. 2004. Heavy-traffic limits for loss proportions in single-server queues. *Queueing Systems* **46**(3) 507–536.