Heavy Traffic Limit Theorems for Two Tandem Polling Stations

Otis B. Jennings
Fuqua School of Business, Duke University, Durham, North Carolina 27708-0120
email: otisj@duke.edu

DRAFT

We consider two tandem single-server stations in heavy traffic. There are two job types and they visit both stations in order. The first station processes jobs in an exhaustive service or gated service fashion and the second station uses an arbitrary nonidling service discipline. The stations have zero switch-over times. We prove heavy traffic limit theorems for the two-dimensional total workload processes, one under exhaustive service and one under gated service. The limiting processes are two-dimensional reflected Brownian motions (RBM) and the limit under exhaustive service is equal in distribution to the limiting process when the first station performs one of two buffer priority policies.

Key words: Heavy traffic; Polling systems
MSC2000 Subject Classification:
OR/MS subject classification:

1. Introduction. Typically, a polling system consists of a single server that visits – or polls – a collection of queues in a cyclic fashion. When switching from one queue to another, the server can incur a travel delay (setup time) or a monetary penalty (setup cost). Either case provides an incentive to poll queues for an extended period of time. The service discipline dictates how long the polling of each queue lasts and in what order the queues are visited.

In this paper we study a network of two single-server workstations. There are two types of jobs, labelled \( k = 1, 2 \). Type \( k \) jobs arrive to the system and immediately enter their own queue at station 1, a polling station with zero setup times. Upon service completion at station 1, jobs move into one of two queues at station 2. We will consider two types of service disciplines for station 1, exhaustive service and gated service. Under exhaustive service, perhaps the most widely studied service discipline for polling systems, the server processes jobs in the currently polled queue until that queue is emptied. Under gated service, the server, on a given visit, processes only those jobs present when the visit begins. Station 2 processes jobs according to an arbitrary, nonidling policy, again with no setup delay.

We prove two heavy traffic limit theorems for the diffusive scaled, two-dimensional, total workload process, one for when station 1 operates under exhaustive service and another when station 1 operates under gated service. The limiting processes are both two-dimensional, reflected Brownian motions (RBMs). The RBM arising under exhaustive service is equal in distribution to the limiting process that would arise if the first station were operated under one of the two possible static buffer priority (SBP) service disciplines. The limiting process under gated service is novel. To the author’s knowledge, this is the first paper that proves a heavy traffic limit theorem for a polling system in a network setting.

The study of polling systems has an extensive history; see, for instance, Takagi \([20, 21]\) for surveys of early work related to telecommunications and Lan and Olsen \([16]\) for a more contemporary review. Most of these works consider single station systems. Polling systems in the network setting has been studied by Bertsimas and Nino-Mora \([2]\), Jennings \([13]\), Andradottir et al. \([1]\), and Dai and Jennings \([8]\). These papers focus predominantly on stability issues, as opposed to performance analysis.

Diffusion approximation is one method of obtaining performance results. Often diffusion approximations are obtained via heavy traffic limits theorems. Such is the case of Coffman et al. \([6, 7]\) for a polling system under exhaustive service, Jennings \([12]\) for a polling system under \( \ell \)-limited service, and Kushner \([15]\) for polling systems under a large variety of settings. Van der Mei and co-authors (e.g. Van der Mei and Olsen \([22]\)) have several results with polling stations under various operating regimes. There are other works that perform heavy traffic analysis of polling systems. For example, Markovitz et al. \([17]\) and Markovitz and Wein \([18]\) assume the form of the would-be heavy traffic limit and proceed to dynamically control the polling dynamics. All of these works are limited to single server systems.
Some of the first papers with heavy traffic limit theorems for queueing networks also consider feedforward networks, as we do here; see, e.g., Iglehart and Whitt \[10\], Harrison \[9\], and Peterson \[19\]. Unlike the results in those papers, the two-dimensional local workload processes do not jointly converge to a two-dimensional reflected Brownian motion. In fact, the scaled, immediate workload for station 2 does not converge at all. Like Peterson \[19\] we instead use the two-dimensional total workload process, whose \(j\)th component measures, as a function of time, the amount of time (or effort) required to process those jobs in the system that have yet to be processed by station \(j\). Unlike Peterson \[19\] we cannot, once a limit on the total workload process is obtained, back out what the immediate workload process would be.

Among studies involving Brownian motion approximations to multiclass queueing networks, state space collapse is an almost ubiquitous phenomenon. When a network exhibits state space collapse, the limiting workload associated with a given station can be decomposed into the contributions by the constituent queues via a constant lifting operator. In a series of important papers, Bramson and Williams \[4, 23, 24\] simplify the procedure for proving a having traffic limit theorem by showing that, under some mild conditions, if the associated fluid model exhibits state space collapse, then the HTLT exists and the limiting diffusion process also exhibits state space collapse. This framework was used, for example, by Bramson and Dai \[5\] and, with some modifications to accommodate closed networks, Kumar \[14\].

As for this paper, both exhaustive service and gated service result in large fluctuations of the queue length process. For example, under exhaustive service, jobs fluctuate from residing entirely in queue 1, at instances when queue 2 has just been emptied, and entirely in queue 2, just as queue 1 is emptied. Under fluid scaling, these fluctuations preclude state space collapse and hence, the Bramson-Williams proof framework is not amenable to our situation.

In our analysis, we consider the fluctuations of jobs at station 1 and the effect these fluctuations have on the total workload and the availability of jobs at station 2. The total workload at the second station has two parts: the immediate workload and what we will refer to here as the upstream workload, made up of the station 2 processing times of the jobs currently residing at station 1. Consider exhaustive service. Over time horizons corresponding to the fluid, or law of large numbers, scaling, the station 1 workload under heavy traffic is constant. Despite being constant, the workload oscillates between residing entirely in queue 1 and then entirely in queue 2. From station 2’s perspective, when the workload is in queue 1 it represents a different amount than when the workload is in queue 2. The crucial point is that the upstream workload oscillates between a high value and a low value during polling cycles at station 1. Because the station 2 total workload is constant (provided station 2 never starves) over this time horizon, the immediate workload oscillates as well; hence, the absence of a limiting process for the station 2 immediate workload. Nevertheless, one can anticipate the form of the reflective boundary for the station 2 limiting total workload process. The boundary keeps the immediate workload from going negative over any given cycle, in particular, when the upstream workload reaches a maximum during the processing cycle at station 1.

Now consider the fluid scaled version of the system operated under gated service. Oscillations of workload will likewise determine the nature of the station 2 total workload constraint. However, under gated service, the oscillations are not as straightforward. For one, except for perhaps the first polled queue, no queue at station 1 ever empties. Still, a cyclic pattern is reached, where, at the beginning of a visit to a given queue, a fixed fraction of the total workload resides in that queue. The other complication is that, unlike with exhaustive service, these oscillations do not settle into their cyclic pattern instantaneously. However, in each successive cycle, the distance from the eventual cyclic pattern decreases geometrically; see Proposition \[5.1\].

Because of their oscillations, the queue length process does not converge. Because of the queue length process does not converge, there can be no state space collapse. Nevertheless, there is an asymptotic relation between the weighted queue lengths and the workload processes that is crucial to the arguments throughout; see Proposition \[3.1\] and Corollary \[3.1\]. These properties will be useful in future analyses of polling models.

The remainder of the paper progresses as follows. In the next section we present the network model and our main result, Theorem \[2.1\]. In preparation of the proof we demonstrate the convergence of some of the model’s building blocks in Sections \[3\] and \[4\]. In Section \[5\], we prove a result needed for the convergence under gated service. The main result is proven is Section \[6\]. Some final remarks and extensions are in
the concluding section.

Throughout this paper, all processes are assumed to have right-continuous with left-hand limits sample paths and considered as random elements of the Skorohod space $D[0,\infty)$ (see, e.g., Jacod and Shiryaev (1987) and Liptser and Shiryaev (1989)), and convergence in distribution for the processes is understood as weak convergence of the induced measures on $D[0,\infty)$. With $\Rightarrow$ we denote convergence in distribution in an appropriate metric space. Also, $\to^p$ denotes convergence in probability.

2. The Model and Main Results. Consider a network of two single-server workstations, labelled $j = 1, 2$. There are two types of jobs, labelled $k = 1, 2$, each with its own exogenous arrival stream. Type $k$ jobs arrive to the system and enter queue $k$ at station 1. After being processed at station 1, jobs immediately move to station 2 and again wait in-queue. After processing at station 2, jobs leave the system.

The servers at both stations are nonidling. That is, when there is no work present at their respective stations, the servers must be busy processing jobs. For station 2, this is the only stipulation we place on the performance of work. Jobs at the station can be processed in any arbitrary fashion, as long as the server is nonidling. As for the first station, the server processes jobs according to one of two service disciplines: either exhaustive service or gated service. We analyze the system under each policy as the server is nonidling. As for the first station, the server processes jobs according to one of two service disciplines: either exhaustive service or gated service. We analyze the system under each policy separately.

