Advances in Applied Probability, 1995

The Probability of Large Queue Lengths and Waiting Times in a Heterogeneous Multiserver Queue Part I: Tight Limits

John. S. Sadowsky[1]
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-5706 USA

Wojciech Szpankowski[2]
Department of Computer Science
Purdue University
West Lafayette, IN 47907 USA

Abstract

We consider a multiserver queuing process specified by i.i.d. interarrival time, batch size and service time sequences. In the case that different servers have different service time distributions we say the system is *heterogeneous*. In this paper we establish conditions for the queuing process to be characterized as a geometrically Harris recurrent Markov chain, and we characterize the stationary probabilities of large queue lengths and waiting times. The queue length is asymptotically geometric and the waiting time is asymptotically exponential. Our analysis is a generalization of the well known characterization of the GI/G/1 queue obtained using classical probabilistic techniques of exponential change of measure and renewal theory.

American Mathematical Society subject classifications: Primary 60K25; Secondary 60F10, 60J05, 60K15, 60K20.

Key Words: Queueing theory, Harris recurrent Markov chains, Foster-Lyapunov theory, Markov additive processes, renewal theory.

[1] J. S. Sadowsky is supported by the National Science Foundation (NCR-9003007).

[2] W. Szpankowski is supported by the National Science Foundation (NCR-9206315 and CCR-9201078), AFOSR (90-0107), and by the National Library of Medicine (R01 LM05118).

1. Introduction and Summary

We consider a multiserver queuing process with batch arrivals that is determined by an i.i.d. interarrival time sequence, an i.i.d. batch size sequence, and an i.i.d. service time sequence for each of c servers, $1 \le c < \infty$. This model occurs in numerous applications, including manufacturing processes, multi-processor computer networks, telecommunications networks and various types of service counters. Different servers may have different service time distributions, in which case we say the system is *heterogeneous*. For example, a bank may have some highly efficient tellers and some less experienced tellers who will generally take longer to provide their services. Another example of a heterogeneous system is a distributed computer network with dissimilar processors.

The results presented here extend the classical GI/G/1 analysis that is generally credited to Feller (1971, Ch. XII). We briefly review the GI/G/1 case in order to place the new results in the proper context.

Let X be a random variable with distribution $F_X(\cdot)$. Let $\Lambda_X(\eta) \stackrel{\text{def}}{=} \log(\mathrm{E}[e^{\eta X}])$ be its *cumulant function*, and put $\mathcal{D}_X \stackrel{\text{def}}{=} \{\eta\colon\ \Lambda_X(\eta) < \infty\}$. As a mapping $\mathbf{R} \to (-\infty, +\infty]$, $\Lambda_X(\cdot)$ is convex (so $\mathcal{D}_X$ is an interval) and l.s.c. (lower semicontinuous). $\Lambda_X(\cdot)$ is analytic on $\mathcal{D}_X^o$, $\Lambda_X(0) = 0$, and when $\mathcal{D}_X^o \ne \emptyset$ we have $\Lambda_X'(0) = \mathrm{E}[X]$ and $\Lambda_X''(0) = \mathrm{var}[X]$ where the derivatives are limits of interior derivatives in the case that 0 is an endpoint of $\mathcal{D}_X$. For any $\eta \in \mathcal{D}_X$, define the *twisted distribution*

$$F_X^{(\eta)}(dx) \quad \stackrel{\text{def}}{=} \quad \exp\Big(\eta x \; - \; \Lambda_X(\eta)\Big)\; F_X(dx). \tag{1}$$

The twisted distribution's cumulant function is $\Lambda_X^{(\eta)}(\zeta) = \Lambda_X(\eta + \zeta) - \Lambda_X(\eta)$, hence, $\Lambda_X'(\eta) = \mathrm{E}^{(\eta)}[X]$ and $\Lambda_X''(\eta) = \mathrm{var}^{(\eta)}[X]$. When $\{X_k\}$ is an i.i.d. sequence with $\mathcal{L}(X_k) = F_X(\cdot)$, we call the i.i.d. sequence distribution with $\mathcal{L}(X_k) = F_X^{(\eta)}(\cdot)$ the *η-conjugate distribution.*

The GI/G/1 queuing process is determined by the i.i.d. interarrival time and service time sequences $\{A_k\}$ and $\{B_k\}$. Define the *arrival rate* $\lambda \stackrel{\text{def}}{=} \mathrm{E}[A]^{-1}$ and the *service rate* $\mu \stackrel{\text{def}}{=} \mathrm{E}[B]^{-1}$. Then it is well known that the *waiting time* sequence $\{W_k\}$ is a Markov chain, and it is positive Harris recurrent, with stationary distribution $\pi(\cdot)$, if and only if $\lambda < \mu$. Define $\Lambda(\eta) \stackrel{\text{def}}{=} \Lambda_A(-\eta) + \Lambda_B(\eta)$, $\theta \stackrel{\text{def}}{=} \sup\{\eta\colon\ \Lambda(\eta) \le 0\}$ and $\omega \stackrel{\text{def}}{=} \exp(-\Lambda_B(\theta))$. One finds that $\Lambda'(0) = (\lambda - \mu)/(\lambda\mu)$, hence, by convexity, $\theta > 0$ implies $\Lambda'(0) < 0$, which in turn implies the stability condition $\lambda < \mu$. The classical result states that there exists constants $0 < c_W, c_Q < \infty$ such that

$$\mathcal{P}_\pi(\, W_k \ge w\,) \quad \sim \quad c_W\, e^{-\theta w} \quad \text{and} \quad \mathcal{P}_\pi(\, Q_k \ge n\,) \quad \sim \quad c_Q\, \omega^n \tag{2}$$

if and only if $0 < \theta < \infty$, $\Lambda(\theta) = 0$ and $\Lambda'(\theta) < \infty$. We refer to (2) as the *tight limits*.

The first step of the classical proof is to relate the queuing problem to a level crossing problem via Lindley's identity. Let $\{X_k\}$ be an i.i.d. sequence with distribution $\mathcal{L}(X_k) = \mathcal{L}(B - A)$, and observe that $\Lambda_X(\cdot) = \Lambda(\cdot)$. Define the random walk $\widetilde{W}_k = \sum_{\kappa=1}^{k} X_\kappa$ and the stopping time $K_w = \inf\{k \geq 1\colon \widetilde{W}_k \geq w\}$. Then $\mathcal{P}_\pi(W_k \geq w) = \mathcal{P}_0(K_w < \infty)$. (See Feller (1971, p. 198).) Next, $\Lambda(\theta) = 0$ and convexity gives $\mathrm{E}^{(\theta)}[X_k] = \Lambda'(\theta) > 0$, which implies $\mathcal{P}^{(\theta)}(K_w < \infty) = 1$. Thus, by change of measure to the θ-conjugate distribution we obtain

$$\mathcal{P}_\pi(W_k \geq w) \quad = \quad \mathcal{P}_0(K_w < \infty) \quad = \quad \mathrm{E}^{(\theta)}\Big[\exp\Big(-\theta\,(\,\widetilde{W}_{K_w} - w\,)\Big)\Big]\ e^{-\theta w}.$$

Finally, since $\mathrm{E}^{(\theta)}[X_k] = \Lambda'(\theta) < \infty$ (by assumption), it is a well known consequence of the renewal theorem that the "overshoot" $\widetilde{W}_{K_w} - w$ has a limiting distribution, and hence, $\mathrm{E}^{(\theta)}[\exp(-\theta(\widetilde{W}_{K_w} - w))\,] \to c_W$ with $0 < c_W < \infty$. If any of the conditions $0 < \theta < \infty$, $\Lambda(\theta) = 0$ and $\Lambda'(\theta) < \infty$ fail, revisiting the above arguments yields $e^{\theta w}\mathcal{P}_\pi(W_k \geq w) \to 0$.

Now consider the heterogeneous multiserver – batch arrival model. $\{A_k\}$ will denote the interarrival time sequence. The kth batch of jobs arrives at time $T_k = \sum_{\kappa=1}^{k} A_\kappa$. N_k is the number of jobs in the kth batch. There are c servers, which we index $i = 1,..,c$. $B_{i,j}$ is the service time of the jth job processed by server i. We assume that the $c+2$ sequences $\{A_k\}$, $\{N_k\}$, and $\{B_{i,j}\colon j = 1,2,..\}$, $i = 1,..,c$, are independent and i.i.d. Thus, the model is specified by the interarrival time distribution $F_A(\cdot)$, the batch size distribution $F_N(\cdot)$, and the c service time distributions $F_i(\cdot)$, $i = 1,..,c$. The corresponding cumulant functions are denoted $\Lambda_A(\cdot)$, $\Lambda_N(\cdot)$ and $\Lambda_i(\cdot)$. A, N and B_i will denote generic random variables having respective distributions $F_A(\cdot)$, $F_N(\cdot)$, and $F_i(\cdot)$. We globally assume the following conditions without reference.

> *Basic Assumptions:* $F_A((0,\infty)) = 1$ and $\mathrm{E}[A] < \infty$; $F_N(\mathbf{Z}^+) = 1$ and $\mathrm{E}[N] < \infty$; and $F_i([0,\infty)) = 1$ and $0 < \mathrm{E}[B_i] < \infty$ for each $i = 1,..,c$.

Note that we may have $F_i(\{0\}) > 0$, which simply means that some jobs require no processing. We define the *arrival rate* $\lambda \stackrel{\text{def}}{=} \mathrm{E}[N]/\mathrm{E}[A]$ and the *service rate* $\mu \stackrel{\text{def}}{=} \sum_{i=1}^{c} \mu_i$ where $\mu_i \stackrel{\text{def}}{=} \mathrm{E}[B_i]^{-1}$.

In most (but not all) of our analysis we will assume that the $F_i(\cdot)$s are *spread out*, that is, some finite convolution power has a nontrivial component that is absolutely continuous with respect to Lebesgue measure. This mild nonsingularity condition is used primarily for the renewal theory. It will be evident that analogous results will also hold for systems having interarrival time and service times restricted to a common lattice, but we will not consider the arithmetic case here.

Our analysis is a generalization of the classical GI/G/1 analysis outlined above. We construct a function $\Lambda(\cdot)$ from $\Lambda_A(\cdot)$, $\Lambda_N(\cdot)$ and the $\Lambda_i(\cdot)$s, and we fully characterize $\Lambda(\cdot)$ as a convex, l.s.c., piecewise analytic function with $\Lambda(0) = 0$ and $\Lambda'(0) = (\mathrm{E}[A]/\mu)(\lambda - \mu)$. The construction also identifies the $c+2$ twisting parameters that determine a θ-conjugate queuing process distribution. Tight limits are obtained by application of Markov additive process renewal theory to the θ-conjugate process distribution. The main difference is that we do not employ a Lindley identity. Instead our results are based on the large deviations behavior of a single busy cycle.

In Part I we obtain the following results. Assuming bounded service times, Theorem 2.1 states that $\theta > 0$ implies *geometric* Harris recurrence. The $\theta = \infty$ case is eliminated by Theorem 2.2, which states that waiting time and queue length processes are actually bounded in this case. The proofs of Theorems 2.1 and 2.2 are both based on construction of Foster-Lyapunov drift conditions that depend on the $\Lambda(\cdot)$ function construction. We thus establish a connection between the Foster-Lyapunov theory and quantities usually associated with large deviations analysis. The final result, Theorem 2.3, establishes the tight limits (2) under a set of sufficient conditions that naturally generalize Feller's conditions $0 < \theta < \infty$, $\Lambda(\theta) = 0$ and $\Lambda'(\theta) < \infty$. Most of these conditions are also shown to be necessary as well as sufficient.

Part II completes the analysis. We give a new proof of irreducibility based on spread out $F_i(\cdot)$s, and we then give a "state dependent" drift criteria proof of positive Harris recurrence. Finally, we complete the stationary distribution characterization. The *logarithmic limits* $\frac{1}{w}\log(\mathcal{P}_\pi(W_k \geq w)) \to -\theta$ and $\frac{1}{n}\log(\mathcal{P}_\pi(Q_k \geq n)) \to \log(\omega)$ are shown to hold assuming no additional hypothesis.

In the process of proving the above results, we will also establish tight and logarithmic limits for the stationary busy period maximum waiting time $\overline{W}$ and maximum queue length $\overline{Q}$, thus generalizing the GI/G/1 results of Iglehart (1989). These limits are useful in the asymptotic analysis of $W_k^{max} = \max_{\kappa \leq k} W_\kappa$ and $Q_k^{max} = \max_{\kappa \leq k} Q_\kappa$. See Sadowsky and Szpankowski (1992).

The tight limits (2) have previously been obtained for the GI/PH/c queue (with single job arrivals) by Takahashi (1981), and Neuts and Takahashi (1981). PH stands for "phase type" distribution (i.e., a class of continuous distributions having rational moment generating functions). The "method-of-phases" is an "analytic" approach based on the matrix geometric form of the stationary queue length distribution and identification of ω as the associated Perron-Frobenious eigenvalue. It has been suggested that the general result might be follow from the PH result and a certain continuity property of queuing process distributions, although this argument would involve an interchange of limits. The results of Part II, however, indicate that this interchange of limits would actually yield the wrong value for ω when $\Lambda(\theta) < 0$!

In comparison, the method-of-phases is an "analytic" approach that is distinctly different approach than the "probabilistic" methods (change in distribution, renewal theory, etc.) used here. Our results are more general (spread out vs. PH service times, weak sufficient conditions, etc.). We suggest that there is also a practical advantage. The constants c_W and c_Q are quite difficult to evaluate, either by our methods or by the method-of-phases. However, the conjugate queuing process distribution developed here can be used in an efficient Monte Carlo algorithm for directly estimating $\mathcal{P}_\pi(W_k \geq w)$ and $\mathcal{P}_\pi(Q_k \geq n)$. See Bucklew et. al. (1990) and Sadowsky (1991, 1993).

Finally, we note that the results are quite different when $c = \infty$. See Aldous, et. al. (1992).

Part I is organized as follows. Section 2 presents the construction of the function $\Lambda(\cdot)$ and then states the main results as outlined above. Section 3 investigates the Markov structure of the queuing process and proves Theorems 2.1 and 2.2. Section 4 proves the tight waiting time limit and Section 5 presents the queue length arguments that together prove Theorem 2.3. The required Markov additive process renewal theory is developed in the appendices.

2. Preliminaries and Presentation of Results

The purpose of this section is to formally state our main results. This is done in Subsection 2.3. We must first construct and characterize the function $\Lambda(\eta)$ and associated quantities (Subsection 2.1), and formally describe the workload and queue length processes (Subsection 2.2).

First set some basic notation. $\mathbf{Z}_+$ and $\mathbf{Z}^+$ will denote, respectively, the nonnegative and positive integers. For a vector $\mathbf{x} \in \mathbf{R}^c$, we write $\vee(\mathbf{x}) \stackrel{\text{def}}{=} \min_{i=1,..,c} x_i$ and $\wedge(\mathbf{x}) \stackrel{\text{def}}{=} \max_{i=1,..,c} x_i$. For $x \in \mathbf{R}$, $[x]^+ = \max\{0, x\}$. For $\boldsymbol{\alpha}$, $\boldsymbol{\beta} \in \mathbf{R}^c$, $\boldsymbol{\alpha} \cdot \boldsymbol{\beta} \stackrel{\text{def}}{=} \sum_{i=1}^c \alpha_i \beta_i$. Define $\underline{a} \stackrel{\text{def}}{=}$ ess inf A ($= \inf\{a$: $F_A((-\infty, a]) > 0\}$), $\overline{a} \stackrel{\text{def}}{=}$ ess sup A, $\underline{n} \stackrel{\text{def}}{=}$ ess inf N, $\overline{n} \stackrel{\text{def}}{=}$ ess sup N, $\underline{b}_i \stackrel{\text{def}}{=}$ ess inf B_i and $\overline{b}_i \stackrel{\text{def}}{=}$ ess sup B_i. Also define $\overline{b} \stackrel{\text{def}}{=} \wedge(\overline{b}_1, .., \overline{b}_c)$ and $\underline{b} \stackrel{\text{def}}{=} \vee(\underline{b}_1, .., \underline{b}_c)$.

2.1: Construction of the Function $\Lambda(\eta)$ and the Twisting Parameters

For a function f: $(-\infty, \infty] \to (-\infty, \infty]$, put $f^{-1}(y) \stackrel{\text{def}}{=} \sup\{x\colon f(x) < y\}$ for each $y \in (-\infty, \infty]$ (where $\sup \emptyset = -\infty$). We will say that $f(\cdot)$ is a *nice increasing function* if $f(x) \downarrow -\infty$ as $x \downarrow -\infty$, $f(x) \uparrow +\infty$ as $x \uparrow +\infty$, and $f(\cdot)$ is continuous and strictly increasing on $(-\infty, \overline{x}]$ where $\overline{x} = \sup\{x$:

$f(x) < \infty\}$. If $f(\cdot)$ is a nice increasing function, then $f^{-1}(\cdot) > -\infty$, $f^{-1}(y) \downarrow -\infty$ as $y \downarrow -\infty$, $f^{-1}(\cdot)$ is strictly increasing on $(-\infty, \overline{f}]$ where $\overline{f} = f(\overline{x})$, $f^{-1}(f(x)) = x$ for all $x \leq \overline{x}$, and $f^{-1}(y) = \overline{x}$ for all $y \geq \overline{f}$. Moreover, if g: $(-\infty, \infty] \to (-\infty, \infty]$ satisfies the properties of an inverse of a nice increasing function (as just listed), then $g^{-1}(\cdot)$ is a nice increasing function.

Extend the domain of the service time cumulants by putting $\Lambda_i(\infty) = \infty$. Then each $\Lambda_i(\cdot)$ is a convex nice increasing function, and is analytic on $(-\infty, \overline{\alpha}_i)$ where $\overline{\alpha}_i \stackrel{\text{def}}{=} \sup\{\alpha\text{: } \Lambda_i(\alpha) < \infty\}$. Put $\overline{\Lambda}_i \stackrel{\text{def}}{=} \Lambda_i(\overline{\alpha}_i)$. We may assume, with no loss of generality, that the servers are indexed so that $\overline{\Lambda}_c \leq \overline{\Lambda}_{c-1} \leq \cdots \leq \overline{\Lambda}_1$.

For each $\eta \in (-\infty, \infty]$, define

$$\Lambda_B(\eta) \quad \stackrel{\text{def}}{=} \quad \inf_{\boldsymbol{\alpha} \in \mathcal{H}(\eta)} \max_{i=1,..,c} \Lambda_i(\alpha_i) \tag{3}$$

where $\mathcal{H}(\eta) \stackrel{\text{def}}{=} \{\boldsymbol{\alpha} \in (-\infty, \infty]^c\text{: } \sum_{i=1}^c \alpha_i = \eta\}$. Also define $\overline{\eta}_B \stackrel{\text{def}}{=} \sup\{\eta\text{: } \Lambda_B(\eta) < \infty\}$, $\overline{\Lambda}_B \stackrel{\text{def}}{=} \Lambda_B(\overline{\eta}_B)$,

$$\alpha_i(\cdot) \quad \stackrel{\text{def}}{=} \quad \Lambda_i^{-1}(\Lambda_B(\cdot)) \tag{4}$$

and $\overline{\eta}_i \stackrel{\text{def}}{=} \Lambda_B^{-1}(\overline{\Lambda}_i)$ for each $i = 1, .., c$. As a vector write $\boldsymbol{\alpha}(\eta) = (\alpha_1(\eta), .., \alpha_c(\eta))$.

Lemma 2.1: Assume $F_i([0, \infty)) = 1$, $F_i(\{0\}) < 1$ and $\overline{\Lambda}_c \leq \cdots \leq \overline{\Lambda}_1$. Then

(i) $\Lambda_B(\cdot)$ is a convex nice increasing function, $\Lambda_B(0) = 0$, $\overline{\eta}_B = \sum_{i=1}^c \overline{\alpha}_i$, and $\overline{\Lambda}_B = \wedge(\overline{\Lambda}_1, .., \overline{\Lambda}_c)$ $(= \overline{\Lambda}_1)$.

(ii) $0 \leq \overline{\eta}_c \leq \cdots \leq \overline{\eta}_1 = \overline{\eta}_B$, each $\alpha_i(\cdot)$ is strictly increasing and continuous on $(-\infty, \overline{\eta}_i]$, $\alpha_i(\eta) = \overline{\alpha}_i$ for all $\eta \geq \overline{\eta}_i$, and $\boldsymbol{\alpha}(0) = \mathbf{0}$. Moreover, for each η such that $\Lambda_B(\eta) < \infty$ we have $\boldsymbol{\alpha}(\eta) \in \mathcal{H}(\eta)$, $\boldsymbol{\alpha}(\eta)$ is the unique solution the minimax problem of definition (3), and

$$\Lambda_i(\alpha_i(\eta)) \quad = \quad \begin{cases} \Lambda_B(\eta) & \text{for } \eta \leq \overline{\eta}_i \\ \overline{\Lambda}_i < \Lambda_B(\eta) & \text{for } \overline{\eta}_i < \eta \leq \overline{\eta}_B \end{cases} . \tag{5}$$

(iii) For each $i = 1, .., c$ such that $\overline{\eta}_{i+1} < \overline{\eta}_i$, where $\overline{\eta}_{c+1} = -\infty$ for the case $i = c$, $\Lambda_B(\cdot)$ and each $\alpha_\iota(\cdot)$ with $\iota \leq i$ are analytic on the interval $(\overline{\eta}_{i+1}, \overline{\eta}_i)$. Moreover, $\Lambda_B'(\eta) = 1/\widehat{\mu}(\eta)$ where

$$\widehat{\mu}(\eta) \quad \stackrel{\text{def}}{=} \quad \sum_{\iota=1}^{i} \Lambda_\iota'(\alpha_\iota(\eta))^{-1} \qquad \text{for } \eta \in (\overline{\eta}_{i+1}, \overline{\eta}_i). \tag{6}$$

Proof: Put $\Lambda_B^*(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \wedge(\Lambda_1(\alpha_1),..,\Lambda_c(\alpha_c))$ and $\mathcal{H}^*(\eta) \stackrel{\text{def}}{=} \mathcal{H}(\eta) \cap (-\infty,\overline{\alpha}_1] \times \cdots \times (-\infty,\overline{\alpha}_c]$. Then Λ_B^*: $(-\infty,\infty]^c \to (-\infty,\infty]$ is convex and l.s.c., and definition (3) is equivalent to $\Lambda_B(\eta) = \inf_{\mathcal{H}^*(\eta)} \Lambda_B^*(\boldsymbol{\alpha})$ where $\inf \emptyset = \infty$. For $\eta > \sum_{i=1}^c \overline{\alpha}_i$, we have $\mathcal{H}^*(\eta) = \emptyset$ and $\Lambda_B(\eta) = \infty$. For $\eta < \sum_{i=1}^c \overline{\alpha}_i$, there exists at least one $\boldsymbol{\alpha} \in \mathcal{H}^*(\eta)$ such that $\alpha_i < \overline{\alpha}_i$ for each i, and hence, $\Lambda_B(\eta) < \infty$. Thus, $\overline{\eta}_B = \sum_{i=1}^c \overline{\alpha}_i$. We also have $\mathcal{H}^*(\overline{\eta}_B) = \{(\overline{\alpha}_1,..,\overline{\alpha}_c)\}$, and hence, $\overline{\Lambda}_B \stackrel{\text{def}}{=} \Lambda_B(\overline{\eta}_B) = \wedge(\overline{\Lambda}_1,..,\overline{\Lambda}_c)$.