The limiting procedure involves a sequence of networks, indexed by $n$. Each network has the same structure as stated before. The only difference is in the parameters involved. For the $n$th network, the sequence of interarrival times for type $k$ jobs are denoted $\{\xi^n_k(i), i \geq 1\}$ and the sequence of service times for type $k$ jobs at station $j$ are denoted $\{\eta^n_{j,k}(i), i \geq 1\}$. The sequences are mutually independent. The elements of each sequence are independent and, with the exception of its first element, identically distributed. The mean and standard deviation of the i.i.d. portion of the interarrival times are denoted $1/\lambda^n_k > 0$ and $\sigma^n_{a,k}$, $k = 1, 2$, respectively. Likewise, the service time means and standard deviations are $m^n_{j,k}$ and $\sigma^n_{s,j,k}$, $j = 1, 2$, $k = 1, 2$, respectively. The quantity $\lambda^n_k$ is referred to as the arrival rate for type $k$ jobs. The quantity $\mu^n_{j,k} \equiv 1/m^n_{j,k}$ is referred to as the service rate for type $k$ jobs at station $j$. One can compute the contribution from each job type to each station’s utilization:

$$\rho^n_{j,k} = \lambda^n_k m^n_{j,k}, \quad j = 1, 2, \quad k = 1, 2,$$

where the total utilization of station $j$ is

$$\rho^n_j = \rho^n_{j,1} + \rho^n_{j,2}.$$

We have the following convergences:

$$\lambda^n_k \to \lambda_k, \quad m^n_k \to m_k, \quad k = 1, 2, \quad (1)$$

and

$$\sigma^n_{a,k} \to \sigma_{a,k}, \quad \text{and} \quad \sigma^n_{s,j,k} \to \sigma_{s,j,k}, \quad j = 1, 2, \quad k = 1, 2, \quad (2)$$

as $n \to \infty$. As is standard in heavy traffic analysis, we have the following heavy traffic conditions:

$$c^n_j = \sqrt{n} (\rho^n - 1) \to c_j, \quad j = 1, 2, \quad (3)$$

as $n \to \infty$, where $c_1$ and $c_2$ are finite. We assume the Lindeberg conditions hold; namely, for $i \geq 1$, $j = 1, 2, k = 1, 2$, and any $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{E} \left[ \xi^n_k(i)^2 \cdot 1_{\{\xi^n_k(i) > \epsilon \sqrt{n}\}} \right] = 0, \quad \text{and} \quad \lim_{n \to \infty} \mathbb{E} \left[ \eta^n_{j,k}(i)^2 \cdot 1_{\{\eta^n_{j,k}(i) > \epsilon \sqrt{n}\}} \right] = 0. \quad (4)$$

For each network, we track the total workload process $W^n = (W^n_1, W^n_2)$, where $W_j$ measures, as a function of time, the amount of time (i.e., effort) required to process those jobs in the system that have yet to be fully processed by station $j$. A similar process is $Z^n = (Z^n_1, Z^n_2)$, where $Z_j^n$ measures the amount of effort required to fully process the jobs currently at station $j$, including the job currently being processed. We refer to $Z^n$ as the immediate workload process and $Z^n_j$ the station $j$ immediate workload process. Clearly, for station 1, the immediate workload and the total workload are identical: $W^n_1 = Z^n_1$. For station 2, the total workload consists of the immediate workload and the upstream workload, $U^n_2$, which is precisely the sum of the station 2 processing times of the jobs currently residing at station 1. The upstream workload captures the work in the system that has yet to reach station 2.
We also track the queue length processes \( Q^n = (Q^n_{j,k}, j = 1, 2, k = 1, 2) \). At time zero the nth network need not be empty. The immediate workloads at time 0 are simply the sum of the first elements of the corresponding service times:

\[
Z^n_j(0) = \sum_k \sum_{i=1} Q^n_{j,k}(0), \quad j = 1, 2.
\] (5)

It is possible that the server is in the process of serving a job at time zero. Similarly, time 0 may mark the middle of some interarrival time. We account for the residual service and interarrival times by allowing the first element of each random variable sequence to be distributed differently from the remainder.

Under the heavy traffic assumption \((3)\), the sequence of networks approaches critical loading; i.e., \( \rho^n \to 1 \), as \( n \to \infty \). Under this scaling, one would expect the workload process to be null recurrent, where, as a function of \( t \), the workload (at least at station 1) would be order \( t \). Accordingly, we define a family the scaled workload processes \( \{\hat{W}^n, n \geq 1\} \), where \( \hat{W}^n = (\hat{W}_1^n, \hat{W}_2^n) \), as follows:

\[
\hat{W}_j^n(t) = \frac{W_j^n(\sqrt{n}t)}{\sqrt{n}}, \quad t \geq 0.
\] (6)

At time \( nt \) we expect the workload to be order \( \sqrt{n} \). Hence the spatial normalizing constant \( \sqrt{n} \) in the denominator. Our main result to appear next, claims that under this scaling, the family of scaled processes obeys a function central limit theorem. We also scale the other processes, though, in general, they will not converge:

\[
\hat{U}_2(t) = \frac{U_2^n(\sqrt{n}t)}{\sqrt{n}}, \quad \hat{Z}_2(t) = \frac{Z_2^n(\sqrt{n}t)}{\sqrt{n}}, \quad \hat{Q}_{j,k}^n(t) = \frac{Q_{j,k}^n(\sqrt{n}t)}{\sqrt{n}}, \quad j = 1, 2, \quad k = 1, 2.
\] (7)

Finally, we place some restriction on the limit of our mean processing times. Without loss of generality, one could assume that \( m_{2,1} \geq m_{1,1} \). We assume that the stations are not identical in the limit, so that

\[
m_{2,1} > m_{1,1}.
\]

Along with \((3)\), this implies

\[
\frac{m_{2,2}}{m_{1,2}} < 1 < \frac{m_{2,1}}{m_{1,1}}.
\] (8)

For sufficiently large \( n \), the mean processing time for type one jobs increases from station 1 to station 2, while the mean processing time for type 2 jobs decreases. In this sense, we refer to type 1 jobs as decelerating and type 2 jobs as accelerating as they move through the network.

### 2.1 The Main Result

**Theorem 2.1** Assume that

\[
\left( \hat{W}_1^n(0), \hat{W}_2^n(0), \hat{U}_2^n(0) \right) \to \left( \hat{W}_1(0), \hat{W}_2(0), \hat{U}_2(0) \right)
\] (9)

and that

\[
\left( \sum_k m_{1,k} \hat{Q}_{1,k}^n(0), \sum_k m_{2,k} \hat{Q}_{1,k}^n(0), \sum_k m_{2,k} \hat{Q}_{1,k}^n(0) \right) \to \left( \hat{W}_1(0), \hat{W}_2(0), \hat{U}_2(0) \right)
\] (10)

as \( n \to \infty \), where

\[
\hat{W}_2(0) \geq \frac{m_{2,1}}{m_{1,1}} \hat{W}_1(0) \geq 0.
\] (11)

Under \((11) - (14)\) and \((8)\), we have

\[
\hat{W}^n \Rightarrow \hat{W} \equiv \hat{X} + \hat{I}.
\] (12)

The two-dimensional total free process \( \hat{X} = (\hat{X}_1, \hat{X}_2) \) has \( \hat{X}_1 \equiv \hat{W}_j(0) + \hat{B}_j + c_j e \), where \( e \) is the identity function (i.e. \( e(t) = t \)) and \( \hat{B} = (\hat{B}_1, \hat{B}_2) \) is a two-dimensional Brownian motion with covariance matrix

\[
\Gamma \equiv \begin{pmatrix}
\sum_k \lambda_k \left( \sigma^2_{1,k} + \rho^2_{1,k} \sigma^2_{a,k} \right) & \sum_k \lambda_k \rho_{1,k} \rho_{2,k} \sigma^2_{a,k} \\
\sum_k \lambda_k \rho_{1,k} \rho_{2,k} \sigma^2_{a,k} & \sum_k \lambda_k \left( \sigma^2_{a,2,k} + \rho^2_{2,k} \sigma^2_{a,k} \right)
\end{pmatrix}.
\]
The idleness process is obtained by applying the reflection map to functions of the free processes:

\[ \hat{I}_1 = \Phi(X_1), \]

and

\[ \hat{I}_2 = \Phi(\tilde{X}_2 - \gamma_p \tilde{W}_1), \]

where the functional \( \Phi(x)(t) = -\inf_{s \leq t}(x(s) \wedge 0) \). The parameter \( \gamma_p \) depends on the service discipline employed at station 1:

- under Exhaustive Service \( \gamma_E = \frac{\rho_{2,1}}{\rho_{1,1}}, \)
- under Gated Service \( \gamma_G = \frac{\rho_{2,1} + \rho_{1,2} \rho_{2,2}}{1 - \rho_{1,1} \rho_{1,2}}. \)

The process \( \tilde{X}_2 - \gamma_p \tilde{W}_1 \) from (14) can be thought of as the virtual, immediate free process. The limiting immediate free process for station 2 does not exist under either station 1 polling policy, neither does the limiting immediate workload process for station 2.

It is instructive to compare the form of the limiting process under exhaustive service to the limiting process when type 2 customers are given priority over type 1 customers at station 1. In this sense, the service at station 1, the local workload at station 2 converges and it has the natural constraint that it is nonnegative. Such a constraint is equivalent to restricting the station 2 total workload from below by \( m_{2,1}/m_{1,1} \) times the station 1 workload, the same constraint that we have under exhaustive service.

Now suppose priority were given to type 1 customers. The lower bound on the station 2 total workload, \( \tilde{W}_1 m_{2,2}/m_{1,2} < \tilde{W}_1 \) would be less restrictive than under the other priority scheme. In this sense, the workload under exhaustive service uncovers the worst possible total workload constraint among the possible priority schemes. This comparison does not hold when there are three or more job types. The reason is that when there are at least three job types, the workload will never reside exclusively in one of the queues, as it can under some priority policies.

3. Preliminaries. In preparation for the proof of our main result, we investigate some of the properties of the basic processes of our network.