Put $\widetilde{\eta}(\cdot) \stackrel{\text{def}}{=} \sum_{i=1}^c \Lambda_i^{-1}(\cdot)$. $\overline{\Lambda}_i < \infty$ implies $\overline{\alpha}_i < \infty$, $\Lambda_i^{-1}(\cdot)$ is continuous and strictly increasing on $(-\infty,\overline{\Lambda}_i]$, and $\Lambda_i^{-1}(\Lambda) = \overline{\alpha}_i$ for $\Lambda \geq \overline{\Lambda}_i$. Thus, it follows that $\widetilde{\eta}(\cdot)$ is continuous and strictly increasing on $(-\infty,\overline{\Lambda}_B]$ (because $\overline{\Lambda}_B = \wedge(\overline{\Lambda}_1,..,\overline{\Lambda}_c)$), and $\widetilde{\eta}(\Lambda) = \overline{\eta}_B$ for all $\Lambda \geq \overline{\Lambda}_B$ (because $\overline{\eta}_B = \sum_{i=1}^c \overline{\alpha}_i$). So, $\widetilde{\eta}_i(\cdot)$ has all of the properties of a nice increasing function inverse, and hence, $\widetilde{\eta}^{-1}(\cdot)$ is a nice increasing function with $\sup\{\eta\text{: } \widetilde{\eta}^{-1}(\eta) < \infty\} = \overline{\eta}_B$. We claim that $\Lambda_B(\cdot) = \widetilde{\eta}^{-1}(\cdot)$. If this claim holds, then the properties of $\Lambda_B(\cdot)$ and $\boldsymbol{\alpha}(\cdot)$ listed in parts (i) and (ii) follow immediately. We have already shown that $\Lambda_B(\eta) = \widetilde{\eta}^{-1}(\eta) = \infty$ for all $\eta > \overline{\eta}_B$, and that $\Lambda_B(\overline{\eta}_B) = \widetilde{\eta}^{-1}(\overline{\eta}_B) = \overline{\Lambda}_B$. Fix any $\eta < \overline{\eta}_B$ and put $\Lambda = \widetilde{\eta}^{-1}(\eta) < \overline{\Lambda}_B$ and $\alpha_i^* = \Lambda_i^{-1}(\Lambda)$. By the definition of $\widetilde{\eta}(\cdot)$, $\boldsymbol{\alpha}^* \in \mathcal{H}^*(\eta)$, $\Lambda_i(\alpha_i^*) \leq \Lambda$ with equality whenever $\alpha_i^* < \overline{\alpha}_i$, and $\alpha_i^* < \overline{\alpha}_i$ for at least one i because $\eta < \overline{\eta}_B = \sum_{i=1}^c \overline{\alpha}_i$. Thus, $\Lambda_B^*(\boldsymbol{\alpha}^*) = \Lambda \geq \Lambda_B(\eta)$, so it is enough to show that $\Lambda_B^*(\boldsymbol{\beta}) > \Lambda$ for any $\boldsymbol{\beta} \in \mathcal{H}(\eta)$ such that $\boldsymbol{\beta} \neq \boldsymbol{\alpha}^*$. For at least one i we must have $\beta_i > \alpha_i^*$. If $\alpha_i^* = \overline{\alpha}_i$, then $\Lambda_B^*(\boldsymbol{\beta}) \geq \Lambda_i(\beta_i) = \infty > \Lambda$. If $\alpha_i^* < \overline{\alpha}_i$, then since $\Lambda_i(\cdot)$ is strictly increasing on $(-\infty,\overline{\alpha}_i]$, we have $\Lambda_B^*(\boldsymbol{\beta}) \geq \Lambda_i(\beta_i) > \Lambda_i(\alpha_i^*) = \Lambda$. Thus, $\Lambda_B(\eta) = \Lambda$ as claimed, and moreover, $\alpha_i(\eta) = \Lambda_i^{-1}(\Lambda_B(\eta)) = \Lambda_i^{-1}(\Lambda) = \alpha_i^*$, and hence, we have also shown that $\boldsymbol{\alpha}(\eta)$ is the unique solution to the minimax problem in definition (3). That $\boldsymbol{\alpha}(0) = \mathbf{0}$ and $\Lambda_B(0) = 0$ follows from $\Lambda_i(0) = 0$ for each i. Fix $\eta_1, \eta_2 \in (-\infty,\overline{\eta}_B]$ and $\sigma \in (0,1)$. By the convexity of $\Lambda_B^*(\cdot)$ we have $\Lambda_B(\sigma\eta_1 + (1-\sigma)\eta_2) \leq \Lambda_B^*(\sigma\boldsymbol{\alpha}(\eta_1) + (1-\sigma)\boldsymbol{\alpha}(\eta_2)) \leq \sigma\,\Lambda_B(\eta_1) + (1-\sigma)\,\Lambda_B(\eta_2)$. Thus, $\Lambda_B(\cdot)$ is also convex. At this point, we have all of parts (i) and (ii).

$\overline{\Lambda}_c \leq \cdots \leq \overline{\Lambda}_1$ implies $\overline{\eta}_c \leq \cdots \leq \overline{\eta}_1$. Suppose $\overline{\eta}_{i+1} < \overline{\eta}_i$ for some i. For $\eta \in (\overline{\eta}_{i+1},\overline{\eta}_i)$ we have

$$\eta \;=\; \sum_{\iota=1}^{c} \alpha_\iota(\eta) \;=\; \sum_{\iota=1}^{i} \Lambda_\iota^{-1}(\Lambda_B(\eta)) \;+\; \sum_{\iota=i+1}^{c} \overline{\alpha}_\iota$$

and the inverses in the first sum on the right above are analytic and strictly increasing on $(\overline{\eta}_{i+1},\overline{\eta}_i)$. Thus, $\Lambda_B(\cdot)$ is analytic on $(\overline{\eta}_{i+1},\overline{\eta}_i)$ by the analytic implicit function theorem. $\Lambda_B'(\eta) = \widehat{\mu}(\eta)^{-1}$ follows by differentiating the above display. This proves (iii). ∎

For each $\eta \in \mathbf{R}$, define

$$\Lambda(\eta) \quad \stackrel{\text{def}}{=} \quad \Lambda_A(-\eta) \quad + \quad \Lambda_N\Big(\Lambda_B(\eta)\Big), \tag{7}$$

and put $\mathcal{D} \stackrel{\text{def}}{=} \{\eta\colon\ \Lambda(\eta) < \infty\}$. By Lemma 2.1, $\Lambda(\cdot)$ is a convex and l.s.c. mapping $\mathbf{R} \to (-\infty, \infty]$, and $\Lambda(0) = 0$. Moreover, $\Lambda(\cdot)$ is analytic on each interval $(\overline{\eta}_{i+1}, \overline{\eta}_i) \cap \mathcal{D}^o$ with derivative

$$\Lambda'(\eta) \quad = \quad \frac{\Lambda_A'(-\eta)}{\widehat{\mu}(\eta)} \Big(\lambda(\eta) \; - \; \widehat{\mu}(\eta) \Big) \tag{8}$$

where $\widehat{\mu}(\eta)$ is as defined in (6) and

$$\lambda(\eta) \quad \stackrel{\text{def}}{=} \quad \frac{\Lambda_N'(\Lambda_B(\eta))}{\Lambda_A'(-\eta)}. \tag{9}$$

In Section 4 we will see that $\lambda(\eta)$ and $\mu(\eta) \stackrel{\text{def}}{=} \sum_{i=1}^{c} \Lambda_i'(\alpha_i(\eta))^{-1}$ are, respectively, the arrival and service rates for the η-conjugate queuing process distribution. In particular, $\lambda(0) = \mathrm{E}[N]/\mathrm{E}[A]$ $= \lambda =$ the arrival rate, and $\mu(0) = \sum_{i=1}^{c} \mathrm{E}[B_i]^{-1} = \mu =$ the service rate. As in the GI/G/1 case, we would like to say that the sign of $\Lambda'(\eta)$ determines η-conjugate stability condition $\lambda(\eta) <$ or $>$ $\mu(\eta)$. But this is not quite true in general. It is $\widehat{\mu}(\eta)$, not $\mu(\eta)$, that appears in the differentiation formula (8), and we may have $\widehat{\mu}(\eta) < \mu(\eta)$ when $\eta > \overline{\eta}_c$ due to the missing terms in (6). However, if $\eta \le \overline{\eta}_c$, then $\widehat{\mu}(\eta) = \mu(\eta)$, and hence, the sign of $\Lambda'(\eta)$ does determine the η-conjugate stability in this case. Notice that this is an issue only for heterogeneous systems. In the homogeneous case, $\overline{\eta}_i = \overline{\eta}_B$ for each $i = 1, .., c$.

2.2: The Workload and Queue Length Markov Chains

If no new jobs are added to the queue after time t, let $t + W_i(t)$ denote the time instant at which server i becomes idle. $W_i(t)$ is called the ith server's workload at time t, and $\mathbf{W}(t) = (W_1(t), .., W_c(t))$ is the *workload process*. The *waiting time* of a single job added to the system at time t is $W(t) = \vee(\mathbf{W}(t))$. At time $T_k = \sum_{\kappa=1}^{k} A_\kappa$ a batch of N_k jobs arrives at the queue. We write $\mathbf{W}_k^- = \mathbf{W}(T_k^-)$ and $\mathbf{W}_k^+ = \mathbf{W}(T_k^+)$ to denote the workload vector just before and just after the kth batch is added to the system.

Let $Q(t)$ denote the number of jobs in queue at time t, not counting the jobs currently being processed. Let $R_i(t)$ denote the service time remaining for the job being processed by server i at time t. $\mathbf{R}(t) = (R_1(t), .., R_c(t))$ is the *residual service time vector*, and we call $(\mathbf{R}(t), Q(t))$ the *queue length process*. As above, write $Q_k^\pm = Q(T^\pm)$ and $\mathbf{R}_k^\pm = \mathbf{R}(T^\pm)$.

We will initialize the queuing process by taking the 0th batch of jobs to be inserted into the system at time $t = 0$. $\mathcal{P}_{\mathbf{w}}(\cdot)$ ($\mathcal{P}_\zeta(\cdot)$) will denote the workload process distribution initialized with $\mathbf{W}_0^- = \mathbf{w}$ ($\mathcal{P}_\zeta(\mathbf{W}_0^- \in \cdot) = \zeta(\cdot)$). Similar notation is used for the queue length process.

It turns out that $\{\mathbf{W}_k^\pm\}$ and $\{(\mathbf{R}_k^\pm, Q_k^\pm)\}$ are Markov chains on the respective state spaces $[0,\infty)^c$ and $[0,\infty)^c \times \mathbf{Z}_+$. This is well known, but we will examine the Markov structure in detail in Section 3.

We require some definitions from Markov chain theory. See Nummelin (1984), or Meyn and Tweedie (1992,1993). Let $\{X_k\}$ be a Markov chain taking values in a locally compact topological space $\mathbf{S}$ with Borel field $\mathcal{S}$. For a measure $\phi(\cdot)$ define $\mathcal{S}_\phi \stackrel{\text{def}}{=} \{B \in \mathcal{S}\colon\ \phi(B) > 0\}$. $\{X_k\}$ is said to be *irreducible* if there exists a positive measure $\phi(\cdot)$, called an *irreducibility measure*, such that $\mathcal{P}_x(X_k \in B) > 0$ for some $k = k(x,B) < \infty$ for all $x \in \mathbf{S}$ and $B \in \mathcal{S}_\phi$. An irreducible chain is *Harris recurrent* if for all $x \in \mathbf{S}$ and $B \in \mathcal{S}_\phi$ we have $\mathcal{P}_x(X_k \in B \text{ i.o.}) = 1$. Harris recurrent chains have a unique (up to a scale factor) σ-finite invariant measure $\pi(\cdot)$. Moreover, $\pi(\cdot)$ is a "maximal" irreducibility measure in the sense that $\phi \ll \pi$ for any other irreducibility measure $\phi(\cdot)$. If $\pi(\cdot)$ is finite, then the chain is *positive Harris recurrent*. In this case we always take $\pi(\cdot)$ to be normalized to a probability distribution and we call $\pi(\cdot)$ the *stationary distribution*.

A set $C \in \mathcal{S}$ is a *petite set* if for some $\epsilon \in (0,1)$ there exists a nontrivial substochastic measure $\nu(\cdot)$ such that $\sum_{k=1}^\infty \epsilon^k\, \mathcal{P}_x(X_k \in \cdot) \geq \nu(\cdot)$ for all $x \in C$. If for some $k_0 < \infty$ we have $\mathcal{P}_x(X_{k_0} \in \cdot)$ $\geq \nu(\cdot)$ for all $x \in C$, then C is a *small set*. Clearly, small sets are petite. Also, these notions are equivalent when $\{X_k\}$ is aperiodic. See Meyn and Tweedie (1993). As will be noted in Part II, our chains are generally not aperiodic.

If $\{X_k\}$ is Harris recurrent and there exists a petite set $C \in \mathcal{S}_\pi$ such that $\sup_{x\in C} \mathrm{E}_x[K_C] < \infty$ where $K_C = \inf\{k \geq 1\colon\ X_k \in C\}$, then $\{X_k\}$ is positive Harris recurrent. If the stronger condition $\sup_{x\in C} \mathrm{E}_x[r^{K_C}] < \infty$ for some $r > 1$ holds, then $\{X_k\}$ is said to be *geometrically* Harris recurrent. This implies geometric convergence to stationary behavior. We again refer the reader to the references for further exposition.

Hereafter, we will use $\pi(\cdot)$ to denote the stationary distribution of $\{\mathbf{W}_k^-\}$, and $\widehat{\pi}(\cdot)$ will denote the stationary distribution of $\{(\mathbf{R}_k^-, Q_k^-)\}$, whenever these chains are positive Harris recurrent.

Perhaps the most common way to show that a queuing process is irreducible is by *idle period regeneration*. The following lemma presents a simple sufficient condition.

Lemma 2.2: If $\overline{a} > \underline{n} \wedge(\underline{b}_1, .., \underline{b}_c)$, then $\{\mathbf{W}_k^-\}$ is irreducible with irreducibility measure $\delta_{\mathbf{0}}(\cdot)$, and all bounded sets are small sets.

Sketch of Proof: Consider only the case $\overline{a} = \infty$. Let C be any bounded set, and put $b^* = \wedge(\underline{b}_1, .., \underline{b}_c)$ and $v = \sup_{\mathbf{w} \in C} \wedge(\mathbf{w}) + \underline{n}\, b^*$. Then $\mathcal{P}_{\mathbf{w}}(\mathbf{W}_1^- \in \cdot) \geq \gamma\, \delta_{\mathbf{0}}(\cdot)$ for all $\mathbf{w} \in C$ where $\gamma = F_A((v + \underline{n}\epsilon, \infty))\, F_N(\{\underline{n}\})\, [\min_i F_i([0, b^* + \epsilon])\,]^{\underline{n}} > 0$. The case $\overline{a} < \infty$ requires more than one step to drive $\mathbf{W}_k^-$ to $\mathbf{0}$ when v is large. ∎

For the single server queue, $\lambda < \mu$ and the independence of A, B and N implies $\mathrm{E}[NB - A] < 0$, which in turn implies ess inf $\{NB - A\} = \underline{n}\,\underline{b} - \overline{a} < 0$. Thus, for $c = 1$ only, $\lambda < \mu$ actually implies irreducibility. For $c > 1$, irreducibility is a somewhat more difficult issue. When $\overline{a} < \underline{b}$ we may still have $\lambda < \mu$, but full idle period regenerations never occur.

In Part II, we will give a new proof of irreducibility for the case that the service time distributions are spread out, and it is also shown that compact sets are petite for that case. Fortunately, between Lemma 2.2 and the results of Part II most systems of practical interest are covered. Outside of the scope of these results it appears to be difficult prove irreducibility and identify the petite sets.

2.3: Statement of Results

Our first result gives a sufficient condition for geometric Harris recurrence. Recall that $\theta \stackrel{\text{def}}{=} \sup\{\eta\colon\, \Lambda(\eta) \leq 0\}$. As discussed above, the irreducibility and petite set hypothesis may be verified using either Lemma 2.2 or results from Part II.

Theorem 2.1: Assume that $\{\mathbf{W}_k^-\}$ is irreducible, all compact sets are petite, $\overline{b} < \infty$ and $\theta > 0$. Then both $\{\mathbf{W}_k^-\}$ and $\{(\mathbf{R}_k^-, Q_k^-)\}$ are geometrically Harris recurrent Markov chains.

We next characterize the $\theta = \infty$ case. To see what to expect, consider the GI/G/1 case. When $\theta = \infty$ we have $\Lambda(\eta) = \log(\,\mathrm{E}[\,\exp(\eta[B - A])\,]\,) \leq 0$ for all $\eta > 0$. This implies $B - A \leq 0$ a.s., and since B and A are independent, ess sup $[B - A] = \overline{b} - \underline{a} \leq 0$. Thus, the queue is either a deterministic $\lambda = \mu$ system (in which case $\Lambda(\cdot) \equiv 0$), or the system eventually empties and $W_k^- \equiv 0$ after that point in time.

We say that "the queue is bounded" if there is a constant $\overline{w} < \infty$ and a stopping time K such that $\mathcal{P}_{\mathbf{w}}(K < \infty) = 1$ for all $\mathbf{w} \in [0,\infty)^c$, and $W_k^+ \leq \overline{w}$ for all $k \geq K$. In other words, $\{\mathbf{w} \in [0,\infty)^c\colon \vee(\mathbf{w}) \leq \overline{w}\}$ is an absorbing set. (Notice that this does not imply that each components $W_{i,k}^+$ is bounded.) Likewise, if the queue has a bounded waiting time as above, then there also exists a constant $\overline{q} < \infty$ such that $Q_k^+ \leq \overline{q}$ for all $k \geq K$.

Theorem 2.2: Assume $\lambda < \mu$. Then the following are equivalent:

(i) $\theta = \infty$.

(ii) There exists a vector $\mathbf{e}^* \in [0,1]^c$ such that $\sum_{i=1}^c e_i^* = 1$ and $e_i^* \,\overline{b}_i\, \overline{n} \;\leq\; \underline{a}$ for each $i = 1,..,c$.

(iii) The queue is bounded.

We remark that the hypothesis $\lambda < \mu$ in Theorem 2.2 is needed only to eliminate the deterministic $\lambda = \mu$ system, for which $\Lambda(\cdot) \equiv 0$.

We require a few more definitions. Define $K_1 \stackrel{\text{def}}{=} \inf\{k \geq 1\colon\; W_k^- = 0\}$, $\overline{W} \stackrel{\text{def}}{=} \max\{W_k^+\colon 0 \leq k \leq K_1\}$ = the maximum waiting time during the first busy period, and $\overline{Q} \stackrel{\text{def}}{=} \max\{Q_k^+\colon k < K_1\}$ = the maximum queue length during the first busy period. A distribution $\zeta(\cdot)$ on $[0,\infty)^c$ is an *admissible* initial distribution for $\mathbf{W}_0^-$ if $\mathrm{E}_\zeta[\,\exp(\boldsymbol{\alpha}(\theta) \cdot \mathbf{W}_0^-)\,] < \infty$ and $\mathcal{P}_\zeta(\overline{W} \geq w) > 0$ for all $w < \infty$. Likewise, $\widehat{\zeta}(\cdot)$ on $[0,\infty)^c \times \mathbf{Z}_+$ is an *admissible* initial distribution for $(\mathbf{R}_0^-, Q_0^-)$ if $\mathrm{E}_{\widehat{\zeta}}\Big[\exp\Big(\boldsymbol{\alpha}(\theta) \cdot \mathbf{R}_0^- \;+\; \Lambda_B(\theta)\, Q_0^-\,\Big)\Big] < \infty$ and $\mathcal{P}_{\widehat{\zeta}}(\overline{Q} \geq n) > 0$ for all $n < \infty$. Obviously, any distribution with compact support satisfies the exponential moment condition. Thus, by Theorem 2.2 admissible initial distributions exist if and only if $\theta < \infty$. Finally, define $\omega \stackrel{\text{def}}{=} \exp(\,-\Lambda_B(\theta)\,)$.

Theorem 2.3: Assume that $\{\mathbf{W}_k^-\}$ and $\{(\mathbf{R}_k^-, Q_k^-)\}$ are positive Harris recurrent, and that the $F_i(\cdot)$s are spread out. Let $\zeta(\cdot)$ and $\widehat{\zeta}(\cdot)$ be any admissible initial distributions.

(i) (Sufficient Conditions) Assume $0 < \theta < \infty$, $\theta \leq \overline{\eta}_c$, $\Lambda(\theta) = 0$, $\Lambda_i'(\alpha_i(\theta)) < \infty$ for each i, and $\Lambda_N'(\Lambda_B(\theta)) < \infty$. Then there exists constants $0 < c_W, c_{\overline{W}}, c_Q, c_{\overline{Q}} < \infty$ such that

$$\mathcal{P}_\zeta(\,\overline{W} \geq w\,) \;\sim\; c_{\overline{W}}\, e^{-\theta w} \quad \text{and} \quad \mathcal{P}_{\widehat{\zeta}}(\,\overline{Q} \geq n\,) \;\sim\; c_{\overline{Q}}\, \omega^n, \tag{10}$$

and

$$\mathcal{P}_\pi(\,W_k^+ \geq w\,) \;\sim\; c_W\, e^{-\theta w} \quad \text{and} \quad \mathcal{P}_{\widehat{\pi}}(\,Q_k^+ \geq n\,) \;\sim\; c_Q\, \omega^n \tag{11}$$

as $w \to \infty$ and $n \to \infty$.