3.1 Basic Processes. For each \( k = 1, 2 \) let

\[ A_k^n(t) = \inf\{ i \geq 1 : \sum_{\ell=1}^{i} \xi_k^n(\ell) \leq t \} \]

denote the number of type \( k \) arrivals to the system (in excess of the jobs present at time 0) to the system by time \( t \geq 0 \). Likewise, for each \( j = 1, 2 \) and \( k = 1, 2 \), let

\[ S_{j,k}^n(i) = \sum_{\ell} \eta_{j,k}^n(\ell) \]

be the sum of the first \( i \) type \( k \) jobs to be processes at station \( j \). As with the total workload process in (6), we can consider diffusive scaled versions of these processes:

\[ \tilde{A}_k^n(t) = \frac{A_k^n(nt) - \lambda_k^n nt}{\sqrt{n}}, \quad \text{and} \quad \tilde{S}_{j,k}^n(t) = \frac{S_{j,k}^n(nt) - m_{j,k}^n nt}{\sqrt{n}}, \]

for each \( j = 1, 2, k = 1, 2, \) and \( t \geq 0 \). Notice that here we subtract the expected value of the processes at time \( nt \) before rescaling the size of the process.

It also helps to perform a fluid scaling on the arrival and queueing processes for each \( j = 1, 2 \) and \( k = 1, 2 \):

\[ \tilde{A}_{j,k}^n(t) = \frac{A_{j,k}^n(t)}{n}, \quad \tilde{Q}_{j,k}^n(t) = \frac{Q_{j,k}^n(nt)}{n}, \quad t \geq 0, \]
where for station \( j \) the process \( A_{j,k}^n \) counts type \( k \) arrivals after time 0 as well as the initial customers that have yet to be fully processed by the station. The process can be expressed as

\[
A_{j,k}^n(t) \equiv A_j^k(t) + Q_{j,k}^n(0), \quad t \geq 0,
\]

where \( Q_{j,k}^n \equiv \sum_{i \leq j} Q_{i,k}^n \). Some limits are expressed in the following lemma.

**Lemma 3.1** Given \((\mathcal{F}_i)\) and \(Q_{j,k}(0) \to P 0 \) for each \( j = 1, 2 \) and \( k = 1, 2 \),

\[
\tilde{A}_k^j \Rightarrow \tilde{A}_k \equiv \sigma_{a,k} \sqrt{\lambda_k} \beta_{a,k}, \quad \tilde{A}_{j,k}^n \to^P \lambda_k e, \quad j = 1, 2, \quad k = 1, 2,
\]

\[
\tilde{S}_{j,k}^n \Rightarrow \tilde{S}_{j,k} \equiv \sigma_{s,j,k} \beta_{s,j,k}, \quad \tilde{S}_{j,k}^n \circ \tilde{A}_{j,k}^n \Rightarrow \sigma_{s,j,k} \sqrt{\lambda_k} \beta_{s,j,k}, \quad j = 1, 2, \quad k = 1, 2,
\]

\[
(1/\sqrt{n}) \min_{i \leq n} \xi_i^k(t) \to^P 0, \quad (1/\sqrt{n}) \min_{i \leq n} \eta_i^j(t) \to^P 0, \quad j = 1, 2, \quad k = 1, 2, \quad \forall \alpha,
\]

as \( n \to \infty \), where \( \beta_{a,k} \) and \( \beta_{s,j,k} \), for each \( j = 1, 2 \) and \( k = 1, 2 \), are independent, standard Brownian motions (i.e. with mean 0 and variance 1).

Again \( e \) is the identity function; i.e., \( e(t) = t \), for all \( t \geq 0 \).

**Proof.** Third result is standard; see, for instance, Lemma 5.1 of Coffman et al. [7]. The first follows from Lemma 14.6 of Billingsley [3]. The second is an immediate consequence of the first. The forth follows from the second and third and the random time change lemma on page 151 of Billingsley [3]. The fifth and sixth are a consequence of the first and an third results. \( \square \)

### 3.2 The Workload Processes.

Consider the \( n \)th network in our sequence. Under any nonidling service discipline, the workload processes can be expressed as

\[
Z_1^n(t) = \sum_k S_{1,k}^n(A_{1,k}^n(t) - t - \int_0^t 1_{\{Z_1^n(s) > 0\}} ds),
\]

\[
Z_2^n(t) = \sum_k S_{2,k}^n(A_{2,k}^n(t) - Q_{1,k}^n(t)) - \int_0^t 1_{\{Z_2^n(s) > 0\}} ds,
\]

\[
U_2^n(t) = \sum_k [S_{2,k}^n(A_{2,k}^n(t)) - S_{2,k}^n(A_{2,k}^n(t) - Q_{1,k}^n(t))],
\]

\[
W_1^n(t) = Z_1^n(t),
\]

and

\[
W_2^n(t) = U_2^n(t) + Z_2^n(t)
\]

For each \( j = 1, 2 \), introduce the total free process,

\[
X_j^n(t) \equiv \sum_k S_{j,k}^n(\tilde{A}_{j,k}^n(t)) - t
\]

and its scaled version

\[
\hat{X}_j^n(t) \equiv \frac{X_j^n(nt)}{\sqrt{n}} = (1/\sqrt{n}) \sum_k S_{j,k}^n(\tilde{A}_{j,k}(nt)) - \sqrt{n}t.
\]

The free process \( X_j^n \) tracks the would-be total workload level, if station \( j \) worked constantly (decreased the workload at rate 1) and, of course, the workload is not restricted to be positive. The difference between the free process and the actually workload process is the idling process \( I_j^n \) which tracks, as a function time, the cumulative amount of idle time for the station:

\[
W_j^n = X_j^n + I_j^n, \quad j = 1, 2.
\]

For station 1, \( I_1^n \) is the minimum nondecreasing function that can be added to \( X_1^n \) so that the resulting sum is nonnegative; i.e.,

\[
I_1^n \equiv \Phi(X_1^n),
\]
where, as mentioned earlier, for any function $x$, $\Phi(x)(t) = -\inf_{s \leq t} (x(s) \wedge 0)$. The workload process can be expressed as the free process, reflected at zero:

$$W^n_j(t) = X^n_j(t) + \Phi(X^n_j)(t), \quad t \geq 0.$$  \hfill (29)

The same equation holds for the station 1, scaled workload process:

$$\tilde{W}^n_1(t) = \tilde{X}^n_1(t) + \Phi(\tilde{X}^n_1)(t), \quad t \geq 0.$$  \hfill (30)

It is this version of the station 1 workload that leads directly to the limit of $\tilde{W}^n_1$ as expressed in (12). The station 2 analog of (29) is not as simple.

We can immediately review some well known convergence results for the scaled, free total workload processes and the scaled station 1 workload process.

**Lemma 3.2** Assume $[\| \cdot \| - 4]$ and $[11]$ hold. Then, as $n \to \infty$,

$$(\tilde{X}^n, \tilde{W}^n_1) \Rightarrow (\tilde{X}, \tilde{W}^n_1).$$  \hfill (31)

**Proof.** It follows from (17), (18), (19), and (26), that

$$\tilde{X}^n_j(t) = \tilde{B}^n_j(t) + c_j^n t + \sum_k m^n_{j,k} Q^n_k(0),$$

where

$$\tilde{B}^n_j(t) = \sum_k \left( \tilde{S}^n_{j,k}(A^n_{j,k}(t)) + m^n_{j,k} \tilde{A}^n_k(t) \right).$$

The convergence $\tilde{B}^n \Rightarrow \tilde{B}$, where $\tilde{B}$ is a two-dimensional Brownian motion with covariance matrix $\Gamma$, follows from Lemma 3.1 and the independence of $\hat{A}^n_1$, $\hat{A}^n_2$, $\tilde{S}^n_1$, and $\tilde{S}^n_2$. The convergence $\tilde{X}^n_j \Rightarrow \tilde{X}_j$ for each $j$ individually follows from (3) and (10). Joint convergence follows from the independence of $\tilde{B}$ and $\tilde{W}$. The operator $\Phi$ is continuous. Hence, by the continuous mapping theorem and (30) the convergence of the scaled station workload occurs jointly with the two-dimensional scaled free process. \hfill $\square$

For station 2, the total workload cannot be zero unless the immediate workload is zero. Hence, in determining the idle time process $I^n_2$ for station 2, one should use the immediate free process at station 2, not the total free process at station 2. The immediate free process for station 2 would track the immediate (and not total) workload process if the resident server worked all of the time. This process should exclude the part of the station 2 total free workload process that is upstream. The appropriate station 2 immediate free process is $X^n_2 - U^n_2$ and the station 2 immediate and total workload processes can be expressed as

$$Z^n_2(t) = X^n_2(t) - U^n_2(t) + I^n_2(t) \quad \text{and} \quad W^n_2(t) = X^n_2(t) + I^n_2(t), \quad t \geq 0,$$

where

$$I^n_2 \equiv \Phi(X^n_2 - U^n_2).$$  \hfill (33)

The scaled analogs hold:

$$\tilde{Z}^n_2(t) = \tilde{X}^n_2(t) - \tilde{U}^n_2(t) + \tilde{I}^n_2(t) \quad \text{and} \quad \tilde{W}^n_2(t) = \tilde{X}^n_2(t) + \tilde{I}^n_2(t), \quad t \geq 0,$$

where

$$\tilde{I}^n_2 \equiv \Phi(\tilde{X}^n_2 - \tilde{U}^n_2)$$  \hfill (34)

In general, $\tilde{U}^n_2$ does not converge; neither does $\tilde{X}^n_2 - \tilde{U}^n_2$. Nevertheless, as is shown in the proof of Theorem 2.1 $I^n_2$ does converge.

### 3.3 C-tightness

Lastly, we review some notions about convergence. Throughout the proof, we exploit the fact that Brownian motion is almost surely continuous so that any sequence of processes that converges to Brownian motion is C-tight:

**Definition 3.1** A family of processes is C-tight if it is tight and all weak limit points of the sequence of their laws are laws of continuous processes.

The following provides an equivalent notion and can be found in Chapter 6 of Jacod and Shiryaev [11]:
LEMMA 3.3 The sequence of processes \( \{Y^n, n \geq 1\} \) is C-tight if and only if, for all \( T > 0 \) and \( \epsilon > 0 \),

\[
\lim_{H \to \infty} \limsup_{n \to \infty} \mathbb{P} (|Y^n(0)| > H) = 0
\]

and

\[
\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P} \left( \sup_{s, t \leq T, |s-t| \leq \delta} |Y^n(s) - Y^n(t)| > \epsilon \right) = 0.
\]

The following is a straightforward consequence that is very helpful throughout the proof of the main theorem.

LEMMA 3.4 Given a C-tight sequence, \( \{Y^n, n \geq 1\} \), for all \( T > 0, K > 0, b > 0 \), and \( \epsilon > 0 \),

\[
\lim_{n \to \infty} \mathbb{P} \left( \sup_{s \leq T, t \leq K/n^b} |Y^n(s + t) - Y^n(s)| > \epsilon \right) = 0.
\]

We can use this lemma to place bounds on the number of arriving jobs, the number of services, the arrival of work, and the change in the free workload processes as well as the station 1 workload.

LEMMA 3.5 For any \( \epsilon > 0 \), \( b \in [0, 1] \), constants \( K \) and \( T \) and sufficiently large \( n \),

\[
\mathbb{P} \left( \sup_{s \leq T, t \leq K} |A^n_k(sn + tn^b) - A^n_k(sn) - \lambda^n_k tn^b| > \epsilon \sqrt{n} \right) < \epsilon, \quad k = 1, 2,
\]

\[
\mathbb{P} \left( \sup_{s \leq T, t \leq K} |S^n_{j,k}(sn + tn^b) - S^n_{j,k}(sn) - \rho^n_{j,k} tn^b| > \epsilon \sqrt{n} \right) < \epsilon, \quad j = 1, 2, \quad k = 1, 2,
\]

\[
\mathbb{P} \left( \sup_{s \leq T, t \leq K} |X^n_j(sn + tn^b) - X^n_j(sn)| > \epsilon \sqrt{n} \right) < \epsilon, \quad j = 1, 2,
\]

and

\[
\mathbb{P} \left( \sup_{s \leq T, t \leq K} |W^n_1(sn + tn^b) - W^n_1(sn)| > \epsilon \sqrt{n} \right) < \epsilon.
\]

For each \( j = 1, 2 \) and \( k = 1, 2 \), the quantities \( \lambda^n_k, m^n_{j,k}, \) and \( \rho^n_{j,k} \) can be replaced by their limits \( \lambda_k, m_{j,k}, \) and \( \rho_{j,k} \), respectively, as long as \( b \leq 0.5 \).

PROOF. In Lemma 3.1 it was shown that each of the sequences of scaled processes \( \{\tilde{A}^n_k, n \geq 1\}, \{\tilde{S}^n_k, n \geq 1\}, \) and \( \{\tilde{S}^n_{j,k} \circ \tilde{A}^n_k, n \geq 1\} \) converged to some Brownian motion for each \( j = 1, 2 \) and \( k = 1, 2 \). The same was done for \( \tilde{X}^n_j \) for each \( j = 1, 2 \) in Lemma 3.2, where we also showed that \( \tilde{W}^n_1 \) converged to a continuous function of Brownian motion. Hence, each of the sequences is C-tight and the claims in the lemma follow from Lemma 3.4. Suppose \( b \leq 0.5 \). For each \( k = 1, 2 \), we can replace \( \lambda^n_k \) with \( \lambda_k \) because \( Kn^b|\lambda^n_k - \lambda_k| < \epsilon \sqrt{n} \) for large enough \( n \). The same argument extends to \( m^n_{j,k} \) and \( \rho^n_{j,k} \) for each \( j = 1, 2 \) and \( k = 1, 2 \).

We can show that the station 1 scaled queue length properties are asymptotically bounded in probability.

LEMMA 3.6 For each \( k = 1, 2 \),

\[
\lim_{M \to \infty} \limsup_{n \to \infty} \mathbb{P} \left( \sup_{s \leq nT} Q^n_{1,k}(s) > M \sqrt{n} \right) = 0.
\]  \( (36) \)

PROOF. By Lemma 3.2 the station 1 scaled workload converges to a reflected Brownian motion, an almost surely continuous process. Hence, over a finite interval, the limiting process is bounded in probability. It follows that

\[
\lim_{M \to \infty} \limsup_{n \to \infty} \mathbb{P} \left( \sup_{s \leq nT} \tilde{W}^n_1(s) > M \right) = 0.
\]  \( (37) \)
Notice that
\[
\mathbb{P}\left( \sup_{s \leq nT} Q^n_1(s) > M\sqrt{n} \right) \leq \mathbb{P}\left( \sup_{i \leq 2\lambda nT + M\sqrt{n}} \left( S^n_{1,k}(i + M\sqrt{n}) - S^n_{1,k}(i) \right) \leq m_{1,k} M\sqrt{n}/2 \right) \\
+ \mathbb{P}\left( \sup_{s \leq nT} W^n_1(s) > m_{1,k} M\sqrt{n}/2 \right) \\
+ \mathbb{P}\left( A^n_k(nT) > 2\lambda nT \right),
\]
so that (36) follows from Lemma 3.1, Lemma 3.5 and (37).

Now that the queue lengths are known to be asymptotically bounded in probability, we can make a stronger claim, that the appropriately weighted queue lengths are approximately equal to the workload processes associated with them.

**Proposition 3.1** For any \( \epsilon > 0 \),
\[
\lim_{n \to \infty} \mathbb{P}\left( \sup_{s \leq T} \left| \sum_k W^n_k(s) - m_{1,k} \bar{Q}^n_{1,k}(s) \right| > \epsilon \sqrt{n} \right) = 0 \tag{38}
\]
and
\[
\lim_{n \to \infty} \mathbb{P}\left( \sup_{s \leq T} \left| \sum_k \bar{U}^n_k(s) - m_{2,k} \bar{Q}^n_k(s) \right| > \epsilon \right) = 0. \tag{39}
\]

Equations (38) and (39) can be thought of as the polling station analog to the so-called *state space collapse* property.

**Proof.** Notice that
\[
\mathbb{P}\left( \sup_{s \leq nT} \left| \sum_k W^n_k(s) - m_{1,k} \bar{Q}^n_{1,k}(s) \right| > \epsilon \sqrt{n} \right) \\
\leq \sum_k \mathbb{P}(A^n_k(nT) > 2\lambda nT) + \sum_k \mathbb{P}\left( \sup_{s \leq nT} Q^n_{1,k}(s) > M\sqrt{n} \right) \\
+ \sum_k \mathbb{P}\left( \sup_{i \leq 2\lambda nT + M\sqrt{n}} \eta^n_{1,k} \sqrt{n} > \epsilon \sqrt{n} \right) \\
+ \sum_k \mathbb{P}\left( \sup_{i \leq 2\lambda nT + M\sqrt{n}, i_0 \leq M\sqrt{n}} \left| S^n_{1,k}(i + i_0) - S^n_{1,k}(i) - i_0 m_{1,k} \right| > \epsilon \sqrt{n} \right),
\]
so that (38) follows from Lemmas each rv is small, 3.1, 3.5, and 3.6. An analogous argument can be made for (39) by replacing \( m_{1,k}, \eta^n_{1,k}, \) and \( S^n_{1,k} \) with \( m_{2,k}, \eta^n_{2,k}, \) and \( S^n_{2,k} \), respectively, in (40). \( \Box \)

We introduce the process \( W^n_{1,k} \), which captures the product type \( k \) contribution to the station 1 immediate workload process, and \( \bar{W}^n_{1,k} \), its scaled analog. Clearly \( W^n_1 = \sum_k W^n_{1,k} \) and \( \bar{W}^n_1 = \sum_k \bar{W}^n_{1,k} \). The scaled queue-level processes do not converge. However, we can prove a version of the state space collapse property for them. The first result of the following corollary was implicitly shown in the preceding proof. The second result is a consequence of the first and Proposition 3.1.

**Corollary 3.1** For any \( \epsilon > 0 \),
\[
\lim_{n \to \infty} \mathbb{P}\left( \sup_{s \leq T} \left| \bar{W}^n_1(s) - m_{1,k} \bar{Q}^n_{1,k}(s) \right| > \epsilon \right) = 0, \quad k = 1, 2,
\]
and
\[
\lim_{n \to \infty} \mathbb{P}\left( \sup_{s \leq T} \left| \sum_k \bar{U}^n_k(s) - \sum_k \frac{m_{2,k}}{m_{1,k}} \bar{W}^n_{1,k}(s) \right| > \epsilon \right) = 0.
\]
4. The Approximate Idle Process. The convergence of the station 1 scaled workload is straightforward; see Lemma 3.2. For station 2, the convergence hinges on the convergence of the scaled idle process, \( \tilde{I}_n^2 = \Phi(X_2^n - U_2^n) \), which, because the scaled upstream workload process does not converge, is not a simple consequence of the continuous mapping theorem. We require an approximate station 2 scaled idle process. To this end, for each system \( n \), define the stopping times \( \tau_{n}^0, \tau_{n}^1, \ldots \) and \( \nu_{n}^0, \nu_{n}^1, \ldots \) as follows. Set \( \tau_{n}^0 = 0 \) and \( \nu_{n}^0 \) equal to the beginning of the first type 2 production run at station 1. For each \( i \geq 1 \), the time \( \tau_{n}^i \) marks the beginning of the first type 1 production run that starts after \( \nu_{n}^{i-1} \), and the time \( \nu_{n}^i \) marks the beginning of the first type 2 production run after time \( \tau_{n}^i \). When interwoven, the times \( \tau_{n}^0 < \nu_{n}^0 < \tau_{n}^1 < \nu_{n}^1 < \tau_{n}^2 < \ldots \) represent the switching times of the server between the two buffers.