(ii) (Necessary Conditions) If any of the conditions $0 < \theta < \infty$, $\theta \le \overline{\eta}_c$ and $\Lambda(\theta) = 0$ fail, then

$$e^{\theta w}\,\mathcal{P}_{\zeta}(\,\overline{W} \ge w\,) \;\to\; 0 \quad \text{and} \quad \omega^{-n}\,\mathcal{P}_{\widehat{\zeta}}(\,\overline{Q} \ge n\,) \;\to\; 0, \tag{12}$$

and likewise,

$$e^{\theta w}\,\mathcal{P}_{\pi}(\,W_k^+ \ge w\,) \;\to\; 0 \quad \text{and} \quad \omega^{-n}\,\mathcal{P}_{\widehat{\pi}}(\,Q_k^+ \ge n\,) \;\to\; 0. \tag{13}$$

We believe that (12) and (13) also hold if either $\Lambda_N'(\Lambda_B(\theta)) = \infty$ or $\Lambda_i'(\alpha_i(\theta)) = \infty$ for some i. If so, then all of the conditions listed under part (i) of Theorem 2.3 are necessary as well as sufficient. The proof would require an excursion into degenerate Markov additive process renewal theory. We will discuss this only briefly at the end of Appendix B.

The above results leave us with two unanswered questions. First, can positive Harris recurrence be obtained under weaker conditions than those of Theorem 2.1? In particular, can we eliminate $\overline{b} < \infty$? Second, when the tight limits fail, do the parameters θ and ω still identify the correct logarithmic behavior, that is, do the quantities in (12) and (13) vanish less than exponentially fast? These are precisely the issues addressed in Part II.

3. The Markov Structure

Subsections 3.1 and 3.2 consider the Markov structure of $\{\mathbf{W}_k^-\}$ and $\{(\mathbf{R}_k^-, Q_k^-)\}$, respectively. Subsections 3.3 and 3.4 prove, respectively, of Theorems 2.1 and 2.2. Both proofs are based on the construction of a "drift criteria." The section is completed by a presentation of the cycle structure that forms the basis for the analysis of Sections 4 and 5.

3.1: The Workload Process

The evolution of $\mathbf{W}(t)$ is best described in terms of c fictitious single server queues, one for each server. FIFO servicing priority is equivalent to "inserting" jobs into the parallel queues system one-by-one by assigning them to the queue with the current minimum waiting time. Put $M_k \stackrel{\text{def}}{=} \sum_{\kappa=0}^{k} N_\kappa$ = the total number of jobs that arrive in batches $\kappa = 0,..,k$. Let $m = 1, 2, ...$ index the jobs in the order that they are inserted into the system, so the jobs in the kth batch are indexed $m = M_{k-1}+1,..,M_k$ (with $M_{-1} = 0$ for $k = 0$). Let $\mathbf{Y}^{\pm}(m)$ denote the workload vector just prior to and just after the insertion of the mth job, so $\mathbf{W}_k^- = \mathbf{Y}^-(M_{k-1}+1)$ and $\mathbf{W}_k^+ = \mathbf{Y}^+(M_k)$. Let

$J_i(m)$ denote the index of the last job inserted in the ith queue just after the mth job is inserted into the system. The process is initialized at time $t=0$ with workload $\mathbf{Y}^-(1) = \mathbf{W}_0^-$, and $J_i(0) = 0$ for $i = 1,..,c$. We have the following recursive set of equations. First determine $I(m)$ such that $Y^-_{I(m)} = \vee(\mathbf{Y}^-(m))$. We assume that there is a tie breaking rule that determines a unique value $I(m)$. For brevity, we consider only deterministic tie breaking rules, however, by augmentation of the sample space in the usual way it is a minor matter to allow randomized tie breaking rules. Next, we have

$$J_i(m) \quad = \quad \begin{cases} J_i(m-1)+1 & \text{for } i = I(m) \\ J_i(m-1) & \text{for } i \neq I(m) \end{cases}, \tag{14}$$

$$Y_i^+(m) \quad = \quad \begin{cases} Y_i^-(m) + B_{i,J_i(m)} & \text{for } i = I(m) \\ Y_i^-(m) & \text{for } i \neq I(m) \end{cases}, \tag{15}$$

and

$$Y_i^-(m+1) \quad = \quad \begin{cases} [Y_i^+(m) - A_{k+1}]^+ & \text{when } M_k = m \text{ for some } k \\ Y_i^+(m) & \text{otherwise} \end{cases}. \tag{16}$$

Since $\mathbf{W}_k^- = \mathbf{Y}^-(M_{k-1}+1)$ and $\mathbf{W}_k^+ = \mathbf{Y}^+(M_k)$, it follows from the above recursive structure that $\{\mathbf{W}_k^-\}$, $\{\mathbf{W}_k^+\}$ and $\{(\mathbf{W}_k^-, \mathbf{W}_k^+)\}$ are Markov chains.

In the GI/G/1 case, $W_k^- = [W_{k-1}^- + B_{k-1} - A_k]^+ =$ a reflected random walk. Much of the behavior of W_k^- can be deduced by studying the random walk $\widetilde{W}_k^- = W_0^- + \sum_{\kappa=1}^k (B_{\kappa-1} - A_\kappa)$, which is "coupled" to the GI/G/1 waiting time process during the first busy period. It is a trivial matter to construct a "tilde process" for our multiserver queuing model. Simply neglect the reflections in (16). Initialize with $\widetilde{\mathbf{Y}}^-(1) = \mathbf{W}_0^-$ and $\widetilde{J}_i(0) = 0$ for $i = 1,..,c$. For $m = 1,2,...$, recursively define $\widetilde{I}(m)$, $\widetilde{J}_i(m)$ and $\widetilde{\mathbf{Y}}^\pm(m)$ as in formulas (14) – (16) with the only modification being that $\widetilde{Y}_i^-(M_k+1) = \widetilde{Y}_i^+(M_k) - A_{k+1}$ in (16). Define $\widetilde{\mathbf{W}}_k^- \stackrel{\text{def}}{=} \widetilde{\mathbf{Y}}^-(M_{k-1}+1)$, $\widetilde{\mathbf{W}}_k^+ \stackrel{\text{def}}{=} \widetilde{\mathbf{Y}}^+(M_k)$ and $\widetilde{W}_k^\pm \stackrel{\text{def}}{=} \vee(\widetilde{\mathbf{W}}_k^\pm)$. Next define the random walks

$$\widetilde{S}_i(j) \quad \stackrel{\text{def}}{=} \quad W_{i,0}^- \; + \; \sum_{j'=1}^{j} B_{i,j'} \tag{17}$$

for $i = 1,..,c$. Then, since there are no reflections in the "tilde process," for $M_{k-1}+1 \le m \le M_k$ we have

$$\widetilde{Y}_i^-(m) \quad = \quad \widetilde{S}_i(\widetilde{J}_i(m-1)) \; - \; T_k \quad \text{and} \quad \widetilde{Y}_i^+(m) \quad = \quad \widetilde{S}_i(\widetilde{J}_i(m)) \; - \; T_k. \tag{18}$$

Thus, the index $\widetilde{I}(m)$ is determined by $\widetilde{S}_{\widetilde{I}(m)}(\widetilde{J}_{\widetilde{I}(m)}(m-1)) = \min_{i=1,..,c} \widetilde{S}_i(\widetilde{J}_i(m-1))$, or equivalently,

$$\widetilde{S}_{\widetilde{I}(1)}(\widetilde{J}_{\widetilde{I}(1)}(0)) \quad \le \quad \widetilde{S}_{\widetilde{I}(2)}(\widetilde{J}_{\widetilde{I}(2)}(1)) \quad \le \quad \widetilde{S}_{\widetilde{I}(3)}(\widetilde{J}_{\widetilde{I}(3)}(2)) \quad \le \quad \cdots$$

with each $\widetilde{S}_i(j)$ occurring exactly once. (The tie breaking rule determines the ordering when there are equalities.) So we see that $\widetilde{I}(m)$ and the $\widetilde{J}_i(m)$s are completely determined by the random walks $\widetilde{S}_i(j)$, $i = 1,..,c$. The unreflected processes $\{\widetilde{\mathbf{W}}_k^{\pm}\}$ are thus completely determined by the $c+2$ random walks T_k, M_k and $\widetilde{S}_i(j)$, $i = 1,..,c$.

Lemma 3.1: We have $\widetilde{W}_k^{\pm} \le W_k^{\pm}$ for all $k \in \mathbf{Z}_+$. On the event $\{W_0^+ > 0\}$ we have $\widetilde{\mathbf{W}}_k^{\pm} = \mathbf{W}_k^{\pm}$ for all $k < K_1 \stackrel{\text{def}}{=} \inf\{k \ge 1\colon W_k^- = 0\} = \inf\{k \ge 1\colon \widetilde{W}_k^- \le 0\}$.

Proof: The fact that $\widetilde{\mathbf{W}}_k^{\pm} = \mathbf{W}_k^{\pm}$ for all $k < K_1$ is obvious. When $I(m) = i$, $J_i(m) = j$ and $M_{k-1} < m \le M_k$, put $U_{i,j} \stackrel{\text{def}}{=} T_k - \sup\{t \le T_k\colon W_i(t) > 0\} \ge 0$. Put $S_i(j) \stackrel{\text{def}}{=} W_{i,0}^- + \sum_{j'=1}^{j}[B_{i,j'} + U_{i,j'}]$. In this manner we have adjusted the random walk definition (17) to account for the reflections in (16) that were neglected to define the "tilde process." Of course, $S_i(j)$ is not a random walk, but for $M_k+1 \le m \le M_{k+1}$ we have $Y_i^+(m) = S_i(J_i(m)) - T_k$, and the original queuing process satisfies the ordering $S_{I(1)}(J_{I(1)}(0)) \le S_{I(2)}(J_{I(2)}(1)) \le \ldots$ with each $S_i(j)$ occurring exactly once. Since $U_{i,j} \ge 0$, we have $\widetilde{S}_i(j) \le S_i(j)$ for all i, j, and hence, $\widetilde{S}_{\widetilde{I}(m)}(\widetilde{J}_{\widetilde{I}(m)}(m-1)) \le S_{I(m)}(J_{I(m)}(m-1))$ and it follows that $\widetilde{W}_k^{\pm} \le W_k^{\pm}$. (Notice that this proof does not imply $\widetilde{I}(m) = I(m)$, $\widetilde{J}_i(m) = J_i(m)$, or even $\widetilde{W}_{i,k}^{*\pm} \le W_{i,k}^{*\pm}$ for fixed i and $k \ge K_1$.) ■

Lemma 3.2: $k^{-1}\,\widetilde{W}_{i,k}^{\pm} \to (\lambda/\mu\ -\ 1)\,\mathrm{E}[A]$ a.s. Moreover, if $\{\mathbf{W}_k^-\}$ is Harris recurrent and $\lambda < \mu$, then $\pi(\{\mathbf{w} : \vee(\mathbf{w}) = 0\}) = \mathcal{P}_\pi(W_0^- = 0) > 0$.

Sketch of Proof: By the law of large numbers, $\widetilde{S}_i(j) \sim j/\mu_i$ a.s. We also have $\widetilde{S}_i(\widetilde{J}_i(m)) \sim \widetilde{S}_1(\widetilde{J}_1(m))$ for each i, and hence, $\widetilde{J}_i(m) \sim (\mu_i/\mu_1)\widetilde{J}_1(m)$ a.s. Since $\sum_{i=1}^{c} \widetilde{J}_i(m) = m$, we have $\widetilde{J}_1(m) \sim (\mu_1/\mu)\,m$. Likewise, $\widetilde{J}_i(m) \sim (\mu_i/\mu)\,m$ for each i, and hence, $\widetilde{S}_i(\widetilde{J}_i(m)) \sim m/\mu$. Since $M_k \sim \mathrm{E}[N]\,k$ and $T_k \sim \mathrm{E}[A]\,k$, $k^{-1}\,\widetilde{W}_{i,k}^{\pm} \to (\lambda/\mu\ -\ 1)\,\mathrm{E}[A]$ follows from (18).

If $\lambda < \mu$, then by the above result we have $\mathcal{P}_{\mathbf{w}}(W_k^- = 0 \text{ for some } k < \infty) = \mathcal{P}_{\mathbf{w}}(\widetilde{W}_k^- \le 0 \text{ for some } k < \infty) = 1$ for all $\mathbf{w} \in [0,\infty)^c$. If $\{\mathbf{W}_k^-\}$ is Harris recurrent, then this implies $\mathcal{P}_\pi(W_0^- = 0) > 0$. ■

As a trivial application of Lemmas 3.1 and 3.2, observe that the system is unstable when $\lambda > \mu$.

That is, $\mathcal{P}_{\mathbf{w}}(W_k^{\pm} \le w \text{ i.o.}) \le \mathcal{P}_{\mathbf{w}}(\widetilde{W}_k^{\pm} \le w \text{ i.o.}) = 0$ for all $w < \infty$.

3.2: The Queue Length Process

The processes $\{(\mathbf{R}_k^{\pm}, Q_k^{\pm})\}$ are also Markov chains. We will not formally specify the Markov structure of these processes, as we have done for the workload processes, but we do formulate the associated "tilde process."

First, observe that queue length and workload processes are connected. We have $W(t) = 0$ if and only if $Q(t) = 0$ and $\vee(\mathbf{R}(t)) = 0$. Define the sequence of stopping times $K_0 = 0$ and

$$K_\ell \quad = \quad \inf\{k > K_{\ell-1} : W_k^- = 0\} \quad = \quad \inf\{k > K_{\ell-1} : Q_k^- = 0 \text{ and } \vee(\mathbf{R}_k^-) = 0\} \tag{19}$$

for $\ell \in \mathbf{Z}^+$. (Note that the $\ell = 1$ case agrees with the previous definition of K_1.) Then, whenever $K_\ell < \infty$, we have $(\mathbf{R}_{K_\ell}^-, Q_{K_\ell}^-) = (\mathbf{W}_{K_\ell}^-, 0)$.

Let $\widetilde{S}_i(j)$ be the random walk as defined in (17) with $W_{i,0}^-$ replaced by $R_{i,0}^-$. (On the event $\{Q_0^- = 0\}$ we have $\mathbf{W}_0^- = \mathbf{R}_0^-$, hence, on this event there is no difference.) Define $\widehat{J}_i^+(t) = \inf\{j \ge 0\colon\ \widetilde{S}_i(j) > t\}$, and

$$\widetilde{R}_i^+(t) \quad = \quad \widetilde{S}_i(\widehat{J}_i^+(t)) \quad - \quad t. \tag{20}$$

In renewal theory, $\widetilde{R}_i^+(t)$ is called a *forward process*. As a vector, write $\widetilde{\mathbf{R}}^+(t) = (\widetilde{R}_1^+(t), .., \widetilde{R}_c^+(t))$, and put $\widetilde{\mathbf{R}}_k^+ = \widetilde{R}(T_k^+)$ and

$$\widetilde{Q}_k^+ \quad = \quad Q_0^- \quad + \quad M_k \quad - \quad \sum_{i=1}^{c} \widehat{J}_i^+(T_k). \tag{21}$$

We state the following is obvious lemma without proof.

Lemma 3.3: We have the alternative expression $K_1 = \inf\{k \ge 1\colon\ \widetilde{Q}_k^- < 0\}$, $(\mathbf{R}_k^{\pm}, Q_k^{\pm}) = (\widetilde{\mathbf{R}}_k^{\pm}, \widetilde{Q}_k^{\pm})$ for all $k < K_1$, and $\widetilde{Q}_k^+ \sim \mathrm{E}[A]\,(\lambda - \mu)\,k$ a.s.

The next lemma indicates the close coupling between the chains $\{\mathbf{W}_k^-\}$ and $\{(\mathbf{R}_k^-, Q_k^-)\}$.

Lemma 3.4: Assume $\lambda < \mu$. Then $K_\ell < \infty$ a.s. for each $\ell \in \mathbf{Z}^+$, and moreover, $\{\mathbf{W}_k^-\}$ is Harris recurrent (positive Harris recurrent) if and only if $\{(\mathbf{R}_k^-, Q_k^-)\}$ is Harris recurrent (positive Harris recurrent).

Proof: That $K_\ell < \infty$ a.s. for each $\ell \in \mathbf{Z}^+$ is obvious. Suppose $\{\mathbf{W}_k^-\}$ is Harris recurrent and put $E \stackrel{\text{def}}{=} \{\mathbf{w} \in [0,\infty)\colon \vee(\mathbf{w}) = 0\}$, so $K_1 = \inf\{k\colon\ \mathbf{W}_k^- \in E\}$. ¿From Lemma 3.2, $\lambda < \mu$ implies $\mathcal{P}_{\mathbf{w}}(K_1 < \infty) = 1$ for all $\mathbf{w} \in [0,\infty)$, which implies $\pi(E) > 0$ and $\mathcal{P}_{\mathbf{w}}(\mathbf{W}_k^- \in E \text{ i.o.}) = 1$ by Harris recurrence, and $\phi(\cdot) \stackrel{\text{def}}{=} \pi(\,\cdot \cap E)$ serves as an irreducibility measure for $\{\mathbf{W}_k^-\}$. Since $(\mathbf{R}_{K_\ell}^-, Q_{K_\ell}^-) = (\mathbf{W}_{K_\ell}^-, 0)$, $\phi{\times}\delta_0(\cdot)$ serves as an irreducibility measure for $\{(\mathbf{R}_k^-, Q_k^-)\}$. By Theorem 2.1 of Nummelin (1984) there exists a small set $C \subset E$ for $\{\mathbf{W}_k^-\}$ with $\phi(C) > 0$. Specifically, for some $k_0 \in \mathbf{Z}^+$ we have $\mathcal{P}_{\mathbf{w}}(W_{k_0}^- \in \cdot) \geq 1_C(\mathbf{w})\,\nu(\cdot)$ where $\nu(\cdot)$ is a nontrivial substochastic measure, and we may assume that $\nu(\cdot)$ is concentrated on E. Then $C{\times}\{0\}$ is a small set for $\{(\mathbf{R}_k^-, Q_k^-)\}$ with $\phi{\times}\delta_0(C{\times}\{0\}) = \phi(C) > 0$, and $\mathcal{P}_{(\mathbf{r},q)}((\mathbf{R}_{k_0}^-, Q_{k_0}^-) \in \cdot) \geq 1_{C\times\{0\}}(\mathbf{r},q)\,\nu{\times}\delta_0(\cdot)$. Moreover, $\mathcal{P}_{(\mathbf{r},0)}((\mathbf{R}_k^-, Q_k^-) \in C{\times}\{0\}$ i.o.$)\ = \mathcal{P}_{\mathbf{r}}(\mathbf{W}_k^- \in C \text{ i.o.})\ = 1$, which implies $\{(\mathbf{R}_k^-, Q_k^-)\}$ is Harris recurrent. A similar argument proves the converse.

Now suppose $\{\mathbf{W}_k^-\}$ is positive Harris recurrent. Applying the small set minorizations of the previous paragraph, from Theorem 5.2 of Nummelin (1984), after appropriate scaling, we find that $\widehat{\pi}(B{\times}\{0\}) = \pi(B)$ for all Borel sets $B \subset E$. Put $K_C = \inf\{k \in \mathbf{Z}^+\colon\ \mathbf{W}_k^- \in C\} = \inf\{k \in \mathbf{Z}^+\colon (\mathbf{R}_k^-, Q_k^-) \in C{\times}\{0\}\}$. Then, using Corollary 5.3(a) of Nummelin (1984), we evaluate

$$
\begin{aligned}
\widehat{\pi}([0,\infty)^c{\times}\mathbf{Z}_+) &= \int_{C\times\{0\}} \mathrm{E}_{(\mathbf{r},q)}[K_C]\,\widehat{\pi}(\,d(\mathbf{r},q)\,) \\
&= \int_C \mathrm{E}_{\mathbf{w}}[K_C]\,\pi(d\mathbf{w}) \quad = \quad \pi([0,\infty)^c) \quad < \quad \infty
\end{aligned}
$$

where the finiteness follows by the assumed positive recurrence of $\{\mathbf{W}_k^-\}$. A similar argument proves the converse. ∎

3.3: Proof of Theorem 2.1

Here we establish geometric recurrence under the assumptions that $\{\mathbf{W}_k^-\}$ is irreducible, compact sets are petite, $\bar{b} < \infty$ and $\theta > 0$. For fixed $\eta \in [0,\infty)$ and $\mathbf{w} \in [0,\infty)^c$ define $v^{(\eta)}(\mathbf{w}) \stackrel{\text{def}}{=} \exp(\,\boldsymbol{\alpha}(\eta)\cdot\mathbf{w}\,)$, and

$$
\overline{h}^{(\eta)}(w) \quad \stackrel{\text{def}}{=} \quad \eta \int_w^\infty F_A^{(-\eta)}([a,\infty))\,e^{\eta(a-w)}\,da, \tag{22}
$$

where $F_A^{(-\eta)}(\cdot)$ is the twisted distribution as defined in (1). The next lemma provides the basis for a "geometric drift condition."

Lemma 3.5: For any $\eta \geq 0$ such that $\Lambda(\eta) < \infty$, we have

$$\mathrm{E}_{\mathbf{w}}[\, v^{(\eta)}(\mathbf{W}_1^-)\,] \quad \leq \quad e^{\Lambda(\eta)} \ \Big(1 \ + \ e^{\Lambda_N(\Lambda_B(\eta))}\, \overline{h}^{(\eta)}(\vee(\mathbf{w}))\ \Big)\ v^{(\eta)}(\mathbf{w}) \tag{23}$$

for all $\mathbf{w} \in [0,\infty)^c$.

Proof: Put $B^*(m) \stackrel{\text{def}}{=} B_{I(m),J_{I(m)}(m)}$ = the service time of the mth job inserted into the system. Then we have $v^{(\eta)}(\mathbf{Y}^+(m)) = \exp(\,\alpha_{I(m)}(\eta)B^*(m)\,)\, v^{(\eta)}(\mathbf{Y}^-(m))$, and hence,

$$\begin{aligned}
\mathrm{E}\Big[\, v^{(\eta)}(\mathbf{Y}^+(m)) \,\Big|\, \mathbf{Y}^-(m)\,\Big] \quad &= \quad \mathrm{E}\Big[\, \exp(\,\alpha_{I(m)}(\eta)B^*(m)\,) \,\Big|\, \mathbf{Y}^-(m)\,\Big]\ v^{(\eta)}(\mathbf{Y}^-(m)) \\
&= \quad \exp\Big(\Lambda_{I(m)}(\alpha_{I(m)}(\eta))\Big)\ v^{(\eta)}(\mathbf{Y}^-(m)) \\
&\leq \quad e^{\Lambda_B(\eta)}\ v^{(\eta)}(\mathbf{Y}^-(m))
\end{aligned}$$

because $\Lambda_i(\alpha_i(\eta)) \leq \Lambda_B(\eta)$ by (5). By iterative conditioning it follows that $\mathrm{E}_{\mathbf{w}}[\, v^{(\eta)}(\mathbf{W}_0^+) \mid N_0 = n\,]$ $\leq e^{n\Lambda_B(\eta)}\, v^{(\eta)}(\mathbf{w})$, and hence, $\mathrm{E}_{\mathbf{w}}[\, v^{(\eta)}(\mathbf{W}_0^+)\,] \leq e^{\Lambda_N(\Lambda_B(\eta))}\, v^{(\eta)}(\mathbf{w})$.