We define a time-change function \( T^n \) based on stopping times associated with the station 1 service cycles:

\[
T^n(t) = \sup_{i \geq 0} \{ \tau_{n}^i : \tau_{n}^i \leq t \}. \tag{41}
\]

The process \( T^n \) is an right-continuous, piecewise constant function and tracks the beginning of the current station 1 polling cycle, as a function of \( t \). Given \( T^n \) we can define our approximate station 2 idle process:

\[
I_2^n \equiv \Phi(X_2^n \circ T_2^n - U_2^n \circ T_2^n)(t), \tag{42}
\]

where for functions \( f \) and \( g \), \( f \circ g(t) = f(g(t)) \), and

\[
U_{2}^n(t) \equiv \left\{ \begin{array}{ll}
g \nu_{n}^1 W_1^n(0) + (U_{2}^n(t) - g \nu_{n}^1 W_1^n(0)) (\gamma t/n^{2/3} - 1) & t < n^{2/3}, \\
\nu_{n}^1 W_2^n(0) & \nu_{n}^1 \leq t < 2n^{2/3}, \\
\nu_{n}^1 W_1^n(0) & t \geq 2n^{2/3}
\end{array} \right.
\]

is an approximation for the station 2 upstream workload process. (We have chosen this particular process, in part, because \( U_2^n(0)/\sqrt{n} \to \nu_1 W_1(0) \).)

Notice that the approximate idle function only increases at the jump times of \( T^n \). Moreover, the interjump times are order \( \sqrt{n} \):

**Lemma 4.1** For any time \( T \), \( \epsilon > 0 \), and sufficiently large \( M \) and \( n \),

\[
\mathbb{P}\left( \sup_{i : \tau_{n}^i \leq nT} (\tau_{n+1}^i \wedge nT) - \tau_{n}^i > M\sqrt{n} \right) < \epsilon. \tag{44}
\]

**Proof.** We can split the jump interval into the contribution from each of the production runs:

\[
\mathbb{P}\left( \sup_{i : \tau_{n}^i \leq nT} (\tau_{n+1}^i \wedge nT) - \tau_{n}^i > M\sqrt{n} \right) \leq \mathbb{P}\left( \sup_{i : \nu_{n}^i \leq nT} (\nu_{n}^i \wedge nT) - \tau_{n}^i > M\rho_1\sqrt{n} \right) \tag{45}
\]

\[
+ \mathbb{P}\left( \sup_{i : \tau_{n}^i \leq nT} (\tau_{n+1}^i \wedge nT) - (\nu_{n}^i \wedge nT) > M\rho_2\sqrt{n} \right).
\]

The type 1 production either starts with \( M\rho_1\rho_2\sqrt{n}/2 \) units of work in buffer 1 or at least \( M\rho_1\sqrt{n} - M\rho_1\rho_2\sqrt{n}/2 \) units of type 1 work arrive before time \( M\rho_1\sqrt{n} \). Otherwise, the queue is empties before time \( M\rho_1\sqrt{n} \). Furthermore, there can only be \( M\rho_1\rho_2\sqrt{n}/2 \) units of type 1 work if total work (type 1 and type 2) is at least this amount. By (37) and Lemma 3.5

\[
\mathbb{P}\left( \sup_{s \leq nT} W_1^n(s) > M\rho_1\rho_2\sqrt{n}/2 \right) \tag{46}
\]

\[
+ \mathbb{P}\left( \sup_{s \leq T} S_1^n(A_1^n(sn + M\rho_1\sqrt{n})) > M\rho_1\sqrt{n} - M\rho_1\rho_2\sqrt{n}/2 \right) < \epsilon/2,
\]

for sufficiently large \( M \) and \( n \). By an analogous argument

\[
\mathbb{P}\left( \sup_{i : \tau_{n}^i \leq nT} (\tau_{n+1}^i \wedge nT) - (\nu_{n}^i \wedge nT) > M\rho_2\sqrt{n} \right) < \epsilon/2, \tag{47}
\]

for some sufficiently large \( M \) and \( n \). The relation (44) follows from (45), (46), and (47). \( \Box \)
Define the scaled processes:
\[
\tilde{I}_2^n(t) = \frac{\int_0^n (nt)^2}{\sqrt{n}} \quad \text{and} \quad \tilde{T}_n(t) = \frac{T^n(nt)}{n}, \quad t \geq 0.
\] (48)

The following lemma is an immediate consequence of Lemma 4.1.

**Lemma 4.2**
\[
\tilde{T}_n \overset{P}{\to} e,
\]

The function \(e\) is the identity function; i.e. \(e(t) = t\).

The following claims that the scaled approximate idle process indeed approximates the scaled (true) idle process.

**Proposition 4.1** For any time \(T\) and \(\epsilon > 0\),
\[
\lim_{n \to \infty} \mathbb{P} \left( \sup_{s \leq T} |\tilde{I}_2^n(s) - \tilde{I}_2^n(s)| > \epsilon \right) = 0.
\] (49)

**Proof.** Because \(\tilde{I}_2^n\) and \(\tilde{I}_2^n\) are nondecreasing,
\[
\mathbb{P} \left( \sup_{s \leq T} |\tilde{I}_2^n(s) - \tilde{I}_2^n(s)| > \epsilon \right) \leq \mathbb{P} \left( I_2^n(2n^{2/3}) > \epsilon \sqrt{n}/3 \right) + \mathbb{P} \left( I_2^n(2n^{2/3}) > \epsilon \sqrt{n}/3 \right)
\] + \mathbb{P} \left( \sup_{2n^{2/3} \leq s \leq nT} |I_2^n(s) - I_2^n(s)| > \epsilon \sqrt{n}/3 \right).

Consider the first term on the right hand side of (50):
\[
\mathbb{P} \left( I_2^n(2n^{2/3}) > \epsilon \sqrt{n}/3 \right) = \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} \left( U_2^n(s) - X_2^n(s) \right) > \epsilon \sqrt{n}/3 \right)
\] ≤ \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} \left( U_2^n(s) - \frac{m_{2,1}}{m_{1,1}} W_1^n(s) \right) > \epsilon \sqrt{n}/12 \right)
\] + \mathbb{P} \left( \frac{m_{2,1}}{m_{1,1}} \sup_{s \leq 2n^{2/3}} \left| W_1^n(s) - W_1^n(0) \right| > \epsilon \sqrt{n}/12 \right)
\] + \mathbb{P} \left( \frac{m_{2,1}}{m_{1,1}} W_1^n(0) - W_2^n(0) > \epsilon \sqrt{n}/12 \right)
\] + \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} \left| W_2^n(0) - X_2^n(s) \right| > \epsilon \sqrt{n}/12 \right).

For the first term on the right hand side of (51), we show a stronger result. Notice that
\[
\mathbb{P} \left( \sup_{s \leq nT} \left( U_2^n(s) - \frac{m_{2,1}}{m_{1,1}} W_1^n(s) \right) > \epsilon \sqrt{n}/12 \right)
\] ≤ \mathbb{P} \left( \sup_{s \leq nT} \left| U_2^n(s) - \sum_k m_{2,k} Q_{1,k}^n(s) \right| > \epsilon \sqrt{n}/24 \right)
\] + \mathbb{P} \left( \frac{m_{2,1}}{m_{1,1}} \sup_{s \leq nT} \left| \sum_k m_{1,k} Q_{1,k}(s) - W_1^n(s) \right| > \epsilon \sqrt{n}/24 \right)
\] + \mathbb{P} \left( \sup_{s \leq nT} \left( \sum_k m_{2,k} Q_{1,k}(s) - \frac{m_{2,1}}{m_{1,1}} \sum_k m_{1,k} Q_{1,k}(s) \right) > 0 \right) < \epsilon/8,

so that, by (8), (52) and Proposition 3.1,
\[
\lim_{n \to \infty} \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} \left( U_2^n(s) - \frac{m_{2,1}}{m_{1,1}} W_1^n(s) \right) > \epsilon \sqrt{n}/12 \right) = 0.
\] (53)
As for the remaining terms on the right hand side of (51),

$$
\lim_{n \to \infty} \mathbb{P} \left( \frac{m_{2,1}}{m_{1,1}} \sup_{s \leq K_n \sqrt{n}} \left| W^n_t(s) - W^n_t(0) \right| > \epsilon \sqrt{n}/12 \right) = 0
$$

(54)

follows from Lemma 3.5

$$
\lim_{n \to \infty} \mathbb{P} \left( \frac{m_{2,1}}{m_{1,1}} \left| W^n_1(0) - W^n_2(0) \right| > \epsilon \sqrt{n}/12 \right) = 0
$$

(55)

follows from (9) and (11), and

$$
\lim_{n \to \infty} \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} \left| W^n_2(0) - X^n_2(s) \right| > \epsilon \sqrt{n}/12 \right) = 0
$$

(56)

follows from Lemma 3.5 and the fact that $X^n_2(0) = W^n_2(0)$. From (51) and (53) - (56)

$$
\lim_{n \to \infty} \mathbb{P} \left( I^n_2(2n^{2/3}) > \epsilon \sqrt{n}/3 \right) = \lim_{n \to \infty} \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} U^n_2(s) - X^n_2(s) > \epsilon \sqrt{n}/3 \right) = 0.
$$

(57)