Next, for $\mathbf{w} \in [0,\infty)^c$ define

$$h^{(\eta)}(\mathbf{w}) \quad = \quad \sum_{i=1}^{c} \alpha_i(\eta) \int_{w_i}^{\infty} F_A^{(-\eta)}([a,\infty))\ \exp\Big(\sum_{\iota=1}^{c} \alpha_\iota(\eta)[a - w_\iota]^+\Big)\ da.$$

Since $W_{i,1}^- = [W_{i,0}^+ - A_1]^+ = W_{i,0}^+ - A_1 + [A_1 - W_{i,0}^+]^+$, using the twisting formula (1) we obtain

$$\begin{aligned}
&\mathrm{E}\Big[\, v^{(\eta)}(\mathbf{W}_1^-) \mid \mathbf{W}_0^+ = \mathbf{w}\,\Big] \\
&\quad = \quad \mathrm{E}\Bigg[\, \exp\Big(-\eta A_1 \ + \ \sum_{i=1}^{c} \alpha_i(\eta)[A_1 - w_i]^+\Big) \,\Bigg|\, \mathbf{W}_0^+ = \mathbf{w}\,\Bigg]\ v^{(\eta)}(\mathbf{w}) \\
&\quad = \quad e^{\Lambda_A(-\eta)} \int_0^{\infty} \exp\Big(\sum_{i=1}^{c} \alpha_i(\eta)[a - w_i]^+\Big)\ F_A^{(-\eta)}(da)\ \ v^{(\eta)}(\mathbf{w}) \\
&\quad = \quad e^{\Lambda_A(-\eta)}\ \Big[1 \ + \ h^{(\eta)}(\mathbf{w})\Big]\ v^{(\eta)}(\mathbf{w})
\end{aligned}$$

where the last equality follows from integration-by-parts. Combining this with the bound of the first paragraph and using $\Lambda(\eta) \stackrel{\text{def}}{=} \Lambda_A(-\eta) + \Lambda_N(\Lambda_B(\eta))$ yields

$$\mathrm{E}_{\mathbf{w}}[\, v^{(\eta)}(\mathbf{W}_1^-)\,] \quad \leq \quad e^{\Lambda(\eta)} \ \Big(v^{(\eta)}(\mathbf{w}) \ + \ \mathrm{E}_{\mathbf{w}}\Big[\, v^{(\eta)}(\mathbf{W}_0^+) h^{(\eta)}(\mathbf{W}_0^+)\,\Big]\Big).$$

¿From (22), observe that $\overline{h}^{(\eta)}(\cdot)$ is nonincreasing, and $\overline{h}^{(\eta)}(0) = e^{-\Lambda_A(-\eta)} - 1$ (which is evaluated by integration-by-parts and using (1)). Next observe that $h^{(\eta)}(\mathbf{w}) \leq h^{(\eta)}(\vee(\mathbf{w}),..,\vee(\mathbf{w})) = \overline{h}^{(\eta)}(\vee(\mathbf{w}))$.

Thus, in the last display, $h^{(\eta)}(\mathbf{W}_0^+) \le \overline{h}^{(\eta)}(\vee(\mathbf{W}_0^+)) \le \overline{h}^{(\eta)}(\vee(\mathbf{W}_0^-)) = \overline{h}^{(\eta)}(\vee(\mathbf{w}))$. ¿From the first paragraph, $\mathrm{E}_{\mathbf{w}}[\, v^{(\eta)}(\mathbf{W}_0^+)\,] \le e^{\Lambda_N(\Lambda_B(\eta))}\, v^{(\eta)}(\mathbf{w})$. Thus, the last display above is reduced to (23). ■

Put $\mathcal{A} = \{\mathbf{w} \in [0,\infty)^c\colon \wedge(\mathbf{w}) - \vee(\mathbf{w}) \le \overline{b}\,\}$. $\mathcal{A}$ is an absorbing set; that is, $\mathcal{P}_{\mathbf{w}}(\mathbf{W}_1^- \in \mathcal{A}) = 1$ for all $\mathbf{w} \in \mathcal{A}$, and $\mathcal{P}_{\mathbf{w}}(K_{\mathcal{A}} < \infty) = 1$ for all $\mathbf{w} \in [0,\infty)^c$ where $K_{\mathcal{A}} = \inf\{k\colon \mathbf{W}_k^- \in \mathcal{A}\}$. (Formally, this is a consequence of the proof of Lemma B.3 in Appendix B.) Thus, we may restrict $\{\mathbf{W}_k^-\}$ to the absorbing set $\mathcal{A}$.

Fix $\eta > 0$ such that $\Lambda(\eta) < 0$. Since $\overline{h}^{(\eta)}(w) \downarrow 0$, we may fix $\overline{w} < \infty$ sufficiently large so that

$$\gamma \quad = \quad e^{\Lambda(\eta)}\ \Big[1 \ + \ e^{\Lambda_N(\Lambda_B(\eta))}\,\overline{h}^{(\eta)}(\overline{w})\Big] \quad < \quad 1.$$

Put $C = \{\mathbf{w} \in \mathcal{A}\colon \vee(\mathbf{w}) \le \overline{w}\}$. Then $C \subset [0, \overline{w} + \overline{b}\,]^c$ is compact, and hence, petite by assumption. ¿From (23) we have

$$\mathrm{E}_{\mathbf{w}}[\, v^{(\eta)}(\mathbf{W}_1^-)\,] \quad \le \quad \gamma\, v^{(\eta)}(\mathbf{w}) \qquad \text{for all } \mathbf{w} \in \mathcal{A}\backslash C.$$

Also, $v^{(\eta)}(\cdot)$ is clearly bounded on C. Geometric Harris recurrence of $\{\mathbf{W}_k^-\}$ now follows by Theorem 6.2 in Meyn and Tweedie (1992).

By Lemma 3.4, $\{(\mathbf{R}_k^-, Q_k^-)\}$ is positive Harris recurrent. To get geometric recurrence, we use the following notion. Let $\{X_k\}$ be a generic positive Harris recurrent Markov chain, and put $K_B = \inf\{k \ge 1\colon X_k \in B\}$. A set $D \in \mathcal{S}_\pi$ is *geometrically regular* if for all $B \in \mathcal{S}_\pi$ there is an $r > 1$ such that $\sup_{x\in D} \mathrm{E}_x[\, r^{K_B}\,] < \infty$. Clearly, any subset of a geometrically regular set is geometrically regular.

By Theorem 6.2 of Meyn and Tweedie (1992), the set C above is geometrically regular for $\{\mathbf{W}_k^-\}$, hence, so is $\widetilde{C} = \{\mathbf{w}\colon \vee(\mathbf{w}) = 0 \text{ and } \wedge(\mathbf{w}) \le \overline{b}\} \subset C$ (if we take $\overline{w} > \overline{b}$). We also have $\mathcal{P}_{\mathbf{w}}(\mathbf{W}_k^- \in \widetilde{C} \text{ i.o.}) = 1$ by Lemma 3.2. Put $\widehat{C} = C \times \{0\}$. Since $(\mathbf{R}_{K_\ell}^-, Q_{K_\ell}^-) = (\mathbf{W}_{K_\ell}^-, 0)$, we have $\mathcal{P}_{(\mathbf{r},q)}((\mathbf{R}_k^-, Q_k^-) \in \widehat{C} \text{ i.o.}) = 1$, which implies $\widehat{C} \in \mathcal{S}_{\widehat{\pi}}$, and by geometric regularity of $\widetilde{C}$ we have

$$\sup_{(\mathbf{r},q)\in\widehat{C}} \mathrm{E}_{(\mathbf{r},q)}\Big[\, r^{K_{\widehat{C}}}\Big] \quad = \quad \sup_{\mathbf{w}\in\widetilde{C}} \mathrm{E}_{\mathbf{w}}\Big[\, r^{K_{\widetilde{C}}}\Big] \quad < \quad \infty.$$

Thus, $\{(\mathbf{R}_k^-, Q_k^-)\}$ is also geometrically Harris recurrent.

3.4: Proof of Theorem 2.2

Define $\mathcal{E} \stackrel{\text{def}}{=} \{\mathbf{e} \in [0,1]^c\colon \sum_{i=1}^c e_i = 1\}$. Assuming $\lambda < \mu$, Theorem 2.2 states that the following are equivalent: (i) $\theta = \infty$; (ii) there exists a vector $\mathbf{e}^* \in \mathcal{E}$ such that $e_i^*\,\overline{b}_i\,\overline{n} \le \underline{a}$ for each i; and (iii) the queue is bounded.

For $e \in [0,1]$ and $\eta \in \mathbf{R}$ define

$$\widehat{\Lambda}_i(\eta; e) \;\stackrel{\text{def}}{=}\; \Lambda_A(-\eta) \;+\; \Lambda_N(\Lambda_i(\eta e)) \;=\; \log\left(\mathrm{E}\left[\exp\left(\eta \left\{ e \sum_{j=1}^{N} B_{i,j} - A \right\} \right) \right] \right),$$

which is the cumulant function of the random variable $e\sum_{j=1}^N B_{i,j} - A$. For $\eta \in \mathbf{R}$ and $\mathbf{e} \in \mathcal{E}$ define $\widehat{\Lambda}(\eta;\mathbf{e}) \stackrel{\text{def}}{=} \max_{i=1,..,c} \widehat{\Lambda}_i(\eta; e_i)$, and note that $\widehat{\Lambda}(\eta;\mathbf{e})$ is jointly l.s.c. in $(\eta;\mathbf{e})$. Also observe that $\Lambda(\eta) = \inf_{\mathbf{e}\in\mathcal{E}} \widehat{\Lambda}(\eta;\mathbf{e})$.

Define $\theta^*(\mathbf{e}) \stackrel{\text{def}}{=} \sup\{\eta \ge 0\colon \widehat{\Lambda}(\eta;\mathbf{e}) \le 0\}$ for each $\mathbf{e} \in \mathcal{E}$. Then $\theta = \sup\{\eta\colon \Lambda(\eta) \le 0\} = \sup_{\mathbf{e}\in\mathcal{E}} \sup\{\eta\colon \widehat{\Lambda}(\eta;\mathbf{e}) \le 0\} = \sup_{\mathbf{e}\in\mathcal{E}} \theta^*(\mathbf{e})$. Moreover, for all $\eta \in [0,\infty)$, $\{\mathbf{e}\colon \theta^*(\mathbf{e}) \ge \eta\} = \{\mathbf{e}\colon \widehat{\Lambda}(\eta;\mathbf{e}) \le 0\}$ = a closed set because $\widehat{\Lambda}(\eta;\cdot)$ is l.s.c., and hence, $\theta^*(\cdot)$ is u.s.c.

Observe that the following statements are equivalent: (a) $\theta^*(\mathbf{e}) = \infty$; (b) $\widehat{\Lambda}_i(\eta; e_i) \le 0$ for all $\eta > 0$ and each $i = 1,..,c$; (c) $e_i \sum_{j=1}^N B_{i,j} - A \le 0$ a.s. for each i; and (d) e_i ess sup $\sum_{j=1}^{N_1} B_{i,j} = e_i\,\overline{n}\,\overline{b}_i \le \underline{a}$ for each i. Statement (ii) of Theorem 2.2 says that (d) holds for some $\mathbf{e}^* \in \mathcal{E}$, but we may equivalently work with (a), (b) or (c).

Assume (ii), equivalently, $\theta^*(\mathbf{e}^*) = \infty$ for some $\mathbf{e}^* \in \mathcal{E}$. Then $\theta \ge \theta^*(\mathbf{e}^*) = \infty$. Thus, (ii) $\Rightarrow$ (i). Suppose that (ii) fails, equivalently, $\theta^*(\mathbf{e}) < \infty$ for all $\mathbf{e} \in \mathcal{E}$. Since $\theta^*(\cdot)$ is u.s.c., we have $\theta = \sup_{\mathbf{e}\in\mathcal{E}} \theta^*(\mathbf{e}) < \infty$, and hence, (i) fails. Thus, (i) $\Leftrightarrow$ (ii).

Next we show (ii) $\Rightarrow$ (iii). Assume (ii) and define $v(\mathbf{w}) \stackrel{\text{def}}{=} \sum_{i=1}^c e_i^*\,w_i$. For any i such that $e_i^* > 0$, we have $W_k^+ \le W_{i,k}^+ \le v(\mathbf{W}_k^+)/e_i^*$, and hence, it is sufficient to show that $v(\mathbf{W}_k^+)$ is bounded. First, as in the proof of Theorem 2.1, we may restrict attention to the absorbing set $\mathcal{A} \stackrel{\text{def}}{=} \{\mathbf{w} \in [0,\infty)^c\colon w_i - \vee(\mathbf{w}) \le \overline{b}_i$ for each $i\}$. Next, write

$$v(\mathbf{W}_k^+) \;=\; v(\mathbf{W}_k^-) \;+\; \sum_{i=1}^{c} \sum_{j=J_i(M_{k-1}+1)}^{J_i(M_k)} e_i^*\,B_{i,j}.$$

By applying $e_i^* B_{i,j} \le e_i^*\overline{b}_i \le \underline{a}/\overline{n}$ and $\sum_{i=1}^c [J_i(M_{k-1}+1) - J_i(M_k)] = N_k \le \overline{n}$, the last display is reduced to the bound $v(\mathbf{W}_k^+) \le v(\mathbf{W}_k^-) + \underline{a}$. Put $C = \{\mathbf{w} \in \mathcal{A}\colon \vee(\mathbf{w}) \le \underline{a}\}$. When $\mathbf{W}_k^+ \in \mathcal{A}\backslash C$ we have $W_{i,k+1}^- = [W_{i,k}^+ - A_{k+1}]^+ \le W_{i,k}^+ - \underline{a}$, and hence, $v(\mathbf{W}_{k+1}^-) \le v(\mathbf{W}_k^+) - \underline{a}$. Thus, combining these bounds yields

$$v(\mathbf{W}_{k+1}^{+}) \quad \leq \quad v(\mathbf{W}_{k}^{+}) \qquad \text{whenever} \quad \mathbf{W}_{k}^{+} \in \mathcal{A}\backslash C.$$

Observe that $\lambda \leq \overline{n}/\underline{a} = \sum_{i=1}^{c} e_i^* \overline{n}/\underline{a} \leq \sum_{i=1}^{c} \overline{b}_i^{-1} \leq \mu$. Since we assume $\lambda < \mu$, one of these inequalities must be strict, and this implies that there exists $\epsilon, \delta > 0$ and an integer $1 \leq m < \infty$ such that $\mathcal{P}\Big(v(\mathbf{W}_{k+m}^{+}) \leq v(\mathbf{W}_{k}^{+}) - \epsilon \,|\mathbf{W}_{k}^{+} = \mathbf{w}\Big) > \delta$ for all $\mathbf{w} \in \mathcal{A}\backslash C$. (For example, $W_{i,k+1}^{-} \leq w_i - (\underline{a} + \epsilon)$ with probability $\delta = F_A([\underline{a} + \epsilon, \infty))$.) Put $K \stackrel{\text{def}}{=} \inf\{k\colon\ \mathbf{W}_k^{+} \in C\}$, so we have $\mathcal{P}(K < \infty | \mathbf{W}_0^{+} = \mathbf{w}) = 1$ for all $\mathbf{w} \notin C$. Whenever $\mathbf{W}_k^{+} \in C$, we have $v(\mathbf{W}_k^{+}) \leq \underline{a} + \sum_{i=1}^{c} e_i^* \overline{b}_i$ and $v(\mathbf{W}_{k+1}^{+}) \leq \overline{v} \stackrel{\text{def}}{=} 2\underline{a} + \sum_{i=1}^{c} e_i^* \overline{b}_i$. Thus, $v(\mathbf{W}_k^{+}) \leq \overline{v}$ for all $k \geq K$; that is, the queue is bounded.

To complete the proof we show that (iii) $\Rightarrow$ (ii) using a contrapositive argument. Assume (ii) fails. If either $\underline{a} = 0$ or $\overline{n} = \infty$, then obviously (iii) fails (the queue is not bounded). So assume $\underline{a} > 0$ and $\overline{n} < \infty$. Put $\tilde{e}_i = \underline{a}/(\overline{b}_i \overline{n})$, so $\sum_{i=1}^{c} \tilde{e}_i < 1$ by the assumption that (ii) fails. If $\sum_{i=1}^{c} \tilde{e}_i = 0$, then each $\overline{b}_i = \infty$, in which case (iii) fails. So, suppose $\sum_{i=1}^{c} \tilde{e}_i > 0$. Put $e_i^* = \tilde{e}_i / \sum_{\iota=1}^{c} \tilde{e}_\iota$, so $\mathbf{e}^* \in \mathcal{E}$ and $\overline{b}_i > \underline{a}/(e_i^* \overline{n})$ for each i since (ii) fails. Fix $\epsilon > 0$ sufficiently small so that $\overline{b}_i > (\underline{a} + \epsilon)/(e_i^* \overline{n})$. Then with positive probability we can have an arbitrarily long run of A_ks $\leq \underline{a} + \epsilon$ and N_ks $= \overline{n}$. For $e_i^* > 0$, we can have an arbitrarily long run of $B_{i,j}$s $\geq (\underline{a} + \epsilon)/(e_i^* \overline{n})$. For $e_i^* = 0$ the $B_{i,j}$'s can be arbitrarily large for an arbitrarily long run. During such a run, $W_k^{\pm}$ is lower bounded by the waiting time of a deterministic queue with $A_k \equiv \underline{a} + \epsilon$, $N_k \equiv \overline{n}$, and $B_{i,j} \equiv (\underline{a} - \epsilon)/(e_i^* \overline{n})$ when $e_i^* > 0$. (We omit the obvious details.) The deterministic queue is unstable, and hence, $W_k^{\pm}$ can obtain arbitrarily large values, that is, (iii) fails. Thus, by contrapositive, (iii) $\Rightarrow$ (ii), and this completes the proof.

3.5: Cycles

We call the time intervals $[T_{K_\ell}, T_{K_{\ell+1}})$ *cycles*. This differs from the usual definition found in the literature that is based on idle period regeneration. In our definition, $W_{K_\ell}^{-} = \vee(\mathbf{W}_{K_\ell}^{-}) = 0$ does not imply the vector equality $\mathbf{W}_{K_\ell}^{-} = \mathbf{0}$.

Put $L_w^{+} \stackrel{\text{def}}{=} \sum_{k=0}^{K_1 - 1} 1_{[w,\infty)}(W_k^{+}) =$ the total number of batches that arrive during the first cycle and produce $W_k^{+} \geq w$. Likewise, $L_n^{+} \stackrel{\text{def}}{=} \sum_{k=0}^{K_1 - 1} 1_{[n,\infty)}(Q_k^{+})$ is the total number of batches during the first cycle that produce queue length $Q_k^{+} \geq n$. The following lemma is an immediate consequence of positive Harris recurrence and Lemma 3.2.

Lemma 3.6: Assume $\lambda < \mu$, and that $\{\mathbf{W}_k^-\}$ and $\{(\mathbf{R}_k^-, Q_k^-)\}$ are positive Harris recurrent. Then

$$\mathcal{P}_\pi(\, W_k^+ \geq w\,) \quad = \quad \frac{\mathrm{E}_\xi[\, L_w^+\,]}{\mathrm{E}_\xi[K_1]} \tag{24}$$

where $\xi(\cdot) = \mathcal{L}_\pi(\mathbf{W}_{K_\ell}^-)$, and

$$\mathcal{P}_{\widehat{\pi}}(\, Q_k^+ \geq n\,) \quad = \quad \frac{\mathrm{E}_{\widehat{\xi}}[\, L_n^+\,]}{\mathrm{E}_{\widehat{\xi}}[K_1]} \tag{25}$$

where $\widehat{\xi}(\cdot) = \mathcal{L}_{\widehat{\pi}}(\mathbf{R}_{K_\ell}^-, Q_{K_\ell}^-) = \xi \times \delta_0(\cdot)$.

Remark: We consider only the stationary distribution of the waiting times just after a batch of jobs arrives. However, we could just as easily work with other the stationary distributions. To get the stationary waiting time distribution sampled just before a batch of jobs arrives, replace L_w^+ by $L_w^- = \sum_{k=0}^{K_1-1} 1_{[w,\infty)}(W_k^-)$. To get the stationary waiting time distribution for arbitrary time, replace L_w^+ by $L_w^* = \int_0^{T_{K_1}} 1_{[w,\infty)}(W(t))\, dt$ = total time during the first cycle that $W(t) \geq w$, and replace $\mathrm{E}_\xi[K_1]$ by $\mathrm{E}_\xi[T_{K_1}]$. Similar statements hold for queue length.

We also remark that (24) and (25) are our main departure from the classical GI/G/1 analysis that relates the queuing problem to a level crossing problem via Lindley's identity. These are not a simple level crossing probabilities, even in the GI/G/1 case. Nonetheless, random walk theory still plays a key role.

4. Workload and Waiting Time Analysis

In this section we prove the waiting time limits of Theorem 2.3. Define $\mathcal{F}_k \stackrel{\text{def}}{=} \sigma(\,\mathbf{W}_0^-\,;\, A_1, .., A_k;\, N_0, .., N_k\,;\, B_{i,j},\, j = 1, .., \widetilde{J}_i(M_k),\, i = 1, .., c\,)$ where $\widetilde{J}_i(m)$ is as defined in Section 3.1. Let K be any $\mathcal{F}_k$-stopping time. Fix $(\boldsymbol{\alpha}, \nu, \eta) \in \mathbf{R}^{(c+2)}$ such that $\Lambda_i(\alpha_i) < \infty$ for each i, $\Lambda_N(\nu) < \infty$ and $\Lambda_A(-\eta) < \infty$, and let $F_A^{(-\eta)}(\cdot)$, $F_N^{(\nu)}(\cdot)$ and $F_i^{(\alpha_i)}(\cdot)$ denote the exponentially twisted distributions as in (1). We define the $(\boldsymbol{\alpha}, \nu, \eta, K)$-*conjugate distribution* $\mathcal{P}_\zeta^{(\boldsymbol{\alpha}, \nu, \eta, K)}(\cdot)$ as follows. First, the distribution is defined on $\mathcal{F}_{-1} = \sigma(\mathbf{W}_0^-)$ as $\mathcal{P}_\zeta^{(\boldsymbol{\alpha}, \nu, \eta, K)}(\mathbf{W}_0^- \in \cdot\,) = \zeta(\cdot)$. Given the definition on $\mathcal{F}_k$ we extend the definition to $\mathcal{F}_{k+1}$ by defining the conditional law of A_{k+1}, N_{k+1} and $B_{i,j}$, $j = \widetilde{J}_i(M_k) + 1, .., \widetilde{J}_i(M_{k+1})$. On the event $\{K > k\} \in \mathcal{F}_k$, this conditional law is just the queuing process governed by $F_A^{(-\eta)}(\cdot)$, $F_N^{(\nu)}(\cdot)$, $F_i^{(\alpha_i)}(\cdot)$, $i = 1, .., c$. On the event $\{K \leq k\} \in \mathcal{F}_k$, the process evolves with the original untwisted distributions $F_A(\cdot)$, $F_N(\cdot)$ and $F_i(\cdot)$.