For the second term in (50), notice that

$$
\mathbb{P} \left( I^n_2(2n^{2/3}) > \epsilon \sqrt{n}/3 \right) \leq \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} \gamma_p W^n_1(0) - X^n_2(s) > \epsilon \sqrt{n}/6 \right) + \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} U^n_2(s) - X^n_2(s) > \epsilon \sqrt{n}/6 \right)
$$

(58)

The second term on the right hand side was handled once already. As for the first term, $\gamma_p \leq \frac{m_{2,1}}{m_{1,1}}$ so that, by (55), (56), (57), and (58), we have

$$
\lim_{n \to \infty} \mathbb{P} \left( I^n_2(2n^{2/3}) > \epsilon \sqrt{n}/3 \right) = 0.
$$

(59)

Finally, for the third term in (50), one can verify

$$
\sup_{2n^{2/3} \leq s \leq nT} (I^n_2(s) - I^n_2(n)) \leq \sup_{2n^{2/3} \leq s \leq nT} |X^n_2(s) - X^n_2(T^n(s))| + \sup_{2n^{2/3} \leq s \leq nT} (U^n_2(s) - U^n_2(T^n(s))) + I^n_2(2n^{2/3}) + I^n_2(2n^{2/3}).
$$

(60)

As for the first term on the right hand side of (60),

$$
\mathbb{P} \left( \sup_{s \leq nT} |X^n_2(s) - X^n_2(T^n(s))| > \epsilon \sqrt{n}/6 \right)
$$

$$
= \mathbb{P} \left( \sup_{i: \tau^n_i \leq nT} |X^n_2((\tau^n_i + t) \wedge nT) - X^n_2(\tau^n_i)| > \epsilon \sqrt{n}/6 \right)
$$

(61)

$$
\leq \mathbb{P} \left( \sup_{s \leq T, t \leq M} |X^n_2(s + \sqrt{n}t) - X^n_2(\tau^n_i)| > \epsilon \sqrt{n}/6 \right) + \mathbb{P} \left( \sup_{i: \tau^n_i \leq nT} \left( \tau^n_{i+1} \wedge nT \right) - \tau^n_i > M \sqrt{n} \right).
$$

(62)

It follows from Lemmas 3.5 and 4.1 and (61) that

$$
\lim_{n \to \infty} \mathbb{P} \left( \sup_{s \leq nT} |X^n_2(s) - X^n_2(T^n(s))| > \epsilon \sqrt{n}/6 \right) = 0.
$$

(63)

As for the second term on the right hand side of (60), we would like to show that

$$
\lim_{n \to \infty} \mathbb{P} \left( \sup_{2n^{2/3} \leq s \leq nT} U^n_2(s) - U^n_2(T^n(s)) > \epsilon \sqrt{n}/6 \right) = 0.
$$

(64)
Without loss of generality, let us assume that, for every \( n \), \( nT \) coincides with the beginning of some service cycle \( (nT = \tau_i^n \text{ for some } i) \). For convenience, define \( J^n \equiv \{ i : 2n^{2/3} \leq \tau_i^n \leq nT \} \). Then

\[
\mathbb{P} \left( \sup_{2n^{2/3} \leq s \leq nT} \left| U_2^n(s) - U_2^n(T^n(s)) \right| > \epsilon \sqrt{n}/6 \right) \\
\leq \mathbb{P} \left( \sup_{i \in J^n} \left| U_2^n(\tau_i^n) - U_2^n(T^n(\tau_i^n+1)) \right| > \epsilon \sqrt{n}/24 \right) \\
+ \mathbb{P} \left( \sup_{s \leq nT} \left| U_2^n(s) - \sum_k \frac{m_{2,k}}{m_{1,k}} W_{1,k}^n(s) \right| > \epsilon \sqrt{n}/24 \right) \\
+ \mathbb{P} \left( \sup_{i \in J^n, s \leq \nu_i^n - \tau_i^n} \sum_k \frac{m_{2,k}}{m_{1,k}} (W_{1,k}^n(\tau_i^n + s) - W_{1,k}^n(\tau_i^n)) > \epsilon \sqrt{n}/24 \right) \\
+ \mathbb{P} \left( \sup_{i \in J^n, s \leq \nu_i^n - \tau_i^n} \sum_k \frac{m_{2,k}}{m_{1,k}} (W_{1,k}^n(\nu_i^n + s) - W_{1,k}^n(\tau_i^{n+1})) > \epsilon \sqrt{n}/24 \right).
\] (65)

The idea behind (65) is to exploit the fluid like nature of the upstream workload over the course of a service cycle. The first term simply states that the upstream workload does not change much over the course of a cycle. In Proposition 4.22 which appears next, we argue that \( U_2^n \circ T^n \) converges. This fact, along with the boundedness of the interjump times of \( T^n \) in Lemma 4.1 imply

\[
\lim_{n \to \infty} \mathbb{P} \left( \sup_{i \in J^n} \left| U_2^n(\tau_i^n) - U_2^n(T^n(\tau_i^n+1)) \right| > \epsilon \sqrt{n}/36 \right) = 0.
\] (66)

The second term states the upstream workload in terms of the station 1 workload contributions from both queues and is handled by Corollary 3.1. The third and fourth terms the right hand side of (65) use the fact that we expect the upstream workload to decrease over \( (\tau_i^n, \nu_i^n) \) and increase over \( (\nu_i^n, \tau_i^{n+1}) \).

Recall that \( W_1^n = W_1^{n+1} + W_1^n \) so that

\[
\mathbb{P} \left( \sup_{i \in J^n, s \leq \nu_i^n - \tau_i^n} \sum_k \frac{m_{2,k}}{m_{1,k}} (W_{1,k}^n(\tau_i^n + s) - W_{1,k}^n(\tau_i^n)) > \epsilon \sqrt{n}/24 \right) \\
\leq \mathbb{P} \left( \sup_{i \in J^n, s \leq \nu_i^n - \tau_i^n} \left( \frac{m_{2,1}}{m_{1,1}} - \frac{m_{2,2}}{m_{1,2}} \right) (W_{1,1}^n(\tau_i^n + s) - W_{1,1}^n(\tau_i^n)) > \epsilon \sqrt{n}/48 \right) \\
+ \mathbb{P} \left( \sup_{i \in J^n, s \leq \nu_i^n - \tau_i^n} \frac{m_{2,2}}{m_{1,2}} (W_{1,1}^n(\nu_i^n + s) - W_{1,1}^n(\tau_i^{n+1})) > \epsilon \sqrt{n}/48 \right).
\] (67)

By Lemma 3.5 and Lemma 4.1,

\[
\lim_{n \to \infty} \mathbb{P} \left( \sup_{i \in J^n, s \leq \nu_i^n - \tau_i^n} \frac{m_{2,2}}{m_{1,2}} (W_{1,1}^n(\nu_i^n + s) - W_{1,1}^n(\tau_i^{n+1})) > \epsilon \sqrt{n}/48 \right) = 0.
\] (68)

Throughout the interval \((\tau_i^n, \nu_i^n)\), the station 1 processes type 1 jobs exclusively and never idles, so that for \( s \leq \nu_i^n - \tau_i^n \),

\[
W_{1,1}^n(\tau_i^n + s) - W_{1,1}^n(\tau_i^n) = S_{1,1}^n(A_{1,1}^n(\tau_i^n + s) - S_{1,1}^n(A_{1,1}^n(\tau_i^n)) - s.
\]

The server processes \( s \) units of work, and we only expect \( \rho_{1,1}^n s \leq s \) units of work to arrive. It follows from (8), Lemma 3.5 and Lemma 4.1 that

\[
\lim_{n \to \infty} \mathbb{P} \left( \sup_{i \in J^n, s \leq \nu_i^n - \tau_i^n} \left( \frac{m_{2,1}}{m_{1,1}} - \frac{m_{2,2}}{m_{1,2}} \right) (W_{1,1}^n(\tau_i^n + s) - W_{1,1}^n(\tau_i^n)) > \epsilon \sqrt{n}/48 \right) = 0.
\] (69)

Then from (67), (68), and (69) it follows that

\[
\lim_{n \to \infty} \mathbb{P} \left( \sup_{i \in J^n, s \leq \nu_i^n - \tau_i^n} \sum_k \frac{m_{2,k}}{m_{1,k}} (W_{1,k}^n(\tau_i^n + s) - W_{1,k}^n(\tau_i^n)) > \epsilon \sqrt{n}/24 \right) = 0.
\] (70)

One can similarly show that

\[
\lim_{n \to \infty} \mathbb{P} \left( \sup_{i \in J^n, s \leq \nu_i^n - \tau_i^n} \sum_k \frac{m_{2,k}}{m_{1,k}} (W_{1,k}^n(\nu_i^n + s) - W_{1,k}^n(\tau_i^{n+1})) > \epsilon \sqrt{n}/24 \right) = 0.
\] (71)
Hence, (63) follows from (65), (66), (70), (71), and Corollary 3.1. Finally, the result (69) follows from (50), (57), (59), (60), (63), and (64).

The convergence of the station 2 scaled approximate idle process depends in part on the convergence of the station 2 scaled approximate sampled upstream workload process, $\bar{U}_2^n \circ T^n$. (The convergence of this process was also used for showing (66)).

**Proposition 4.2**

$\bar{U}_2^n \circ T^n \Rightarrow \gamma_p \bar{W}_1^n$

**Proof.** We will show that for any $\epsilon > 0$ and time $T$,

$$\lim_{n \to \infty} \mathbb{P} \left( \sup_{s \leq T} \left| \bar{U}_2^n(T(s)) - \gamma_p \bar{W}_1^n(s) \right| > \epsilon \right) = 0,$$

(72)

regardless of the polling discipline employed at station 1. We start by using the definition of the approximate upstream function. We exploit the fact that the approximate function, for each time $s$ in the interval $[n^{2/3}, 2n^{2/3}]$, is a convex combination of the constant $W_1^n(0)$ and the true scaled upstream workload function, so that

$$\mathbb{P} \left( \sup_{s \leq T} \left| \bar{U}_2^n(T(s)) - \gamma_p \bar{W}_1^n(s) \right| > \epsilon \right) \leq \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} \gamma_p \left| W_1^n(0) - W_1^n(s) \right| > \epsilon \sqrt{n}/2 \right)$$

(73)

$$+ \mathbb{P} \left( \sup_{n^{2/3} \leq s \leq nT} \left| U_2^n(T(s)) - \gamma_p W_1^n(s) \right| > \epsilon \sqrt{n}/2 \right).$$

By Lemma 3.5,

$$\lim_{n \to \infty} \mathbb{P} \left( \sup_{s \leq 2n^{2/3}} \gamma_p \left| W_1^n(0) - W_1^n(s) \right| > \epsilon \sqrt{n}/2 \right) = 0,$$

(74)

For the second term on the right hand side of (73), we would like to show that

$$\lim_{n \to \infty} \mathbb{P} \left( \sup_{n^{2/3} \leq s \leq nT} \left| U_2^n(T(s)) - \gamma_p W_1^n(s) \right| > \epsilon \sqrt{n}/2 \right) = 0,$$

(75)

thereby completing the proof. As in the proof of Proposition 4.1, we define $J^n = \{ i : n^{2/3} \leq \tau_i^n \leq nT \}$ and whenever we write $\tau_i^n$ and $\nu_i^n$ we really mean $\tau_i^n \wedge nT$ and $\nu_i^n \wedge nT$. It helps to expand the second term slightly so that we only compare the station 2 sampled upstream function with the station 1 sampled workload function:

$$\mathbb{P} \left( \sup_{n^{2/3} \leq s \leq nT} \left| U_2^n(T(s)) - \gamma_p W_1^n(s) \right| > \epsilon \sqrt{n}/2 \right)$$

$$\leq \mathbb{P} \left( \sup_{i \in J^n} \left| U_2^n(\tau_i^n) - \gamma_p W_1^n(\tau_i^n) \right| > \epsilon \sqrt{n}/4 \right)$$

$$+ \mathbb{P} \left( \sup_{i \in J^n, n^{2/3} \leq s \leq \tau_i^n} \gamma_p \left| W_1^n(\tau_i^n + s) - W_1^n(\tau_i^n) \right| > \epsilon \sqrt{n}/4 \right).$$

By Lemma 3.5 and Lemma 4.1, the second term goes to zero as $n \to \infty$ and (75) will hold once we demonstrate

$$\lim_{n \to \infty} \mathbb{P} \left( \sup_{i \in J^n} \left| U_2^n(\tau_i^n) - \gamma_p W_1^n(\tau_i^n) \right| > \epsilon \sqrt{n}/4 \right) = 0.$$

(76)

To do so, and complete the proof, we must consider each polling discipline separately.

Case 1: Exhaustive Service. Note that $\gamma_E = m_{2,1}/m_{1,1}$. Starting with cycle 1 (which occurs strictly after time 0) the type 2 queue is empty at the beginning of each service cycle; i.e., $Q_{2,1}(\tau_i^n) = W_{1,2}^n(\tau_i^n) = 0$ for each $i \geq 1$. Equation (76) follows immediately from Corollary 3.1.

Case 2: Gated Service. Note that

$$\mathbb{P} \left( \sup_{i \in J^n} \left| U_2^n(\tau_i^n) - \gamma_p W_1^n(\tau_i^n) \right| > \epsilon \sqrt{n}/4 \right) \leq \mathbb{P} \left( \sup_{i \in J^n} \left| U_2^n(\tau_i^n) - \sum_{k} \frac{m_{2,k}}{m_{1,k}} W_{1,k}^n(\tau_i^n) \right| > \epsilon \sqrt{n}/8 \right)$$

$$+ \mathbb{P} \left( \sup_{i \in J^n} \left| \sum_{k} \frac{m_{2,k}}{m_{1,k}} W_{1,k}^n(\tau_i^n) - \gamma_G W_1^n(\tau_i^n) \right| > \epsilon \sqrt{n}/12 \right).$$
Equation (76) follows from Corollary 3.1 and Proposition 5.1 in the following section.

5. Properties of Gated Service. The key to Proposition 4.2 from the previous section was Equation (76), which related the station 2 upstream workload with the station 1 immediate workload. What is intriguing here is that one only needs to compare these processes at the beginning of the station 1 service cycle. It turns out that the upstream workload achieves a maximum at these times (in the limit). For exhaustive service, the analysis is straightforward. At the beginning of the station 1 service cycle, all of the station 1 jobs are type 1. For gated service the analysis is more involved. For one, it is not the case that all of the station 1 jobs are type 1 at the beginning of the station 1 service cycle. Our expectation is that, in equilibrium, given the immediate workload for station 1 at the beginning of its service cycle, roughly a fixed fraction of this workload will be from the sum of the service times of the type 1 jobs present. The problem is that the system does not achieve an equilibrium immediately, as is the case under exhaustive service. Thus, we have to let the process settle. Hence, the $n^{2/3}$ term throughout many of the equations from Section 4. The following proposition captures this settling property.

**Proposition 5.1** For any $\epsilon > 0$ and time $T$,

$$\lim_{n \to \infty} P \left( \sup_{i: \tau_i^n \leq nT} \left| \frac{W_{1,1}^n(\tau_i^n)}{1-\rho_{1,1} \rho_{1,2}} - P \right| \geq \epsilon \sqrt{n} \right) = 0 \quad (77)$$

and

$$\lim_{n \to \infty} P \left( \sup_{i: \tau_i^n \leq nT} \left| \frac{W_{1,2}^n(\tau_i^n)}{1-\rho_{1,1} \rho_{1,2}} - P \right| \geq \epsilon \sqrt{n} \right) = 0 \quad (78)$$

Note: Equations (77) and (78) could be stated equivalently using stopping times $\nu_i^n$ rather than $\tau_i^n$, where the fraction $\frac{\rho_{1,1}}{1-\rho_{1,1} \rho_{1,2}}$ is replaced by $\frac{\rho_{1,2}}{1-\rho_{1,1} \rho_{1,2}}$ and $\frac{\rho_{1,2}}{1-\rho_{1,1} \rho_{1,2}}$ is replaced by $\frac{\rho_{1,1}}{1-\rho_{1,1} \rho_{1,2}}$.

The key to the proof of Proposition 5.1 is the following lemma. It states that error bounds on the deviation in (77) get geometrically smaller between cycles.

**Lemma 5.1** Assume $\sup_{s \leq T} W_1^s(s) \leq M \sqrt{n}$. For any $\epsilon > 0$ and sufficiently large $n$,

$$P \left( \sup_{i: \tau_i^n \leq nT} \left| \frac{W_{1,1}^n(\tau_i^n)}{1-\rho_{1,1} \rho_{1,2}} - \rho_{1,1} \rho_{1,2} \right| \frac{W_{1,1}^n(\tau_i^n)}{1-\rho_{1,1} \rho_{1,2}} > \epsilon \sqrt{n} \right) < \epsilon. \quad (79)$$

**Proof.** Fix $\epsilon$. The length of the $i$th cycle, $\tau_i^n - \tau_i^n$ is equal to the total workload at the beginning of the cycle plus the type 2 workload that arrives during the type 1 production run:

$$\tau_i^n - \tau_i^n = W_1^n(\tau_i^n) + W_{1,2}^n(\tau_i^n) - \rho_{1,2} W_{1,2}^n(\tau_i^n).$$

It follows that

$$\left| \frac{W_{1,1}^n(\tau_i^n)}{1-\rho_{1,1} \rho_{1,2}} - \rho_{1,1} \rho_{1,2} \right| \frac{W_{1,1}^n(\tau_i^n)}{1-\rho_{1,1} \rho_{1,2}} \leq \frac{\rho_{1,1}}{1-\rho_{1,1} \rho_{1,2}} \left| W_{1,1}^n(\tau_i^n) - W_{1,1}^n(\tau_i^n) \right| + \rho_{1,1} |W_{1,2}^n(\tau_i^n) - W_{1,2}^n(\tau_i^n)|.$$

We bound each of the terms on the right hand side of (80) separately. Because the total workload is bounded by $M \sqrt{n}$ the length of the cycle is bounded by $2M \sqrt{n}$. Hence, by Lemma 3.5,

$$P \left( \sup_{i: \tau_i^n \leq nT} \left| \frac{W_{1,1}^n(\tau_i^n)}{1-\rho_{1,1} \rho_{1,2}} - \rho_{1,1} \rho_{1,2} \right| > \epsilon \sqrt{n} / 3 \right) < \epsilon / 3. \quad (81)$$

As for the second term on the right hand side of (80), the amount of work in buffer 1 at the beginning of cycle $i+1$ is equal to the type 1 work that accumulates during cycle $i$. Because the cycle length is
bounded by $2M\sqrt{n}$, the accumulated work should be roughly proportional to the cycle length. More precisely, by Lemma 3.5
\[
P\left( \sup_{i: \tau_n^i \leq nT} \left| W_{1,1}^n(\tau_n^i) - \rho_{1,1} \left( \tau_n^i - \tau_n^0 \right) \right| > \epsilon \sqrt{n}/3 \right) < \epsilon/3.
\] (82)