The restrictions of $\mathcal{P}_\zeta^{(\boldsymbol{\alpha},\nu,\eta,K)}(\cdot)$ and $\mathcal{P}_\zeta(\cdot)$ to the event $\{K < \infty\}$ are equivalent measures, and on this event the likelihood ratio is

$$\frac{d\mathcal{P}_\zeta}{d\mathcal{P}_\zeta^{(\boldsymbol{\alpha},\nu,\eta,K)}} \;=\; \exp\Bigg(\eta\, T_K \;+\; K\,\Lambda_A(-\eta) \;-\; \nu\, M_K \;+\; (K+1)\,\Lambda_N(\nu) \\ \qquad\qquad + \sum_{i=1}^{c} \Big[-\alpha_i \,\Big(\widetilde{S}_i(\widetilde{J}_i(M_K)) - W_{i,0}^- \Big) \;+\; \widetilde{J}_i(M_K)\,\Lambda_i(\alpha_i) \Big] \Bigg). \tag{26}$$

The above likelihood ratio formula is significantly simplified by taking $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\eta)$ and $\nu = \Lambda_B(\eta)$. ¿From Lemma 2.1, $\Lambda_i(\alpha_i(\eta)) \le \Lambda_B(\eta)$ with equality if and only if $\eta \le \overline{\eta}_c$. Thus, since $\sum_{i=1}^c \widetilde{J}_i(M_K) = M_K$, we have

$$-\;\Lambda_B(\eta)\, M_K \;+\; \sum_{i=1}^{c} \widetilde{J}_i(M_K)\,\Lambda_i(\alpha_i(\eta)) \;=\; \sum_{i=1}^{c} \widetilde{J}_i(M_K)\,[\,\Lambda_i(\alpha_i(\eta)) \;-\; \Lambda_B(\eta)\,] \;\le\; 0. \tag{27}$$

Next, $\widetilde{W}_{i,k}^+ = \widetilde{Y}_i^+(M_k) = \widetilde{S}_i(\widetilde{J}_i(M_k)) - T_k$. Thus, using $\sum_{i=1}^c \alpha_i(\eta) = \eta$ we obtain

$$\eta\, T_K \;-\; \sum_{i=1}^{c} \alpha_i(\eta)\; \Big(\widetilde{S}_i(\widetilde{J}_i(M_K)) - W_{i,0}^- \Big) \\ \;=\; -\sum_{i=1}^{c} \alpha_i(\eta)\,(\widetilde{W}_{i,K}^+ - W_{i,0}^-) \;=\; -\,\boldsymbol{\alpha}(\eta)\cdot(\widetilde{\mathbf{W}}_K^+ - \mathbf{W}_0^-\,).$$

The remaining terms are

$$K\,\Lambda_A(-\eta) \;+\; (K+1)\,\Lambda_N(\Lambda_B(\eta)) \;=\; K\,\Lambda(\eta) \;+\; \Lambda_N(\Lambda_B(\eta)).$$

Finally, observe that $\Lambda_i(\alpha_i(\eta)) < \infty$ for each i, $\Lambda_N(\Lambda_B(\eta)) < \infty$ and $\Lambda_A(-\eta) < \infty$, if and only if $\eta \in \mathcal{D} \stackrel{\text{def}}{=} \{\eta\colon\ \Lambda(\eta) < \infty\}$.

We hereafter refer to the $(\boldsymbol{\alpha}(\eta), \Lambda_B(\eta), \eta, K)$-conjugate distribution as simply the (η, K)-conjugate distribution, and we modify the notation accordingly. By the above arguments, we have the following change of distribution formula.

Lemma 4.1: For any $\eta \in \mathcal{D}$, the restrictions of $\mathcal{P}_\zeta^{(\eta,K)}(\cdot)$ and $\mathcal{P}_\zeta(\cdot)$ to the event $\{K < \infty\}$ are equivalent, and on this event we have

$$\frac{d\mathcal{P}_\zeta}{d\mathcal{P}_\zeta^{(\eta,K)}} \;=\; \exp\Bigg(-\;\boldsymbol{\alpha}(\eta)\cdot(\widetilde{\mathbf{W}}_K^+ - \mathbf{W}_0^-\,) \;+\; K\,\Lambda(\eta) \;+\; \Lambda_N(\Lambda_B(\eta)) \\ \qquad\qquad + \sum_{i=1}^{c} \widetilde{J}_i(M_K)\,[\,\Lambda_i(\alpha_i(\eta)) \;-\; \Lambda_B(\eta)\,] \Bigg) \tag{28}$$

and the last term vanishes when $\eta \le \overline{\eta}_c$.

We will also have cause to consider the distribution $\mathcal{P}_\zeta^{(\eta)}(\cdot)$ having $\mathcal{L}(A_k) = F_A^{(-\eta)}(\cdot)$ and $\mathcal{L}(N_k) = F_N^{(\Lambda_B(\eta))}(\cdot)$ for all k, and $\mathcal{L}(B_{i,j}) = F_i^{(\alpha_i(\eta))}(\cdot)$ for all i, j. We will call this simply the η-conjugate distribution, and we may think of it as the (η, K)-conjugate distribution with $K \equiv \infty$. Notice that $\mathcal{P}_\zeta^{(\eta,K)}(\cdot) = \mathcal{P}_\zeta^{(\eta)}(\cdot)$ on $\mathcal{F}_K$.

Recall $\lambda(\eta) \stackrel{\text{def}}{=} \Lambda'_N(\Lambda_B(\eta))/\Lambda'_A(-\eta)$ as defined in (9). Since $\mathrm{E}^{(-\eta)}[A] = \Lambda'_A(-\eta)$ and $\mathrm{E}^{(\Lambda_B(\eta))}[N] = \Lambda'_N(\Lambda_B(\eta))$, we now observe that $\lambda(\eta)$ is indeed the η-conjugate arrival rate. Likewise, since $\mathrm{E}^{(\alpha_i(\eta))}[B_i] = \Lambda'_i(\alpha_i(\eta))$, we now see that $\mu(\eta) \stackrel{\text{def}}{=} \sum_{i=1}^c \Lambda'_i(\alpha_i(\eta))^{-1}$ is the η-conjugate service rate.

Now, assume that $0 < \theta < \infty$, $\theta \le \overline{\eta}_c$ and $\Lambda(\theta) = 0$. By convexity we have $\Lambda'(\theta) > 0$, and hence, from the differentiation formula (8) we have $\lambda(\theta) > \widehat{\mu}(\theta) = \mu(\theta)$. (See the remarks at the end of Section 2.1 regarding $\widehat{\mu}(\eta) \le \mu(\eta)$.) Thus, by Lemma 3.2, the θ-conjugate workload process is unstable. For $w > 0$ define

$$K_w^+ \stackrel{\text{def}}{=} \inf\{k : \widetilde{W}_k^+ \ge w\}, \tag{29}$$

which is an $\mathcal{F}_k$-stopping time. Since $\mathbf{W}_k^\pm = \widetilde{\mathbf{W}}_k^\pm$ for $k < K_1$ (by Lemma 3.1), the random variables K_1, $\overline{W}$ and L_w^+ that were originally defined in terms of $\{(\mathbf{W}_k^-, \mathbf{W}_k^+)\}$ can also be determined by $\{(\widetilde{\mathbf{W}}_k^-, \widetilde{\mathbf{W}}_k^+)\}$. We have $K_1 = \inf\{k \ge 1\colon\ \widetilde{W}_k^- \le 0\}$, $\overline{W} = \max_{k<K_1} \widetilde{W}_k^+$, and $L_w^+ = \sum_{k=0}^{K_1-1} 1_{[w,\infty)}(\widetilde{W}_k^+)$. Moreover, $\{\overline{W} \ge w\} = \{K_w^+ < K_1\}$, and $L_w^+ = 0$ on the event $\{\overline{W} < w\}$. Apply Lemma 4.1 with $K = K_w^+$. Since we assume $\Lambda(\theta) = 0$, we may delete $K_w^+\,\Lambda(\theta) = 0$ in (28). Since we assume $\theta \le \overline{\eta}_c$, we may delete the $\widetilde{J}_i(M_K)[\Lambda_i(\alpha_i(\theta)) - \Lambda_B(\theta)] = 0$ terms in (28). Thus, the expectation in the numerator of (24) written as a (θ, K_w^+)-conjugate expectation is

$$\mathrm{E}_\zeta[\,L_w^+\,] \;=\; e^{\Lambda_N(\Lambda_B(\theta))}\,\mathrm{E}_\zeta^{(\theta,K_w^+)}\Big[\,L_w^+\,\exp\Big(-\boldsymbol{\alpha}(\theta)\cdot(\widetilde{\mathbf{W}}_{K_w^+}^+ - \mathbf{W}_0^-)\Big)\;;\;\overline{W} \ge w\,\Big].$$

Define $\widetilde{\Delta}_{i,k}^+ \stackrel{\text{def}}{=} \widetilde{W}_{i,k}^+ - \widetilde{W}_k^+ =$ the excess of the ith server's workload over $\widetilde{W}_k^+ = \vee(\widetilde{\mathbf{W}}_k^+)$. As a vector write $\widetilde{\boldsymbol{\Delta}}_k^+ \stackrel{\text{def}}{=} (\widetilde{\Delta}_{1,k}^+, .., \widetilde{\Delta}_{c,k}^+)$. Since $\sum_{i=1}^c \alpha_i(\theta) = \theta$, we have $\boldsymbol{\alpha}(\theta)\cdot\widetilde{\mathbf{W}}_k^+ = \theta\widetilde{W}_k^+ + \boldsymbol{\alpha}(\theta)\cdot\widetilde{\boldsymbol{\Delta}}_k^+$. Thus, we may rewrite the last display as

$$\begin{aligned}\mathrm{E}_\zeta[\,L_w^+\,] \;=\; e^{\Lambda_N(\Lambda_B(\theta))}\,\mathrm{E}_\zeta^{(\theta,K_w^+)}\Big[\,L_w^+\,\exp\Big(&-\theta(\widetilde{W}_{K_w^+}^+ - w) \;-\; \boldsymbol{\alpha}(\theta)\cdot\widetilde{\boldsymbol{\Delta}}_{K_w^+}^+ \\ &+\; \boldsymbol{\alpha}(\theta)\cdot\mathbf{W}_0^-\Big)\;;\;\overline{W} \ge w\,\Big]\;e^{-\theta w}.\end{aligned}$$

On the event $\{\overline{W} \ge w\} \in \mathcal{F}_{K_w^+}$, L_w^+ is determined by the two $\mathcal{F}_{K_w^+}$-measurable quantities, $\widetilde{\boldsymbol{\Delta}}_{K_w^+}^+$ and $\widetilde{W}_{K_w^+}^+ - w$, and by the random variables $\{A_k,\, N_k,\, B_{i,j};\, j > \widetilde{J}_i(M_k),\, k > K_w^+\}$ that are independent of $\mathcal{F}_{K_w^+}$ and governed by the untwisted process distribution. For $z \ge 0$ and $\boldsymbol{\delta} \in [0,\infty)^c$ define

$$g_w^{(\eta)}(\boldsymbol{\delta}, z) \;\stackrel{\text{def}}{=}\; \mathrm{E}\Big[\,L_w^+\;\Big|\;\widetilde{\boldsymbol{\Delta}}_0^+ = \boldsymbol{\delta}, W_0^+ = w + z\,\Big]\;\exp\Big(-\,\eta z \;-\; \boldsymbol{\alpha}(\eta)\cdot\boldsymbol{\delta}\Big) \tag{30}$$

where the expectation is with respect to the untwisted distribution. Also, recall $v^{(\eta)}(\mathbf{w}) \stackrel{\text{def}}{=} \exp(\boldsymbol{\alpha}(\eta)\cdot\mathbf{w})$. Then by conditioning on $\mathcal{F}_{K_w^+}$ we get

$$\mathrm{E}_\zeta[\,L_w^+\,] \;=\; e^{\Lambda_N(\Lambda_B(\theta))}\,\mathrm{E}_\zeta^{(\theta)}\Big[\,g_w^{(\theta)}\Big(\widetilde{\boldsymbol{\Delta}}^+_{K_w^+},\widetilde{W}^+_{K_w^+}-w\Big)\;v^{(\theta)}(\mathbf{W}_0^-)\;;\;\overline{W}\geq w\,\Big]\;e^{-\theta w}\,. \tag{31}$$

Notice that (31) is written as the expectation with respect to the θ-conjugate distribution, instead of the (θ, K_w^+)-conjugate distribution. Since $\widetilde{\boldsymbol{\Delta}}^+_{K_w^+}$, $\widetilde{W}^+_{K_w^+}$ and $\mathbf{W}_0^-$, and the event $\{\overline{W}\geq w\}$ are all $\mathcal{F}_{K_w^+}$-measurable the θ-conjugate and (θ, K_w^+)-conjugate expectations are identical. Also, we reiterate that (31) holds only when $0<\theta<\infty$, $\theta\leq\overline{\eta}_c$ and $\Lambda(\theta)=0$.

It turns out that $\{(\widetilde{\boldsymbol{\Delta}}^+_k, \widetilde{W}^+_k)\}$ is a *Markov additive process*. Lemma 4.2 below, which is proved in Appendices A and B, is a consequence of the renewal theory for Markov additive processes.

Lemma 4.2: Assume the $F_i(\cdot)$s are spread out, $0<\theta<\infty$, $\lambda(\theta)>\mu(\theta)$, $\Lambda_i'(\alpha_i(\theta))<\infty$ for each i and $\Lambda_N'(\Lambda_B(\theta))<\infty$. Then there exists a unique probability distribution $\psi^+(\cdot)$ on $[0,\infty)^{c+1}$ such that the following holds. Let g: $[0,\infty)^{c+1}\to\mathbf{R}$ be bounded and measurable, and let $g_w(\cdot,\cdot)$ be uniformly bounded measurable functions such that $g_w(\cdot,\cdot)\to g(\cdot,\cdot)$ pointwise. Let $\zeta(\cdot)$ be a distribution and $v(\cdot)\geq 0$ be a measurable function on $[0,\infty)^c$ such that $\mathrm{E}_\zeta[v(\mathbf{W}_0^-)]<\infty$. Then

$$\begin{aligned}\lim_{w\to\infty}\;&\mathrm{E}_\zeta^{(\theta)}\Big[\,g_w\Big(\widetilde{\boldsymbol{\Delta}}^+_{K_w^+},\widetilde{W}^+_{K_w^+}-w\Big)\;v(\mathbf{W}_0^-)\;;\;\overline{W}\geq w\,\Big]\\ &=\;\mathrm{E}_{\psi^+}\Big[\,g\Big(\widetilde{\boldsymbol{\Delta}}^+_\infty,V_\infty^+\Big)\,\Big]\;\mathrm{E}_\zeta^{(\theta)}\Big[\,v(\mathbf{W}_0^-)\;;\;\overline{W}=\infty\,\Big]\end{aligned} \tag{32}$$

where $\mathcal{L}\Big(\widetilde{\boldsymbol{\Delta}}^+_\infty, V_\infty^+\Big)=\psi^+(\cdot)$.

Comparing (31) and (32), our task now is to verify the conditions of Lemma 4.2. First observe, as indicated in Section 2.1, that the conditions $0<\theta<\infty$, $\Lambda(\theta)=0$ and $\theta\leq\overline{\eta}_c$ required for (31) also imply the condition $\lambda(\theta)>\mu(\theta)$ needed by Lemma 4.2. With the additional conditions $\Lambda_i'(\alpha_i(\theta))<\infty$ and $\Lambda_N'(\Lambda_B(\theta))<\infty$, we have all of the conditions listed in part (i) of Theorem 2.3. Next, recall that $\zeta(\cdot)$ is admissible if $\mathrm{E}_\zeta[v^{(\theta)}(\mathbf{W}_0^-)]<\infty$ and $\mathcal{P}_\zeta(\overline{W}\geq w)>0$ for all $w<\infty$. $\lambda(\theta)>\mu(\theta)$ implies that the θ-conjugate distribution is unstable. Since $\mathcal{P}_\zeta(\,\cdot\,;K_w^+<\infty)$ and $\mathcal{P}_\zeta^{(\theta)}(\,\cdot\,;K_w^+<\infty)$ are equivalent, $\mathcal{P}_\zeta(\overline{W}>w)>0$ for all $w<\infty$ implies $\mathcal{P}_\zeta^{(\theta)}(\overline{W}>w)>0$ for all $w<\infty$, which in turn implies $\mathcal{P}_\zeta^{(\theta)}(\overline{W}=\infty)>0$ because of the θ-conjugate instability. We will conclude that for any admissible initial distribution

$$0\;<\;\mathrm{E}_\zeta^{(\theta)}\Big[\,v^{(\theta)}(\mathbf{W}_0^-)\;;\;\overline{W}=\infty\,\Big]\;<\;\infty. \tag{33}$$

Thus, we must establish two facts: 1) $\xi(\cdot) \stackrel{\text{def}}{=} \mathcal{L}_\pi(\mathbf{W}^-_{K_\ell})$ (from Lemma 3.6) is admissible, and 2) $g^{(\theta)}_w(\cdot,\cdot) \uparrow g^{(\theta)}(\cdot,\cdot)$ = a strictly positive and bounded measurable function.

Lemma 4.3: Assume $\{\mathbf{W}^-_k\}$ and $\{(\mathbf{R}^-_k, Q^-_k)\}$ are positive Harris recurrent. Then $\mathrm{E}_\xi[v^{(\eta)}(\mathbf{W}^-_0)]$ $= \mathrm{E}_{\hat\xi}[v^{(\eta)}(\mathbf{R}^-_0)] < \infty$ for any $\eta \in \mathcal{D}$.

Proof: Recall from Section 3 that $(\mathbf{R}^-_0, Q^-_0) = (\mathbf{W}^-_0, 0)$ on the event $\{W^-_0 = 0\}$. Since $\xi(\cdot) = \mathcal{P}_\pi(\mathbf{W}^-_0 \in \cdot\,|W^-_0 = 0)$, and $\hat\xi(\cdot) = \xi \times \delta_0(\cdot)$, it is sufficient to show $\mathrm{E}_{\hat\pi}[v^{(\eta)}(\mathbf{R}^-_0)] < \infty$. Put $K(t) \stackrel{\text{def}}{=} \sup\{k\colon\, T_k \le t\}$. Then $v^{(\eta)}(\mathbf{R}(t)) \ge e^{-\eta A_k}\, v^{(\eta)}(\mathbf{R}^-_{k-1})$ for $t \in [T_{k-1}, T_k]$, and hence,

$$\frac{1}{t}\int_0^t \mathrm{E}_{(\mathbf{0},0)}[\,v^{(\eta)}(\mathbf{R}(s))\,]\,ds \quad\ge\quad \mathrm{E}_{(\mathbf{0},0)}\!\left[\,\frac{K(t)}{t}\,\frac{1}{K(t)}\sum_{k=1}^{K(t)} e^{-\eta A_k}\, v^{(\eta)}(\mathbf{R}^-_{k-1})\,\right].$$

By the law of large numbers, $K(t)/t \to 1/\mathrm{E}[A]$ a.s. By positive Harris recurrence, and since $v^{(\eta)}(\cdot) \ge 0$, we have $\frac{1}{k}\sum_{\kappa=1}^{k} e^{-\eta A_\kappa}\, v^{(\eta)}(\mathbf{R}^-_{\kappa-1}) \to e^{\Lambda_A(-\eta)}\,\mathrm{E}_{\hat\pi}[v^{(\eta)}(\mathbf{R}^-_0)]$ a.s. $\mathcal{P}_{(\mathbf{0},0)}$, even if $\mathrm{E}_{\hat\pi}[v^{(\eta)}(\mathbf{R}^-_0)] = \infty$. The same is true if we replace k by $K(t)$ and let $t \uparrow \infty$. An application of Fatou's Lemma yields

$$\frac{e^{\Lambda_A(-\eta)}}{\mathrm{E}[A]}\,\mathrm{E}_{\hat\pi}[v^{(\eta)}(\mathbf{R}^-_0)] \quad\le\quad \liminf_{t\to\infty}\frac{1}{t}\int_0^t \mathrm{E}_{(\mathbf{0},0)}[v^{(\eta)}(\mathbf{R}(s))]\,ds \quad\le\quad \sup_{t>0}\mathrm{E}_{(\mathbf{0},0)}[\,v^{(\eta)}(\mathbf{R}(t))\,].$$

So, it is sufficient to show that $\sup_{t>0}\mathrm{E}_{(\mathbf{0},0)}[v^{(\eta)}(\mathbf{R}(t))] < \infty$.

For $\mathcal{I} \subset \{1,..,c\}$ put $E_{\mathcal{I}} \stackrel{\text{def}}{=} \{\mathbf{r}\colon\; r_i > 0$ for $i \in \mathcal{I}$ and $r_i = 0$ for $i \notin \mathcal{I}\}$ and $v^{(\eta)}_{\mathcal{I}}(\mathbf{r}) \stackrel{\text{def}}{=} v^{(\eta)}(\mathbf{r})\,1_{E_{\mathcal{I}}}(\mathbf{r})$. Then, $v^{(\eta)}(\mathbf{r}) = \sum_{\mathcal{I}} v^{(\eta)}_{\mathcal{I}}(\mathbf{r})$, so we may consider $\mathrm{E}_{(\mathbf{0},0)}[v^{(\eta)}_{\mathcal{I}}(\mathbf{R}(t))]$ separately for each index set $\mathcal{I}$. If $\overline{\alpha}_i = 0$ for each $i \in \mathcal{I}$, then $v^{(\eta)}_{\mathcal{I}}(\cdot) \le 1$. So, putting $\alpha_{\mathcal{I}}(\eta) \stackrel{\text{def}}{=} \min_{\overline{\alpha}_i>0,\, i\in\mathcal{I}} \alpha_i(\eta)$, we only need to consider subsets $\mathcal{I}$ such that $\alpha_{\mathcal{I}}(\eta) > 0$.

Let $\widetilde{S}_i(j)$ and $S_i(j)$ be as defined in Section 3.1. To abbreviate the notation, write $S^*_i(m) \stackrel{\text{def}}{=} S_i(J_i(m))$, $\mathbf{S}^*(m) = (S^*_1(m),..,S^*_c(m))$, $\widetilde{S}^*_i(m) \stackrel{\text{def}}{=} \widetilde{S}_i(\widetilde{J}_i(m))$, etc. For $\mathbf{u} \in [0,\infty)^c$, put $\widehat{v}^{(\eta)}_{\mathcal{I}}(\mathbf{u})$ $= \prod_{i\in\mathcal{I}} \mathrm{E}[e^{\alpha_i(\eta)(B_i-u_i)}; B_i > u_i]$ whenever $\min_{i\in\mathcal{I}} u_i \ge \max_{i\notin\mathcal{I}} u_i \ge 0$, otherwise put $\widehat{v}^{(\eta)}_{\mathcal{I}}(\mathbf{u}) = 0$. Then, whenever $\mathcal{I}$ is a strict subset of $\{1,..,c\}$, we have

$$\begin{aligned}\mathrm{E}_{(\mathbf{0},0)}[v^{(\eta)}_{\mathcal{I}}(\mathbf{R}(t))] \quad&=\quad \sum_{m=0}^{\infty} \mathrm{E}_{(\mathbf{0},0)}[\,v^{(\eta)}_{\mathcal{I}}(t - S^*_1(m),..,t-S^*_c(m))\;;\; M_{K(t)} - \#\mathcal{I} = m\,]\\ &\le\quad \mathrm{E}_{(\mathbf{0},0)}\!\left[\,\sum_{m=0}^{\infty} \widehat{v}^{(\eta)}_{\mathcal{I}}(t - S^*_1(m),..,t-S^*_c(m))\,\right]\end{aligned}$$

where $\#\mathcal{I}$ denotes the number of indices in $\mathcal{I}$. For the case $\mathcal{I} = \{1,..,c\}$, simply replace $M_{K(t)} - \#\mathcal{I} = m$ by $M_{K(t)} - c \ge m$ in the first equality above. The upper bound is the same. We obtain a further upper bound by applying the crude bound

$$\widehat{v}_{\mathcal{I}}^{(\eta)}(\mathbf{u}) \quad \le \quad \exp\Big(\#\mathcal{I}\,\Lambda_B(\eta) \;-\; \alpha_{\mathcal{I}}(\eta) \vee (\mathbf{u}) \Big)\; 1_{[0,\infty)}(\vee(\mathbf{u})).$$

Define $U(s) \stackrel{\text{def}}{=} \mathrm{E}_{(\mathbf{0},0)}[$ number of points $\mathbf{S}^*(m) \in [0,s]^c$ $]$. Then the bound is

$$\mathrm{E}_{(\mathbf{0},0)}\Big[v_{\mathcal{I},\epsilon}^{(\eta)}(\mathbf{R}(t)) \Big] \quad \le \quad e^{\#\mathcal{I}\Lambda_B(\eta)} \int_0^t e^{-\alpha_{\mathcal{I}}(\eta)(t-s)}\, dU(s).$$

Integration-by-parts, and then the change of variable $\sigma = t - s$ yields

$$\int_0^t e^{-\alpha_{\mathcal{I}}(\eta)(t-s)}\, dU(s) \quad = \quad e^{-\alpha_{\mathcal{I}}(\eta)t}\, U(t) \;+\; \alpha_{\mathcal{I}}(\eta) \int_0^t [U(t) - U(t-\sigma)]\, e^{-\alpha_{\mathcal{I}}(\eta)\sigma}\, d\sigma.$$

Put $\pi_s(\cdot) = \mathcal{L}_{(\mathbf{0},0)}(\mathbf{R}(s), Q(s), T_{K(s)+1} - s)$. Since $\widetilde{S}_i(j) \le S_i(j)$ (see the proof of Lemma 3.1), we have

$$\begin{aligned}
U(t) \;-\; U(t-\sigma) \quad &= \quad \mathrm{E}_{(\mathbf{0},0)}[\text{ number of points } \mathbf{S}^*(m) \in [0,t]^c \backslash [0,t-\sigma]^c \text{ }] \\
&= \quad \mathrm{E}_{\pi_{t-\sigma}}[\text{ number of points } \mathbf{S}^*(m) \in (0,\sigma]^c \text{ }] \\
&\le \quad \mathrm{E}_{\pi_{t-\sigma}}[\text{ number of points } \widetilde{\mathbf{S}}^*(m) \in (0,\sigma]^c \text{ }] \\
&\le \quad \mathrm{E}_{(\mathbf{0},0)}[\text{ number of points } \widetilde{\mathbf{S}}^*(m) \in [0,\sigma]^c \text{ }] \\
&\stackrel{\text{def}}{=} \quad \widetilde{U}(\sigma).
\end{aligned}$$

Since the $\widetilde{S}_i(j)$s are conditionally independent random walks (given $\mathbf{R}_0^- = \mathbf{0}$), we have $\widetilde{U}(\sigma) \sim \mu\sigma$. Thus,

$$\sup_{t\ge 0} \int_0^t e^{-\alpha_{\mathcal{I}}(\eta)(t-s)}\, dU(s) \quad \le \quad \sup_{t\ge 0} e^{-\alpha_{\mathcal{I}}(\eta)t}\, \widetilde{U}(t) \;+\; \alpha_{\mathcal{I}}(\eta) \int_0^\infty \widetilde{U}(\sigma)\, e^{-\alpha_{\mathcal{I}}(\eta)\sigma}\, d\sigma \quad < \quad \infty,$$

which yields the desired result. ∎

Lemma 4.4: Assume $\lambda < \mu$ and $\{\mathbf{W}_k^-\}$ is positive Harris recurrent. Fix $\eta \in \mathcal{D}$ such that $\lambda(\eta) > \mu(\eta)$. Then $\mathcal{P}_\xi^{(\eta)}(\overline{W} = \infty) > 0$, and $\mathcal{P}_\xi(\overline{W} \ge w) > 0$ for all $w < \infty$.

Proof: Applying Lemmas 3.1 and 3.2 to the η-conjugate process, we conclude that there are only finitely many cycles with W_k^+ drifting off to $+\infty$ during the last cycle a.s. Put $\xi_\ell^{(\eta)}(\cdot) \stackrel{\text{def}}{=} \mathcal{P}_\xi^{(\eta)}(W_{K_\ell}^- \in \cdot\ ;\ K_\ell < \infty)$ = a substochastic measure. Then

$$\sum_{\ell=0}^\infty \mathcal{P}_\xi^{(\eta)}(\, K_\ell < \infty\,,\; K_{\ell+1} = \infty\,) \quad = \quad \sum_{\ell=0}^\infty \mathcal{P}_{\xi_\ell^{(\eta)}}^{(\eta)}(\overline{W} = \infty) \quad = \quad 1,$$

and hence, $\mathcal{P}^{(\eta)}_{\xi^{(\eta)}_\ell}(\overline{W} = \infty) > 0$ for some $\ell < \infty$. Clearly $\xi^{(\eta)}_\ell(\cdot) \ll \xi(\cdot) = \mathcal{P}_\xi(W^+_{K_\ell} \in \ \cdot\ )$. Thus, $\mathcal{P}^{(\eta)}_{\xi^{(\eta)}_\ell}(\overline{W} = \infty) > 0$ for some $\ell < \infty$ implies $\mathcal{P}^{(\eta)}_{\xi}(\overline{W} = \infty) > 0$. The restrictions of $\mathcal{P}^{(\eta)}_{\xi}(\cdot)$ and $\mathcal{P}_\xi(\cdot)$ to the event $\{K^+_w < \infty\}$ are equivalent, so we have $\mathcal{P}_\xi(\overline{W} \geq w) > 0$ for all $w < \infty$. ■

Corollary 4.1: Assume $\lambda < \mu$, $\{\mathbf{W}^-_k\}$ is positive Harris recurrent, and $\lambda(\eta) > \mu(\eta)$ for some $\eta > 0$. Then $\xi(\cdot) = \mathcal{L}_\pi(\mathbf{W}^-_{K_\ell})$ is admissible and (33) holds.

Finally, we turn our attention to the functions $g^{(\theta)}_w(\cdot,\cdot)$ defined by (30).

Lemma 4.5: Assume $0 < \theta < \infty$. Then as $w \uparrow \infty$ we have the pointwise limit $g^{(\theta)}_w(\cdot,\cdot) \uparrow g^{(\theta)}(\cdot,\cdot)$ = a bounded measurable function.

Proof: Define $\widetilde{K}_w \stackrel{\text{def}}{=} \inf\{k\colon \widetilde{W}^-_k \leq -w\}$ and $\widetilde{L}_w \stackrel{\text{def}}{=} \sum_{k=0}^{\widetilde{K}_w - 1} 1_{[0,\infty)}(\widetilde{W}^+_k)$. By translation of the initial condition we have $\mathrm{E}[\, L^+_w \,|\, \widetilde{\boldsymbol{\Delta}}^+_0 = \boldsymbol{\delta}, W^+_0 = w + z] = \mathrm{E}[\, \widetilde{L}_w \,|\, \widetilde{\boldsymbol{\Delta}}^+_0 = \boldsymbol{\delta}, W^+_0 = z]$. Since $\widetilde{K}_w \uparrow \infty$ as $w \uparrow \infty$, we have $\widetilde{L}_w \uparrow \widetilde{L}_\infty = \sum_{k=0}^\infty 1_{[0,\infty)}(\widetilde{W}^+_k)$ a.s. The expectation converges by monotone convergence. Thus, $g^{(\theta)}_w(\cdot,\cdot) \uparrow g^{(\theta)}(\cdot,\cdot)$ pointwise, where

$$g^{(\theta)}(\boldsymbol{\delta}, z) \quad = \quad \mathrm{E}\left[\ \widetilde{L}_\infty \ \Big|\ \widetilde{\boldsymbol{\Delta}}^+_0 = \boldsymbol{\delta}, W^+_0 = z\ \right]\ \exp(\, -\theta z \ - \ \boldsymbol{\alpha}(\theta) \cdot \boldsymbol{\delta}\,). \tag{34}$$

Fix $\eta \in (0,\theta)$. Put $K^*_k = \inf\{\kappa \geq k\colon \widetilde{W}^+_\kappa \geq 0\}$ with $\inf \emptyset = \infty$, and observe that $\{\widetilde{L}_\infty \geq k\} \subset \{K^*_k < \infty\}$. We now use a modified version of the process distributions and the likelihood ratio formula (28). The modification is that we initialize with $\mathbf{W}^+_0$ instead of $\mathbf{W}^-_0$. The resulting modification of (28) is that $\mathbf{W}^-_0$ is replaced by $\mathbf{W}^+_0$, and the term $\Lambda_N(\Lambda_B(\eta))$ is deleted. To obtain an upper bound we may delete the terms $\widetilde{J}_i(M_{K^*_k})[\Lambda_i(\alpha_i(\eta)) - \Lambda_B(\eta)] \leq 0$, and, since $\widetilde{W}^+_{i,K^*_k} \geq \widetilde{W}^+_{K^*_k} \geq 0$, we can also delete the term $-\boldsymbol{\alpha}(\eta) \cdot \widetilde{\mathbf{W}}^+_{K^*_k} \leq 0$. Fix $\eta \in (0,\theta)$, so $\Lambda(\eta) < 0$. Then

$$\begin{aligned}
&\mathcal{P}(\, \widetilde{L}_\infty \geq k \,|\, \widetilde{\boldsymbol{\Delta}}^+_0 = \boldsymbol{\delta}, W^+_0 = z\,) \\
&\qquad \leq \ \ \mathcal{P}(\, K^*_k < \infty \,|\, \widetilde{\boldsymbol{\Delta}}^+_0 = \boldsymbol{\delta}, W^+_0 = z\,) \\
&\qquad \leq \ \ \mathrm{E}^{(\eta, K^*_k)}\Big[\ \exp\Big(\ \boldsymbol{\alpha}(\eta) \cdot \mathbf{W}^+_0 \ + \ K^*_k\, \Lambda(\eta)\ \Big)\ ;\, K^*_k < \infty \ \Big|\ \widetilde{\boldsymbol{\Delta}}^+_0 = \boldsymbol{\delta}, W^+_0 = z\,\Big] \\
&\qquad \leq \ \ \exp\Big(\ \eta z \ + \ \boldsymbol{\alpha}(\eta) \cdot \boldsymbol{\delta} \ + \ \Lambda(\eta)\, k\ \Big)
\end{aligned}$$

where the last bound follows because $K^*_k \geq k$ and $\Lambda(\eta) < 0$, and $W^+_{i,0} = \delta_i + z$. Summing over k, we have

$$\mathrm{E}\Big[\,\widetilde{L}_\infty\,\Big|\,\widetilde{\boldsymbol{\Delta}}_0^+ = \boldsymbol{\delta}, W_0^+ = z\,\Big] \;\le\; \frac{\exp\left(\,\eta z \,+\, \boldsymbol{\alpha}(\eta)\cdot\boldsymbol{\delta}\,\right)}{1 \,-\, e^{\Lambda(\eta)}}.$$

Since $\eta < \theta$ implies $\alpha_i(\eta) \le \alpha_i(\theta)$, we have

$$g^{(\theta)}(\boldsymbol{\delta}, z) \;\le\; \frac{\exp(\,-(\theta-\eta)z \,-\, (\boldsymbol{\alpha}(\theta)-\boldsymbol{\alpha}(\eta))\cdot\boldsymbol{\delta}\,)}{1 \,-\, e^{\Lambda(\eta)}} \;\le\; \frac{1}{1 \,-\, e^{\Lambda(\eta)}} \;<\; \infty \tag{35}$$

for all $\boldsymbol{\delta} \in [0,\infty)^c$ and $z \ge 0$. ■

Corollary 4.2: Assume $\{\mathbf{W}_k^-\}$ is positive Harris recurrent, the $F_i(\cdot)$s are spread out, and the conditions listed in part (i) of Lemma 2.3 hold. Let $\zeta(\cdot)$ be any admissible initial distribution. Then

$$\lim_{w\to\infty} e^{\theta w}\,\mathrm{E}_\zeta[\,L_w^+\,] \;=\; e^{\Lambda_N(\Lambda_B(\theta))}\,\mathrm{E}_\zeta^{(\theta)}\Big[\,v^{(\theta)}(\mathbf{W}_0^-)\,;\,\overline{W}=\infty\,\Big]\;\mathrm{E}_{\psi^+}\Big[\,g^{(\theta)}(\widetilde{\boldsymbol{\Delta}}_\infty^+, V_\infty^+)\,\Big] \tag{36}$$

where $\psi^+(\cdot)$ is the limiting distribution in Lemma 4.2, and moreover, the limit is finite and positive.

At this point we have established the tight $\mathcal{P}_\pi(\,W_k^+ \ge w\,)$ limit of Theorem 2.3 part (i). The constant c_W is just the right side of (36) divided by $\mathrm{E}_\xi[K_1]$. To get the limit for $\mathcal{P}_\zeta(\,\overline{W} \ge w\,)$, simply replace L_w^+ by $1_{\{\overline{W}\ge w\}}$ in the preceding arguments.

Part (ii) of Theorem 2.3 addresses the necessity of the conditions $0 < \theta < \infty$, $\theta \le \overline{\eta}_c$ and $\Lambda(\theta) = 0$. If any of these conditions fail, the claim is that $e^{\theta w}\,\mathcal{P}_\pi(\,W_k^+ \ge w\,) \to 0$, etc. We only give a quick sketch of the proof. The case $\theta = 0$ is trivial. The case $\theta = \infty$ follows from Theorem 2.2. So assume $0 < \theta < \infty$, in which case, by Lemma 4.5, $g_w^{(\theta)}(\cdot,\cdot)$ is a uniformly bounded sequence of functions. If $\Lambda(\theta) < 0$, then the additional likelihood ratio term $\Lambda(\theta)K_w^+$ must be included in the expectation in expression (31). Since $K_w^+ \uparrow \infty$ as $w \uparrow \infty$, the expectation will vanish, by dominated convergence. Likewise, if $\theta > \overline{\eta}_c$, then by Lemma 2.1 we have $\Lambda_i(\alpha_i(\theta)) < \Lambda_B(\theta)$ for at least one i, and the expectation in (31) will vanish due to the additional likelihood ratio terms $J_i(M_{K_w^+})[\Lambda_i(\alpha_i(\theta)) - \Lambda_B(\theta)] \downarrow -\infty$ a.s.

We believe that the conditions $\Lambda_N'(\Lambda_B(\theta)) < \infty$ and $\Lambda_i'(\alpha_i(\theta)) < \infty$ for each i are also necessary, but this is more difficult to prove. Assuming the other conditions hold, one must reconsider the renewal limit for the Markov additive process $\{(\widetilde{\boldsymbol{\Delta}}_k^+, \widetilde{W}_k^+)\}$. If $\Lambda_N'(\Lambda_B(\theta)) = \infty$, then $\mathrm{E}_\pi[\widetilde{W}_{k+1} -^+ \widetilde{W}_k^+] = \infty$. If $\Lambda_i'(\alpha_i(\theta)) < \infty$ for any i, then the Markov chain $\{\widetilde{\boldsymbol{\Delta}}_k^+\}$ is presumably null Harris recurrent. In either case, the limit of Lemma 4.2 should be degenerate. Appendix B provides some further discussion of the a degenerate case.

5. Queue Length Analysis

In this section we consider the proof of the queue length limits of Theorem 2.3. The arguments are almost identical to the workload analysis of the previous section, so we only give a quick sketch.

Define $\widehat{\mathcal{F}}_k \stackrel{\text{def}}{=} \sigma(\mathbf{R}_0^-,\, Q_0^-,\, A_1,..,A_k,\, N_0,..,N_k,\, B_{i,j}\colon\; j \le \widehat{J}_i^+(T_k),\; i = 1,..,c)$ where $\widehat{J}_i^+(t)$ is as defined in Section 3.2. Let K be an $\widehat{\mathcal{F}}_k$-stopping time. Define the (η, K)-conjugate distribution $\mathcal{P}_\zeta^{(\eta,K)}(\cdot)$ exactly as in Section 4, except that we use $\widehat{\mathcal{F}}_k$ in place of $\mathcal{F}_k$. The difference is that in Section 4 service times are assigned to jobs at the instant that they enter the system, while in this section service times are assigned to jobs only when they exit the queue and processing begins. Thus, in Section 4 the service times for jobs in queue at time T_K^+ are samples from the twisted distributions $F_i^{(\alpha_i(\eta))}(\cdot)$, while here they are ultimately sampled from the untwisted distributions $F_i(\cdot)$.

The restrictions of $\mathcal{P}_\zeta^{(\eta,K)}(\cdot)$ and $\mathcal{P}_\zeta(\cdot)$ to the event $\{K < \infty\}$ are equivalent, and the likelihood ratio is

$$\frac{d\mathcal{P}_\zeta}{d\mathcal{P}_\zeta^{(\eta,K)}} \;=\; \exp\Bigg(\eta\, T_K \;+\; K\,\Lambda_A(-\eta) \;-\; \Lambda_B(\eta)\, M_K \;+\; (K+1)\,\Lambda_N(\Lambda_B(\eta)) \\ +\; \sum_{i=1}^{c} \Big[-\alpha_i(\eta)\, (\widetilde{S}_i(\widehat{J}_i^+(T_K)) - R_{i,0}^-) \;+\; \widehat{J}_i^+(T_K)\, \Lambda_i(\alpha_i(\eta)) \Big] \Bigg).$$

By (20), (21) and other definitions in Section 3.2 we have $M_K = \widetilde{Q}_K^+ - Q_0^- + \sum_{i=1}^c \widehat{J}_i^+(T_K)$ and $\widetilde{S}_i(\widehat{J}_i^+(T_K)) = T_K + \widetilde{R}_{i,K}^+$. Using these relationships along with $\sum_{i=1}^c \alpha_i(\eta) = \eta$ and $\Lambda_i(\alpha_i(\eta)) = \Lambda_B(\eta)$ for $\eta \le \overline{\eta}_c$, the above likelihood ratio is reduced and we obtain the following analog of Lemma 4.1.

Lemma 5.1: Assume $\eta \in \mathcal{D} \cap (-\infty, \overline{\eta}_c]$. The restrictions of $\mathcal{P}_\zeta^{(\eta,K)}(\cdot)$ and $\mathcal{P}_\zeta(\cdot)$ to the event $\{K < \infty\}$ are equivalent, and on this event we have

$$\frac{d\mathcal{P}_\zeta}{d\mathcal{P}_\zeta^{(\eta,K)}} \;=\; \exp\Big(-\Lambda_B(\eta)\,(\widetilde{Q}_K^+ - Q_0^-) \;-\; \boldsymbol{\alpha}(\eta)\cdot(\widetilde{\mathbf{R}}_K - \mathbf{R}_0^-) \\ +\; K\,\Lambda(\eta) \;+\; \Lambda_N(\Lambda_B(\eta)) \Big). \tag{37}$$

Define $K_n^+ \stackrel{\text{def}}{=} \inf\{k \ge 1\colon\; \widetilde{Q}_k^+ \ge n\}$ and observe that $\{\overline{Q} \ge n\} = \{K_n^+ < K_1\}$. Also, the random variable L_n^+ in Lemma 3.6 can be written as $L_n^+ = \sum_{k=0}^{K_1-1} 1_{[n,\infty)}(\widetilde{Q}_k^+)$. Define $\widehat{v}^{(\eta)}(\mathbf{r}, q) \stackrel{\text{def}}{=}$

$\exp(\,\Lambda_B(\eta)\,q\;+\;\boldsymbol{\alpha}(\eta)\cdot\mathbf{r}\,)$. Then, Lemma 5.1 yields

$$\begin{aligned}\mathrm{E}_\zeta[\,L_n^+\,] \;&=\; e^{\Lambda_N(\Lambda_B(\eta))}\,\mathrm{E}_\zeta^{(\theta,K_n^+)}\Big[\,L_n^+\,\exp\Big(-\Lambda_B(\theta)\,\widetilde{Q}^+_{K_n^+}\;-\;\boldsymbol{\alpha}(\theta)\cdot\widetilde{\mathbf{R}}_{K_n^+}\,\Big)\\ &\qquad\qquad\qquad\qquad\qquad\qquad \widehat{v}^{(\eta)}(\mathbf{R}_0^-,Q_0^-)\;;\;\overline{Q}\geq n\,\Big]\,.\end{aligned}$$

Define

$$\widehat{g}_n^{(\theta)}(\mathbf{r},q) \;\stackrel{\mathrm{def}}{=}\; \mathrm{E}\Big[\,L_n^+\;\Big|\;\widetilde{\mathbf{R}}_0=\mathbf{r},\widetilde{Q}_0^+=n+q\,\Big]\,\exp(\,-\Lambda_B(\theta)\,q\;-\;\boldsymbol{\alpha}(\theta)\cdot\mathbf{r}\,) \tag{38}$$

where the expectation with respect to the untwisted probability law. Then by conditioning on $\widehat{\mathcal{F}}_{K_n^+}$ we obtain

$$\mathrm{E}_\zeta[\,L_n^+\,] \;=\; e^{\Lambda_N(\Lambda_B(\eta))}\,\mathrm{E}_\zeta^{(\theta)}\Big[\,\widehat{g}_n^{(\theta)}(\widetilde{\mathbf{R}}_{K_n^+},\widetilde{Q}^+_{K_n^+}-n)\,\widehat{v}^{(\eta)}(\mathbf{R}_0^-,Q_0^-)\;;\;\overline{Q}\geq n\,\Big]\,\omega^n \tag{39}$$

where $\omega\stackrel{\mathrm{def}}{=}\exp(-\Lambda_B(\theta))$. Formula (39) is obviously the queue length analog of the waiting time expression (31). As in Section 4, we characterize the limiting behavior of (39) using renewal theory. Lemma 5.2 is proved in Appendices A and C.