Likewise, the amount of type 2 work that arrives during the type 1 production run of cycle $i$ is proportional to the length of that production run. Again, by Lemma 3.5
\[
P\left( \sup_{i: \tau_n^i \leq nT} \left| W_{1,2}^n(\tau_n^i) - \rho_{1,2} W_{1,1}^n(\tau_n^i) \right| > \epsilon \sqrt{n}/3 \right) < \epsilon/3.
\] (83)

The result (79) follows from (80) – (83) and the fact that $\rho_{1,1} < 1 - \rho_{1,1}\rho_{1,2}$. \qed

Now we return to Proposition 5.1

**Proof of Proposition 5.1.** Without loss of generality, fix $\epsilon < M$. Set
\[
I \equiv \left\lceil \frac{\log(M/\epsilon)}{\log 4} \right\rceil + 2 \quad \text{and} \quad \epsilon' \equiv \frac{\epsilon}{2I} < \frac{\epsilon}{4}.
\] (84)

Consider the event in (77). If, for some cycle $i \geq I$, $W_{1,1}^n(\tau_n^i)$ deviates from $\frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^i)$ by more than $\epsilon \sqrt{n}$, there is a first cycle in which this happens. That is,
\[
P\left( \sup_{i \geq I, \tau_n^i \leq nT} \left| W_{1,1}^n(\tau_n^i) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^i) \right| > \epsilon \sqrt{n} \right)
\]
\[
\leq P\left( \left| W_{1,1}^n(\tau_n^I) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^I) \right| > \epsilon \sqrt{n} \right)
+ P \left( \exists i \geq I : \tau_n^i \leq nT, \left| W_{1,1}^n(\tau_n^{i+1}) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^{i+1}) \right| > \epsilon \sqrt{n},
\right.
\]
\[
\left. \quad \text{and} \quad \left| W_{1,1}^n(\tau_n^i) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^i) \right| \leq \epsilon \sqrt{n} \right)
\] (85)

It follows from Lemma 5.1, the bound on $W_{1,1}^n$, and the fact that $\rho_{1,1}\rho_{1,2} \leq 1/4$, that for sufficiently large $n$,
\[
P\left( \left| W_{1,1}^n(\tau_n^{i+1}) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^{i+1}) \right| > 4^{-I} M \sqrt{n} + 2\epsilon' \sqrt{n} \right) < i\epsilon', \quad i = 1, 2, \ldots
\]

In particular,
\[
P\left( \left| W_{1,1}^n(\tau_n^I) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^I) \right| > 4^{-I+1} M \sqrt{n} + 2\epsilon' \sqrt{n} \right) < (I-1)\epsilon'.
\] (86)

It is follows from Lemma 5.1 that
\[
P\left( \exists i \geq I : \tau_n^i \leq nT, \left| W_{1,1}^n(\tau_n^{i+1}) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^{i+1}) \right| > \epsilon \sqrt{n},
\right.
\]
\[
\left. \quad \text{and} \quad \left| W_{1,1}^n(\tau_n^i) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^i) \right| \leq \epsilon \sqrt{n} \right)
\]
\[
\leq P\left( \sup_{i \geq I: \tau_n^i \leq nT} \left| W_{1,1}^n(\tau_n^{i+1}) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^{i+1}) \right|
\]
\[
- \left| W_{1,1}^n(\tau_n^i) - \frac{\rho_{1,1}}{1 - \rho_{1,1}\rho_{1,2}} W_{1,1}^n(\tau_n^i) \right| / 2 > \epsilon \sqrt{n}/2 \right) < \epsilon/2.
\] (87)

Thus (77) follows from (84) – (87). \qed

The following lemma relates the maximum queue size at station 1 (once the process achieves its cyclic pattern) to the workload at that station.
Lemma 5.2 Suppose \( \sup_{s \leq n T} W^n_1(s) \leq M \sqrt{n} \). For any \( \epsilon > 0 \) and sufficiently large \( n \) (which depends on \( \epsilon \)),

\[
\mathbb{P} \left( \sup_{\tau^n_1 \leq s \leq n T} m_{1,1} Q^n_{1,1}(s) - \frac{\rho_{1,1}}{1 - \rho_{1,1} \rho_{1,2}} W^n_1(s) > \epsilon \sqrt{n} \right) < \epsilon.
\]

Proof. We replace the queue length quantity with its workload equivalent:

\[
\mathbb{P} \left( \sup_{\tau^n_1 \leq s \leq n T} m_{1,1} Q^n_{1,1}(s) - \frac{\rho_{1,1}}{1 - \rho_{1,1} \rho_{1,2}} W^n_1(s) > \epsilon \sqrt{n} \right)
\]

\[
\leq \mathbb{P} \left( \sup_{\tau^n_1 \leq s \leq n T} \left| m_{1,1} Q^n_{1,1}(s) - W^n_{1,1}(s) \right| > \epsilon \sqrt{n}/4 \right) \quad (89)
\]

\[
+ \mathbb{P} \left( \sup_{\tau^n_1 \leq s \leq n T} W^n_{1,1}(s) - \frac{\rho_{1,1}}{1 - \rho_{1,1} \rho_{1,2}} W^n_1(s + \tau^n_1) > \epsilon \sqrt{n}/2 \right)
\]

By Lemma 3.5 and the fact that \( W^n_1(\cdot) \leq M \sqrt{n} \), we have

\[
\mathbb{P} \left( \sup_{\tau^n_1 \leq s \leq n T} Q_{1,1}(s) > 2M \sqrt{n}/m_{1,1} \right) < \epsilon/4.
\]

Applying Lemma 3.5 again, yields

\[
\mathbb{P} \left( \sup_{\tau^n_1 \leq s \leq n T} \left| m_{1,1} Q^n_{1,1}(s) - W^n_{1,1}(s) \right| > \epsilon \sqrt{n}/4 \right) < \epsilon/2.
\]

As for the second term on the right hand side of (89), we will show that

\[
\mathbb{P} \left( \sup_{\tau^n_1 \leq s \leq n T} W^n_{1,1}(s) - \frac{\rho_{1,1}}{1 - \rho_{1,1} \rho_{1,2}} W^n_1(s) > \epsilon \sqrt{n}/2 \right)
\]

\[
\leq \mathbb{P} \left( \sup_{i \geq I^*} \left| W^n_{1,1}(\tau^n_1) - \frac{\rho_{1,1}}{1 - \rho_{1,1} \rho_{1,2}} W^n_1(\tau^n_1) \right| > \epsilon \sqrt{n}/6 \right) \quad (91)
\]

\[
+ \mathbb{P} \left( \sup_{i \geq I^*} \sup_{\tau^n_1 \leq s \leq \tau^n_{i+1} - \tau^n_1} \left| W^n_1(\tau^n_1) - W^n_1(\tau^n_1 + s) \right| > \epsilon \sqrt{n}/6 \right)
\]

\[
+ \mathbb{P} \left( \sup_{i \geq I^*} \sup_{\tau^n_1 \leq s \leq \nu^n - \tau^n_1} W^n_{1,1}(\tau^n_1 + s) - W^n_{1,1}(\tau^n_1) > \epsilon \sqrt{n}/6 \right)
\]

\[
+ \mathbb{P} \left( \sup_{i \geq I^*} \sup_{\tau^n_1 \leq s \leq \nu^n - \tau^n_1} W^n_{1,1}(\tau^n_1 + s) - W^n_{1,1}(\tau^n_{i+1}) > 0 \right) < \epsilon/2.
\]

First notice we have split the interval \( [\tau^n_1, nT] \) into smaller intervals corresponding with the service cycles. The first term on the right hand side observes the type 1 workload process \( W^n_{1,1} \) only at the beginning of each cycle. At these times, by Proposition 5.1 the type 1 workload process is asymptotically close to a fixed fraction of total station workload; the probability of the first term is less than \( \epsilon/6 \). The remainder of the terms show that, asymptotically, the type 1 workload achieves maxima at these cycle start times. By Lemma 3.5 the second term is less than \( \epsilon/6 \). Since during \( [\tau^n_1, \nu^n] \) all server effort is dedicated to type 1 work, Lemma 3.5 can be employed again, bounding the third term by \( \epsilon/6 \) as well. As for the last term, notice that \( W^n_{1,1}(\cdot) \) is nondecreasing on \([\nu^n, \tau^n_{i+1} - \tau^n_1]\). So the probability of the last term is zero. Thus (91) holds. The result (88) follows from (89) – (91).

6. Proof of Theorem 2.1 We now prove our main result.

Proof of Theorem 2.1. It was already shown that the station 1 total workload process converges to a reflected Brownian motion. As for the second station’s workload, we need to show that the scaled idle time \( I^n_1 = \Phi(\bar{X}^n_1 - \bar{U}^n_1) \) converges. To do this, we employ the approximate scaled idle time process \( \tilde{I}^n_2 = \Phi(\bar{X}^n_2 \circ T^n - \bar{U}^n_2 \circ T^n) \). By Lemmas 3.2 and 4.2 Proposition 4.2 and the Continuous Mapping Theorem, we have

\[
\left( \tilde{W}^n_1, \bar{X}^n, \tilde{I}^n_2 \right) \Rightarrow \left( \bar{W}_1, \bar{X}, \bar{I}_2 \right).
\]

(92)
where
\[ \tilde{I}_2 \equiv \Phi(\tilde{X}_2 - \gamma p \tilde{W}_1). \]
It follows from Proposition 4.1 and Theorem 3.1 of Billingsley [3] that
\[ \tilde{I}_2^n \Rightarrow \tilde{I}_2 \]
and this convergence takes place jointly with the terms in (92). The convergence in (12) holds.

\[ \square \]

References

[12] Jennings, O. B. A heavy traffic limit theorem for a polling station under limited service. \emph{In Preparation}.