Lemma 5.2: Assume the $F_i(\cdot)$s are spread out, $0<\theta<\infty$, $\lambda(\theta)>\mu(\theta)$, $\Lambda_i'(\alpha_i(\theta))<\infty$ for each $i=1,..,c$ and $\Lambda_N'(\Lambda_B(\theta))<\infty$. Then there exists a unique probability distribution $\psi^+(\cdot)$ on $[0,\infty)^c\times\mathbf{Z}_+$ such that the following holds. Let g: $[0,\infty)^c\times\mathbf{Z}_+\to[0,\infty)$ be bounded and measurable, and let $g_n(\cdot,\cdot)$ be uniformly bounded measurable functions such that $g_n(\cdot,\cdot)\to g(\cdot,\cdot)$ pointwise. Let $\zeta(\cdot)$ be a distribution and let $v(\cdot)\geq 0$ be a measurable function on $[0,\infty)^c\times\mathbf{Z}_+$ such that $\mathrm{E}_\zeta[\,v(\mathbf{R}_0^-,Q_0^-)\,]<\infty$. Then

$$\begin{aligned}&\lim_{n\to\infty}\mathrm{E}_\zeta^{(\theta)}\Big[\,g_n(\widetilde{\mathbf{R}}^+_{K_n^+},\widetilde{Q}^+_{K_n^+}-n)\,v(\mathbf{R}_0^-,Q_0^-)\;;\;\overline{Q}\geq n\,\Big]\\ &\qquad=\;\mathrm{E}_{\psi^+}\Big[\,g(\widetilde{\mathbf{R}}^+_\infty,V^+_\infty)\,\Big]\;\mathrm{E}_\zeta^{(\theta)}\Big[\,v(\mathbf{R}_0^-,Q_0^-)\;;\;\overline{Q}=\infty\,\Big]\,.\end{aligned} \tag{40}$$

where $(\widetilde{\mathbf{R}}^+_\infty,V^+_\infty)$ have joint distribution $\psi^+(\cdot)$.

Verification of the conditions of Lemma 5.2 follow almost exactly as in Section 4. Lemmas 4.3 and 4.4 are directly applicable to the queue length analysis. Thus, we find that $\widehat{\xi}(\cdot)$ is admissible. Following the proof of Lemma 4.5, it is straightforward to show that $g_n^{(\theta)}(\cdot,\cdot)\uparrow g^{(\theta)}(\cdot,\cdot)=$ a bounded measurable function.

Appendix A: Markov Additive Processes Renewal Theory

Let Γ denote either $\mathbf{R}$ or $\mathbf{Z}$, and put $\Gamma_+ \stackrel{\text{def}}{=} \Gamma \cap [0, \infty)$. Let $(\mathbf{S}, \mathcal{S})$ be a measurable state space. Let $\{(X_k, Z_k)\}$ be a Markov chain taking values on $\mathbf{S} \times \Gamma$ such that $\mathcal{P}(\,(X_{k+1}, Z_{k+1}) \in \cdot \mid (X_\kappa, Z_\kappa), \kappa \le k)$ $= \mathcal{P}(\,(X_{k+1}, Z_{k+1}) \in \cdot \mid X_k)$. Then $\{(X_k, Y_k)\}$, where $Y_k = Y_0 + \sum_{\kappa=1}^{k} Z_\kappa$, is called a *Markov additive process*. We will work with the following conditions:

(i) There is a function $\overline{f}$: $\Gamma \to [0,1]$ such that $\overline{f}(z)$ is nonincreasing for $z \ge 0$, $\overline{f}(z) = 0$ for $z < 0$, $\mathcal{P}_x(\,|Z_1| \ge z\,) \le \overline{f}(z)$ for all $(x, z) \in \mathbf{S} \times \Gamma_+$ and $c_Z \stackrel{\text{def}}{=} \int_0^\infty \overline{f}(z)\, dz < \infty$. (Notice that $\mathrm{E}_x[\,|Z_1|\,] \le c_Z$ for all $x \in \mathbf{S}$.)

(ii) $\{X_k\}$ is positive Harris recurrent with stationary distribution $\pi(\cdot)$, and $\mathrm{E}_\pi[Z_1] > 0$.

(iii) There is a positive integer $k_0 < \infty$, a $\gamma > 0$, a probability measure $\nu(\cdot)$ on $\mathbf{S} \times \Gamma$, and a set $C \in \mathcal{S}_\pi$ such that

$$\mathcal{P}_{(x,0)}(\,(X_{k_0}, Y_{k_0}) \in \cdot\,) \quad \ge \quad \gamma\, 1_C(x)\, \nu(\cdot). \tag{41}$$

(iv) In the case $\Gamma = \mathbf{R}$, $\nu(dx \times dz) = \widetilde{\nu}(dx)\, (b-a)^{-1}\, 1_{[a,b]}(z)\, dz$ for some $a < b$.

Condition (iii) is the basis for a regeneration scheme for the skeletons $\{(X_{\kappa_0 + \kappa k_0}, Y_{\kappa_0 + \kappa k_0})$: $\kappa = 0, 1, 2 \ldots\}$, where $\kappa_0 = 0, .., k_0 - 1$ is fixed. Briefly, the probability space is augmented with a sequence of 0–1 valued "coin toss" random variables ξ_κ constructed as follows. If $X_{\kappa_0 + \kappa k_0} \notin C$, then $\xi_\kappa = 0$. If $X_{\kappa_0 + \kappa k_0} \in C$, then $\xi_\kappa = 1$ with probability γ, otherwise $\xi_\kappa = 0$. The event $\{\xi_\kappa = 1\}$ is the regeneration event. The regeneration times for the κ_0th skeleton are $K_0(\kappa_0) = 0$ and $K_\ell(\kappa_0)$ $= \inf\{\kappa > K_{\ell-1}(\kappa_0)\colon\ \xi_\kappa = 1\}$ for $\ell \ge 1$. The following Lemma is proved in Ney and Nummelin (1984) and Nummelin (1984).

Lemma A.1: Assume $\{X_k\}$ is Harris recurrent and (iii) holds. Then for fixed $\kappa_0 \in \{0, .., k_0 - 1\}$, the blocks $\Big\{ \Big(X_{\kappa_0 + \kappa k_0}, [Y_{\kappa_0 + \kappa k_0} - Y_{\kappa_0 + K_\ell(\kappa_0) k_0}]\Big) :\ \kappa = K_\ell(\kappa_0) + 1, .., K_{\ell+1}(\kappa_0) \Big\}$, for $\ell \ge 0$, are independent. Moreover, for $\ell \ge 1$ the blocks are i.i.d. with distribution determined by $\mathcal{L}(X_{\kappa_0 + (K_\ell(\kappa_0)+1)k_0}, [Y_{\kappa_0 + (K_\ell(\kappa_0)+1)k_0} - Y_{\kappa_0 + K_\ell(\kappa_0) k_0}]\,) = \nu(\cdot)$.

The basic ideas of the proof of Lemma A.2 below are the same as the proof of Theorem 4.1 of Athreya, et. al. (1978). The much stronger hypotheses (conditions (i) – (iv)) used here allows us to relax the continuity requirement on the function $f(\cdot,\cdot)$. Also, for our application we must work with $k_0 > 1$, which was not worked out by Athreya et. al.

Lemma A.2: Assume conditions (i) – (iv). Let f: $\mathbf{S} \times \Gamma \to [-1,1]$ be a (bounded) measurable function such that $|f(x,z)| \le \overline{f}(z)$ for all $(x,z) \in \mathbf{S} \times \Gamma_+$. Then we have the limit

$$\lim_{y\to\infty} \mathrm{E}_{(x_0,y_0)}\left[\sum_{k=0}^{\infty} f(X_k, y - Y_k) \right] \quad = \quad \frac{\int_{-\infty}^{\infty} \mathrm{E}_\pi[\, f(X,z)\,]\, dz}{\mathrm{E}_\pi[\, Z_1\,]}. \tag{42}$$

Proof: Begin by writing $Y_{\kappa_0+K_\ell(\kappa_0)k_0} = Y_{\kappa_0+K_1(\kappa_0)k_0} + \sum_{m=1}^{\ell-1} \overline{Z}_m$ where $\overline{Z}_m = Y_{\kappa_0+K_{m+1}(\kappa_0)k_0} - Y_{\kappa_0+K_m(\kappa_0)k_0}$. By lemma A.1, $\{\overline{Z}_m\}$ is an i.i.d. sequence with marginal distribution $\mathcal{L}(\overline{Z}_m) = \mathcal{L}_\nu(Y_{K_1(0)k_0})$, which does not depend on (x_0, y_0, κ_0). Define $F_0(y) \stackrel{\text{def}}{=} \mathrm{E}_\nu\Big[\sum_{\kappa=1}^{K_1(0)} f(X_{\kappa k_0}, y - Y_{\kappa k_0}) \Big]$. Then using Lemma A.1 we have

$$\begin{aligned}
&\mathrm{E}_{(x_0,y_0)}\left[\sum_{\kappa=K_\ell(\kappa_0)+1}^{K_{\ell+1}(\kappa_0)} f(X_{\kappa_0+\kappa k_0}, y - Y_{\kappa_0+\kappa k_0}) \right] \\
&\quad = \quad \mathrm{E}_{(x_0,y_0)}\left[F_0\!\left(y - Y_{\kappa_0+K_1(\kappa_0)k_0} - \sum_{m=1}^{\ell-1} \overline{Z}_m \right) \right] \quad = \quad \mathrm{E}\left[F\!\left(y - \sum_{m=1}^{\ell-1} \overline{Z}_m \,; x_0, y_0, \kappa_0 \right) \right]
\end{aligned}$$

where $F(y; x_0, y_0, \kappa_0) = \mathrm{E}_{(x_0,y_0)}[F_0(y - Y_{\kappa_0+K_1(\kappa_0)k_0})]$. Thus,

$$\begin{aligned}
&\mathrm{E}_{(x_0,y_0)}\left[\sum_{\kappa=0}^{\infty} f(X_{\kappa_0+\kappa k_0}, y - Y_{\kappa_0+\kappa k_0}) \right] \\
&\quad = \quad \mathrm{E}_{(x_0,y_0)}\left[\sum_{\kappa=0}^{K_1(\kappa_0)} f(X_{\kappa_0+\kappa k_0}, y - Y_{\kappa_0+\kappa k_0}) \right] \; + \; \mathrm{E}\left[\sum_{\ell=1}^{\infty} F\!\left(y - \overline{Y}_\ell \,; x_0, y_0, \kappa_0 \right) \right]
\end{aligned} \tag{43}$$

where $\overline{Y}_\ell = \sum_{m=1}^{\ell} \overline{Z}_m$. Since $|f(x,z)| \le \overline{f}(z) \le 1$ and $\mathrm{E}_{x_0}[K_1(\kappa_0)] < \infty$ (by positive Harris recurrence), the first term on the right side of (43) will vanish as $y \to \infty$ by dominated convergence. Thus, the limit (42) will follow by application of the i.i.d. key renewal theorem to the second term in (43), and then adding up the k_0 limits for $\kappa_0 = 0, .., k_0 - 1$. We only need to verify the hypothesis of the i.i.d. key renewal theorem.

By positive Harris recurrence we have $\mathrm{E}[\,|\overline{Z}_1|\,] = \mathrm{E}_\nu[\,|Y_{K_1(0)k_0}|\,] \le c_Z\, k_0\, \mathrm{E}_\nu[K_1(0)] < \infty$, and

$\mathrm{E}[\overline{Z}_1] = \mathrm{E}_\nu[Y_{K_1(0)k_0}] = \mathrm{E}_\pi[Z_1]\,k_0\,\mathrm{E}_\nu[K_1(0)] > 0$ because $\mathrm{E}_\pi[Z_1] > 0$. Next, we have

$$\int_{-\infty}^{\infty} |F_0(y)|\,dy \quad \leq \quad \mathrm{E}_\nu\left[\sum_{\kappa=1}^{K_1(0)} \int_{-\infty}^{\infty} |f(X_{\kappa k_0}, y - Y_{\kappa k_0})|\,dy\right] \quad \leq \quad c_z\,\mathrm{E}_\pi[\,K_1(0)\,] \quad < \quad \infty,$$

and hence, $F(\,\cdot\,; x_0, y_0, \kappa_0)$ is also integrable. We only need to show direct Riemann integrability for the case $\Gamma = \mathbf{R}$. By condition (iv), and using the Markov additive process defining property,

$$\mathrm{E}_\nu\left[\sum_{\kappa=1}^{k_1} f(x_\kappa, y - [z_\kappa + Y_{k_0}]) \;\middle|\; X_{\kappa k_0} = x_\kappa,\ Y_{\kappa k_0} - Y_{k_0} = z_\kappa,\ K_1(0) = k_1\right]$$
$$= \int_a^b \sum_{\kappa=1}^{k_1} f(x_\kappa, y - [z_\kappa + z])\,\frac{dz}{b-a}$$

which is a continuous function of y. Taking the expectation, a standard dominated convergence argument yields that $F_0(\cdot)$ is also continuous, and hence, directly Riemann integrable. Likewise, so is $F(\cdot; x_0, y_0, \kappa_0)$. ∎

Lemma A.3: Assume conditions (i) – (iv) and define $K_y \stackrel{\text{def}}{=} \inf\{k : Y_k \geq y\}$. Then there exists a probability distribution $\psi(\cdot)$ on $\mathbf{S} \times \Gamma_+$ such that for all $(x_0, y_0) \in \mathbf{S} \times \mathbf{R}$ we have the limit

$$\lim_{y\to\infty} \mathrm{E}_{(x_0,y_0)}\Big[\,g(X_{K_y}, Y_{K_y} - y)\,\Big] \quad = \quad \mathrm{E}_\psi[\,g(X, Z)\,] \tag{44}$$

whenever g: $\mathbf{S} \times \Gamma_+ \to [0, \infty)$ is a bounded measurable function.

Sketch of Proof: Put $K_y^0 = 0$ and $K_y^{i+1} = \inf\{k > K_y^i\!: \ Y_{k-1} < y;\ Y_k \geq y\}$ with $\inf\emptyset = \infty$. In particular, $K_y = K_y^1$. Also, define $I = \sup\{i : K_y^i < \infty\}$. (We have $I < \infty$ a.s.) Without loss of generality, suppose $\sup_{(x,y)} = g(x, y) = 1$.

For $z > 0$, put $f_a(x, z) = \mathrm{E}_{(x,y-z)}\Big[g(X_1, Y_1 - y); K_y^1 = 1\Big]$ (which does not depend on $y \geq 0$), and for $z \leq 0$ put $f_a(x, z) = 0$. Then, $f_a(x, z) \leq \mathcal{P}_x(Z_1 \geq z) \leq \overline{f}(z)$, and a straightforward computation yields

$$\mathrm{E}_{(x_0,y_0)}\left[\sum_{k=0}^{\infty} f_a(X_k, y - Y_k)\right] \quad = \quad \mathrm{E}_{(x_0,y_0)}\left[\sum_{i=1}^{I} g\Big(X_{K_y^i}, Y_{K_y^i} - y\Big)\right].$$

For $z > 0$, put $f_b(x, z) = \mathrm{E}_{(x,y-z)}\Big[\,g\Big(X_{K_y^2}, Y_{K_y^2} - y\Big)\ ;\ K_y^1 = 1, K_y^2 < \infty\Big]$ (which does not depend on the value of $y \geq 0$), and for $z \leq 0$ put $f_b(x, z) = 0$. Then, $f_b(x, z) \leq \mathcal{P}_x(Z_1 \geq z) \leq \overline{f}(z)$, and a straightforward computation yields

$$\mathrm{E}_{(x_0,y_0)}\left[\sum_{k=0}^{\infty} f_b(X_k, Y_k - y)\right] = \mathrm{E}_{(x_0,y_0)}\left[\sum_{i=2}^{I} g\Big(X_{K_y^i}, Y_{K_y^i} - y\Big) \ ; \ I \geq 2\right].$$

Thus, putting $f(x,z) = f_a(x,z) - f_b(x,z)$, we have

$$\mathrm{E}_{(x_0,y_0)}\left[\sum_{k=0}^{\infty} f(X_k, Y_k - y)\right] = \mathrm{E}_{(x_0,y_0)}[\ g(X_{K_y}, Y_{K_y} - y)\ ]$$

and the limit (44) follows by application of Lemma A.2. ∎

Lemma A.4 addresses the degenerate case. For $g(\cdot,\cdot) \geq 0$, we have $\mathrm{E}_{(x_0,y_0)}[g(X_{K_y}, Y_{K_y} - y)] \leq \mathrm{E}_{(x_0,y_0)}[\sum_{k=1}^{\infty} g(X_k, Y_k - y)]$. Thus, the proof of Lemma A.4 is obtained by reconsidering the proof of Lemma A.2 with $\mathrm{E}[\overline{Z}_\ell] = \infty$. For brevity, we omit this proof.

Lemma A.4: Assume conditions (iii) and (iv), $\{X_k\}$ is Harris recurrent, $\int[-\mathrm{E}_x[Z_1]]^+ \pi(dx) < \infty$, $\int[\mathrm{E}_x[Z_1]]^+ \pi(dx) = \infty$, and $g(\cdot,\cdot) \geq 0$ is such that $\int_{\mathbf{S}} \int_{-\infty}^{\infty} g(x,z)\, dz\, \pi(dx) < \infty$. Then $\mathrm{E}_{(x_0,y_0)}[g(X_{K_y}, Y_{K_y} - y)] \to 0$ as $y \uparrow \infty$.

Appendix B: The Workload Markov Additive Process

In this appendix we prove Lemma 4.2 as an application of Lemma A.3. Notice that since there are no reflections in the "tilde process," we have $\widetilde{\boldsymbol{\Delta}}_k^+ = \widetilde{\boldsymbol{\Delta}}_{k+1}^-$, hence, we may work with either $\{\widetilde{\boldsymbol{\Delta}}_k^-\}$ or $\{\widetilde{\boldsymbol{\Delta}}_k^+\}$.

Lemma B.1: Assume that $F_i(\cdot)$ is spread out. Fix any $\overline{w} \in (0,\infty)$. Then there exists a $\beta > 0$ and a $j_i^* < \infty$ such that for each $j \geq j_i^*$ there are constants $a_i(j) < b_i(j)$ and $\gamma_i(j) > 0$ such that

$$\mathcal{P}_{\mathbf{w}}\Big(\ \widetilde{S}_i(j) \in ds\ ;\ \widetilde{S}_i(j) - \widetilde{S}_i(j-1) > \beta\ \Big) \geq \gamma_i(j)\ 1_{[a_i(j), b_i(j)]}(s)\, ds \tag{45}$$

for all $\mathbf{w} \in [0,\overline{w}]^c$. Moreover, $b_i(j) - a_i(j) = O(j)$.

Proof: Let $F^{*j}(\cdot)$ denote the jth convolution power of $F(\cdot)$. If $F(\cdot)$ is spread out, there exists $\gamma > 0$, $j^\dagger < \infty$ and $a < b$ such that $F^{*j^\dagger}(ds) \geq \gamma\, 1_{[a,b]}(s)\, ds$. For each $j \geq j^\dagger$ we have $a^\dagger(j) < b^\dagger(j)$ and $\gamma(j) > 0$ such that $F^{*j}(ds) \geq \gamma(j)\, 1_{[a^\dagger(j), b^\dagger(j)]}(s)\, ds$. Moreover, by convolution of uniform densities, we can fix the $a^\dagger(j)$s and $b^\dagger(j)$s so that for each $j_0 = 0,..,j^\dagger - 1$ we have $a^\dagger(j_0 + mj^\dagger) \leq$

$a^\dagger(j_0) + (m+1)a^\dagger(j^\dagger)$ and $b^\dagger(j_0 + mj_0) \geq b^\dagger(j_0) + (m-1)b^\dagger(j^\dagger)$ for all $m \geq 1$. Take $j_i^* < \infty$ sufficiently large so that $b_i^\dagger(j-1) - a_i^\dagger(j-1) > \overline{w} + \beta$. The lemma follows by applying the above minorizations to $F_i^{*(j-1)} * F_i^\beta(\cdot)$ where $F_i^\beta(\cdot) = F_i(\cdot \cap (\beta, \infty))$. ∎

Lemma B.2: Assume that each $F_i(\cdot)$ is spread out. Fix any $\overline{\delta} \in (0, \infty)$. There exist a positive integer $k_0 < \infty$, a $\gamma > 0$ and a probability measure $\nu(\cdot)$ on $[0,\infty)^c \times \mathbf{R}$ such that

$$\mathcal{P}\Big(\, (\widetilde{\boldsymbol{\Delta}}^-_{k_0}, \widetilde{W}^-_{k_0}) \, \in \, \cdot \, \Big| \, \widetilde{\boldsymbol{\Delta}}^\pm_0 = \boldsymbol{\delta}, \widetilde{W}^\pm_0 = 0 \, \Big) \quad \geq \quad \gamma \, 1_{[0,\overline{\delta}]^c}(\boldsymbol{\delta}) \, \nu(\cdot) \tag{46}$$

and $\nu(\cdot)$ satisfies condition (iv) in Appendix A.

Proof: Fix $\beta > 0$ such that $F_i((\beta, \infty)) > 0$ for each i. Let the $a_i(j)$s and $b_i(j)$s be as constructed in the proof of Lemma B.1. Then there exists an $m_0 < \infty$ such that for each $m \geq m_0$ there is an integer vector $(j_1, .., j_c)$ such that $j_i \geq j_i^*$ for each i, $\sum_{i=1}^c j_i = m$, and $a(m) \stackrel{\text{def}}{=} \wedge(a_1(j_1), .., a_c(j_c)) < b(m) \stackrel{\text{def}}{=} \vee(b_1(j_1), .., b_c(j_c))$. By taking m_0 sufficiently large, we also have $b(m) - a(m) \geq \beta$. Fix $k_0 < \infty$ so that $\mathcal{P}(M_{k_0 - 1} = m^*) > 0$ for some $m^* \geq m_0$. Let $\mathbf{U}$ be a random vector that is uniformly distributed on the cube $[a(m^*), a(m^*) + \beta]^c$ and is independent of T_{k_0}. Then (46) holds with $\nu(\cdot) = \mathcal{L}((U_1 - \vee(\mathbf{U}), .., U_c - \vee(\mathbf{U}), \vee(\mathbf{U}) - T_{k_0})$ and $\gamma = \mathcal{P}(M_{k_0-1} = m^*) \, \beta^c \, \min_i \gamma_i(j_i)$. ∎

Lemma B.3: Assume that $\mathrm{E}[B_i] < \infty$ for each i. Then for all sufficiently large $\overline{\delta} < \infty$ there exists an $\epsilon > 0$ such

$$\mathrm{E}\Big[\wedge(\widetilde{\boldsymbol{\Delta}}^\pm_c) \, \Big| \, \widetilde{\boldsymbol{\Delta}}^\pm_0 = \boldsymbol{\delta} \Big] \quad \leq \quad \wedge(\boldsymbol{\delta}) \; - \; \epsilon \tag{47}$$

whenever $\boldsymbol{\delta} \notin [0, \overline{\delta}]^c$ and $\vee(\boldsymbol{\delta}) = 0$.

Proof: Define $\widetilde{S}_i^*(m) \stackrel{\text{def}}{=} \widetilde{S}_i(\widetilde{J}_i(m))$ and $\widetilde{\mathbf{S}}^*(m) \stackrel{\text{def}}{=} (\widetilde{S}_1^*(m), .., \widetilde{S}_c^*(m))$. Then $\widetilde{\Delta}^-_{i,k} = \widetilde{S}_i^*(M_{k-1}) - \vee(\widetilde{\mathbf{S}}_i^*(M_{k-1}))$. Also, put $B^*(m) = B_{i,j}$ with $(i,j) = (\widetilde{I}(m), \widetilde{J}_{\widetilde{I}(m)}(m))$.

Let $\widetilde{\mathbf{S}}^*(0) = \boldsymbol{\delta}$ with $\vee(\boldsymbol{\delta}) = 0$. Then we claim that $\vee(\widetilde{\mathbf{S}}^*(c)) \geq \min_{m=1,..,c} B^*(m)$. Let $\delta_{[i]}$, $i = 1, .., c$, denote the components of the vector $\boldsymbol{\delta}$ ordered so that $\delta_{[1]} \leq \delta_{[2]} \leq ... \leq \delta_{[c]}$. After job $m = 1$ is inserted there are two possibilities; either $\vee(\widetilde{\mathbf{S}}^*(1)) = \delta_{[1]} + B^*(1)$, or $\vee(\widetilde{\mathbf{S}}^*(1)) = \delta_{[2]}$. In the first case, the claim holds. So proceed assuming latter case. After job $m = 2$ is inserted, either $\vee(\widetilde{\mathbf{S}}^*(2)) = \delta_{[i]} + B^*(i)$ for $i = 1$ or $i = 2$, or $\vee(\widetilde{\mathbf{S}}^*(2)) = \delta_{[3]}$. Again, the claim holds in the first case, so proceed assuming the second, and so on. After $c - 1$ jobs have been inserted we have either

$\vee(\widetilde{S}_i^*(c-1)) = \delta_{[i]} + B^*(i)$ for some $i \le c-1$ or $\vee(\widetilde{\mathbf{S}}^*(c-1)) = \delta_{[c]}$. In the latter case, insertion of the cth job yields $\vee(\widetilde{\mathbf{S}}^*(c)) = \min_{i=1,..,c}[\delta_{[i]} + B^*(i)] \ge \min_{i=1,..,c} B^*(i)$.

Now, write

$$\begin{aligned} \mathrm{E}\Big[\wedge(\widetilde{\boldsymbol{\Delta}}_c^-) \;\Big|\; \widetilde{\boldsymbol{\Delta}}_0^- = \boldsymbol{\delta} \Big] \;&=\; \mathrm{E}\Big[\wedge(\widetilde{\boldsymbol{\Delta}}_c^-) \,;\, \wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) = \wedge(\boldsymbol{\delta}) \;\Big|\; \widetilde{\mathbf{S}}^*(0) = \boldsymbol{\delta} \Big] \\ &\quad + \; \mathrm{E}\Big[\wedge(\widetilde{\boldsymbol{\Delta}}_c^-) \,;\, \wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) > \wedge(\boldsymbol{\delta}) \;\Big|\; \widetilde{\mathbf{S}}^*(0) = \boldsymbol{\delta} \Big]. \end{aligned} \tag{48}$$

We upper bound the two terms above separately.

To upper bound the first term on the right side of (48), observe that $\wedge(\widetilde{\boldsymbol{\Delta}}_c^-) = \wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) - \vee(\widetilde{\mathbf{S}}^*(M_{c-1}))$, and $\vee(\widetilde{\mathbf{S}}^*(M_{c-1})) \ge \vee(\widetilde{\mathbf{S}}^*(c)) \ge \min_{m=1,..,c} B^*(m)$ because $M_{c-1} \ge c$. Thus, on the event $\{\wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) = \wedge(\boldsymbol{\delta})\}$, we have $\wedge(\widetilde{\boldsymbol{\Delta}}_c^-) \le \wedge(\boldsymbol{\delta}) - \min_{m=1,..,c} B^*(m)$. Define $\underline{G}(b) \stackrel{\text{def}}{=} \min_{i=1,..,c} F_i([b,\infty))$ and $\overline{G}(b) \stackrel{\text{def}}{=} \max_{i=1,..,c} F_i([b,\infty))$ for $b \in [0,\infty)$. By a standard construction, we may construct i.i.d. sequences $\{\underline{B}^*(m)\}$ and $\{\overline{B}^*(m)\}$ such that $\underline{B}^*(m) \le B^*(m) \le \overline{B}^*(m)$ a.s., $\mathcal{P}(\underline{B}^*(m) \ge b) = \underline{G}(b)$ and $\mathcal{P}(\overline{B}^*(m) \ge b) = \overline{G}(b)$. Then $\Big\{\sum_{m=1}^{M_{c-1}} \overline{B}^*(m) < \wedge(\boldsymbol{\delta})\Big\} \subset \Big\{\sum_{m=1}^{M_{c-1}} B^*(m) < \wedge(\boldsymbol{\delta})\Big\} \subset \{\wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) = \wedge(\boldsymbol{\delta})\}$. Thus, for $\boldsymbol{\delta} \notin [0,\overline{\delta}]^c$ we have

$$\begin{aligned} &\mathrm{E}\Big[\min_{m=1,..,c} B^*(m) \,;\, \wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) = \wedge(\boldsymbol{\delta}) \;\Big|\; \widetilde{\mathbf{S}}^*(0) = \boldsymbol{\delta} \Big] \\ &\qquad \ge \; \mathrm{E}\left[\min_{m=1,..,c} \underline{B}^*(m) \,;\, \sum_{m=1}^{M_{c-1}} \overline{B}^*(m) < \overline{\delta} \right] \;\uparrow\; \int_0^\infty \underline{G}(b)^c \, db \end{aligned}$$

as $\overline{\delta} \uparrow \infty$, and hence, when $\boldsymbol{\delta} \notin [0,\overline{\delta}]^c$ with sufficiently large $\overline{\delta} < \infty$, we may upper bound the first term in (48) as

$$\mathrm{E}[\, \wedge(\widetilde{\boldsymbol{\Delta}}_c^-) \,;\, \wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) = \wedge(\boldsymbol{\delta}) \mid \widetilde{\mathbf{S}}^*(0) = \boldsymbol{\delta} \,] \;\le\; \wedge(\boldsymbol{\delta}) \;-\; \frac{1}{2} \int_0^\infty \underline{G}^c(b)\, db. \tag{49}$$

Next consider the second term in (48). Since $\wedge(\boldsymbol{\delta}) = \delta_{[c]} = \delta_{i^*}$ for some $i^* \in \{1,..,c\}$, we have $\{\wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) > \wedge(\boldsymbol{\delta})\} = \{\widetilde{S}_i^*(M_{c-1}) > \delta_{[c]}$ for some $i \ne i^*\}$. The insertion of the mth job can increase $\max_{i \ne i^*} \widetilde{S}_i^*(m)$ by at most $B^*(m)$. Thus, for $\boldsymbol{\delta} \notin [0,\overline{\delta}]^c$ we have $\{\wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) > \wedge(\boldsymbol{\delta})\} \subset \{\sum_{m=1}^{M_{c-1}} B^*(m) \ge \overline{\delta}\}$. When $\wedge(\widetilde{\mathbf{S}}^*(m)) > \wedge(\widetilde{\mathbf{S}}^*(m-1))$ we have $\wedge(\widetilde{\mathbf{S}}^*(m)) - \vee(\widetilde{\mathbf{S}}^*(m)) \le B^*(m)$. Thus, on the event $\{\wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) > \wedge(\boldsymbol{\delta})\}$ we have $\wedge(\widetilde{\boldsymbol{\Delta}}_c^-) \le \max_{m=1,..,M_{c-1}} B^*(m)$, and the second term in (48) is upper bounded by

$$\mathrm{E}\Big[\ \wedge(\widetilde{\boldsymbol{\Delta}}_c^-)\ ;\ \wedge(\widetilde{\mathbf{S}}^*(M_{c-1})) > \wedge(\boldsymbol{\delta})\ \Big|\ \widetilde{\mathbf{S}}^*(0) = \boldsymbol{\delta}\ \Big]$$

$$\leq \mathrm{E}\left[\max_{m=1,..,M_{c-1}} B^*(m)\ ;\ \sum_{m=1}^{M_{c-1}} B^*(m) \geq \overline{\delta}\ \middle|\ \widetilde{\mathbf{S}}^*(0) = \boldsymbol{\delta}\right]$$

$$\leq \mathrm{E}\left[\max_{m=1,..,M_{c-1}} \overline{B}^*(m)\ ;\ \sum_{m=1}^{M_{c-1}} \overline{B}^*(m) \geq \overline{\delta}\right] \tag{50}$$

for all $\boldsymbol{\delta} \notin [0,\overline{\delta}]^c$. (Notice that the last bound does not depend on the initial condition $\boldsymbol{\delta}$.) By dominated convergence the bound (50) will vanish as $\overline{\delta} \uparrow \infty$.

Lemma B.3 follows by applying (49) and (50) to (48) with sufficiently large $\overline{\delta} < \infty$. ■

Lemma B.4: Assume $\mathrm{E}[B_i] < \infty$ and $F_i(\cdot)$ is spread out for each i. Then $\{\widetilde{\boldsymbol{\Delta}}_k^{\pm}\}$ is a positive Harris recurrent Markov chain, and for any $\overline{\delta} \in (0,\infty)$, $[0,\overline{\delta}]^c$ is a small set.

Proof: By Lemma B.2, to establish Harris recurrence it is sufficient to show that $\mathcal{P}_{\mathbf{w}}(\widetilde{\boldsymbol{\Delta}}_k^- \in [0,\overline{\delta}]^c$ Applying Lemma B.3, fix $\overline{\delta} < \infty$ so that (47) holds for some $\epsilon > 0$. Put $K = \inf\{k\text{: } \widetilde{\boldsymbol{\Delta}}_k^- \in [0,\overline{\delta}]^c\}$, and $K^* = \inf\{k = \kappa c\text{: } \widetilde{\boldsymbol{\Delta}}_k^- \in [0,\overline{\delta}]^c\}$. For $\boldsymbol{\delta} \notin [0,\overline{\delta}]^c$, (47) and Theorem 4.2 in Meyn and Tweedie (1992) gives $\mathrm{E}[K|\widetilde{\boldsymbol{\Delta}}_0^- = \boldsymbol{\delta}] \leq \mathrm{E}[K^*|\widetilde{\boldsymbol{\Delta}}_0^- = \boldsymbol{\delta}] \leq c \wedge (\boldsymbol{\delta})/\epsilon$, and this proves Harris recurrence with small set $[0,\overline{\delta}]^c$ and irreducibility measure $\nu(\cdot \times \mathbf{R})$.

Using the same arguments as in the proof of Lemma B.3, for all $\boldsymbol{\delta} \in [0,\overline{\delta}]^c$ we have

$$\mathrm{E}\Big[\ \wedge(\widetilde{\boldsymbol{\Delta}}_c^-)\ \Big|\ \widetilde{\boldsymbol{\Delta}}_0^- = \boldsymbol{\delta}\ \Big] \quad \leq \quad \overline{\delta}\ +\ \mathrm{E}\Big[\max_{m=1,..,M_{c-1}} \overline{B}^*(m)\Big] \quad < \quad \infty.$$

By Proposition 5.10 in Nummelin (1984), $\{\widetilde{\boldsymbol{\Delta}}_{\kappa c}^-\text{: } \kappa = 1,2,...\}$ is positive Harris recurrent. By irreducibility, $\{\widetilde{\boldsymbol{\Delta}}_k^-\}$ is positive Harris recurrent. ■

Lemma B.5: Assume $\mathrm{E}[N] < \infty$, $\mathrm{E}[B_i] < \infty$ and $F_i(\cdot)$ is spread out for each i, and $\lambda > \mu$. Then the Markov additive process $\{(\widetilde{\boldsymbol{\Delta}}_k^{\pm}, \widetilde{W}_k^{\pm})\}$ satisfies conditions (i) through (iv) in Appendix A.

Proof: Put $Z_k^- = \widetilde{W}_k^- - \widetilde{W}_{k-1}^-$, and let $\tilde{\pi}(\cdot)$ denote the stationary distribution of the positive Harris recurrent chain $\{\widetilde{\boldsymbol{\Delta}}_k^-\}$. Clearly, $|Z_1^-| \leq \overline{Z} = A_1 + \sum_{i=1}^{c}\sum_{j=1}^{N_0} B_{i,j}$. Put $\overline{f}(z) = \mathcal{P}(\overline{Z} \geq z)$, so $c_Z = \mathrm{E}[A] + \mathrm{E}[N]\sum_{i=1}^{c}\mathrm{E}[B_i] < \infty$. By and Lemma 3.2 we have $\mathrm{E}_{\mathbf{w}}[\widetilde{W}_k^-] \sim \mathrm{E}_{\tilde{\pi}}[Z_1^-]\,k \sim$

$(\lambda/\mu\, -\, 1)\, \mathrm{E}[A]\, k$, hence, $\mathrm{E}_{\tilde{\pi}}[Z_1^-] = (\lambda/\mu\, -\, 1)\, \mathrm{E}[A] > 0$. Lemmas B.2, B.3 and B.4 verify the other conditions. ∎

Proof of Lemma 4.2: First, recall that $\{K_w^+ < K_1\} = \{\overline{W} \geq w\}$. For $v < w$,

$$\mathrm{E}_{\mathbf{w}}\Big[\, g(\widetilde{\boldsymbol{\Delta}}^+_{K_w^+}, \widetilde{W}^+_{K_w^+} - w)\ ;\ \overline{W} \geq v\, \Big]\ -\ ||g||\, \mathcal{P}_{\mathbf{w}}(v \leq \overline{W} < w)$$

$$\leq\ \mathrm{E}_{\mathbf{w}}\Big[\, g(\widetilde{\boldsymbol{\Delta}}^+_{K_w^+}, \widetilde{W}^+_{K_w^+} - w)\ ;\ \overline{W} \geq w\, \Big]$$

$$\leq\ \mathrm{E}_{\mathbf{w}}\Big[\, g(\widetilde{\boldsymbol{\Delta}}^+_{K_w^+}, \widetilde{W}^+_{K_w^+} - w)\ ;\ \overline{W} \geq v\, \Big]\ +\ ||g||\, \mathcal{P}_{\mathbf{w}}(v \leq \overline{W} < w)$$

where $||g|| = \sup g(\boldsymbol{\delta}, z)$. Successive conditioning on $\mathcal{F}_{K_v^+}$, the dominated convergence theorem and Lemma A.3 yield

$$\lim_{w\to\infty}\ \mathrm{E}_{\mathbf{w}}\Big[\, g(\widetilde{\boldsymbol{\Delta}}^+_{K_w^+}, \widetilde{W}^+_{K_w^+} - w)\ ;\ \overline{W} \geq v\, \Big]$$

$$=\ \mathrm{E}_{\mathbf{w}}\Big[\, \lim_{w\to\infty}\ \mathrm{E}\Big[\, g(\widetilde{\boldsymbol{\Delta}}^+_{K_w^+}, \widetilde{W}^+_{K_w^+} - w)\ |\ \widetilde{\mathbf{W}}^+_{K_v^+}\, \Big]\ ;\ \overline{W} \geq v\, \Big]$$

$$=\ \mathrm{E}_{\psi}[\, g(\widetilde{\boldsymbol{\Delta}}^+_{\infty}, V^+_{\infty})\,]\, \mathcal{P}_{\mathbf{w}}(\overline{W} \geq v).$$

Letting $w \uparrow \infty$, the last two displays yields

$$\liminf_{w\to\infty}\ \mathrm{E}_{\mathbf{w}}\Big[\, g(\widetilde{\boldsymbol{\Delta}}^+_{K_w^+}, W^+_{K_w^+} - w)\ ;\ K_w^+ < K_1\, \Big]$$

$$\geq\ \mathrm{E}_{\psi}[\, g(\widetilde{\boldsymbol{\Delta}}^+_{\infty}, V^+_{\infty})\,]\, \mathcal{P}_{\mathbf{w}}(\overline{W} \geq v)\ \ -\ \ ||g||\, \mathcal{P}_{\mathbf{w}}(v \leq \overline{W} < \infty)$$

Let $v \uparrow \infty$, so $\mathcal{P}_{\mathbf{w}}(\overline{W} \geq v) \downarrow \mathcal{P}_{\mathbf{w}}(\overline{W} = \infty)$ and $\mathcal{P}_{\mathbf{w}}(v \leq \overline{W} < \infty) \downarrow 0$. A similar argument upper bounds the $\limsup$.

Replacing $g(\cdot,\cdot)$ by $g_w(\cdot,\cdot)$ follows by dominated convergence. To obtain Lemma 4.2, simply apply the above arguments to the θ-conjugate distribution. ∎

Finally, we comment on the case that the necessary conditions $0 < \theta < \infty$, $\theta \leq \overline{\eta}_c$ and $\Lambda(\theta) = 0$ hold, but either $\Lambda_N'(\Lambda_B(\theta)) = \infty$, or $\Lambda_i'(\alpha_i(\theta)) = \infty$ for at least one i. In these cases we argue that the degenerate renewal limit of Lemma A.4 is applicable, ultimately yielding the same conclusion as part (ii) of Theorem 2.3. In the case $\mathrm{E}^{(\Lambda_B(\theta))}[N] = \Lambda_N'(\Lambda_B(\theta)) = \infty$, which clearly yields $\mathrm{E}^{(\theta)}_{\boldsymbol{\delta}}[Z_1] = \infty$ for all $\boldsymbol{\delta}$. If $\mathrm{E}^{(\alpha_i(\theta))}[B_i] = \Lambda_i'(\alpha_i(\theta)) = \infty$ for some i, then evidently $\{\widetilde{\boldsymbol{\Delta}}^+_k\}$ is null Harris recurrent with σ-finite invariant measure $\pi(\cdot)$. It must also be shown that $\int\int g^{(\theta)}(\boldsymbol{\delta}, z)\, dz\, \pi(d\boldsymbol{\delta}) < \infty$. This task is reduced by applying the bound (35) developed in the proof of Lemma 4.5. ¿From (35), we easily obtain $\int g^{(\theta)}(\boldsymbol{\delta}, z)\, dz \le C\, e^{-\boldsymbol{\beta}\cdot\boldsymbol{\delta}}$ where $C < \infty$, $\boldsymbol{\beta} = \boldsymbol{\alpha}(\theta) - \boldsymbol{\alpha}(\eta)$ for some $\eta \in (0, \theta)$. Moreover, note that since $0 < \theta \le \overline{\eta}_c$, we have $\beta_i > 0$ for each i. Thus, it would be sufficient to show that $\int e^{-\boldsymbol{\beta}\cdot\boldsymbol{\delta}} \pi(d\boldsymbol{\delta}) < \infty$.

Appendix C: The Queue Length Markov Additive Process

In this Appendix we provide only the key elements of the proof of Lemma 5.2. The full proof follows essentially the same arguments as in Appendix B.

Lemma C.1: Assume that each $F_i(\cdot)$ is spread out and fix $\overline{r} \in (0, \infty)$. Then there exist a positive integer $k_0 < \infty$, a $\gamma > 0$ and a probability measure $\nu(\cdot)$ on $[0, \infty)^c \times \mathbf{Z}$ such that

$$\mathcal{P}\Big(\, (\widetilde{\mathbf{R}}^-_{k_0}, \widetilde{Q}^-_{k_0}) \;\in\; \cdot \mid \widetilde{\mathbf{R}}^-_0 = \mathbf{r}, \widetilde{Q}^-_0 = 0 \,\Big) \quad \ge \quad \gamma\, 1_{[0,\overline{r}]^c}(\mathbf{r})\, \nu(\cdot). \tag{51}$$

Proof: Let the $a_i(j)$s and $b_i(j)$s and β be as constructed in the proof of Lemma B.1 (using $\overline{r}$ in place of $\overline{\delta}$). Fix some $\epsilon > 0$, and for each $k < \infty$, find a $t_0(k) > 0$ such that $\mathcal{P}(T_k \in [t_0(k), t_0(k) + \beta]) > 0$. There exists a $k_0 < \infty$ such that for all $k \ge k_0$ there is a vector $\mathbf{j}(k)$ such that $j_i(k) \ge j_i^*$ (from Lemma B.1) and $[t_0(k), t_0(k) + \beta] \subset [a_i(j_i(k)), b_i(j_i(k)) - \beta]$. Then (51) holds with

$$\nu(\, d\mathbf{r} \times \{q\}\,) \quad = \quad \beta^{-c}\, 1_{[0,\beta]^c}(\mathbf{r})\, d\mathbf{r}\;\; \mathcal{P}\Bigg(\, M_{k_0-1} - \sum_{i=1}^{c} j_i(k_0) \;=\; q \,\Bigg)$$

and $\gamma = \beta^c\, \min_i \gamma_i(j_i)$. ∎

Lemma C.2: For sufficiently large $\overline{r} < \infty$, there exists an $\epsilon > 0$ such

$$\mathrm{E}[\, \wedge(\widetilde{\mathbf{R}}^-_1) \mid \widetilde{\mathbf{R}}^-_0 = \mathbf{r}\,] \quad \le \quad \wedge(\mathbf{r}) \;-\; \epsilon \tag{52}$$

whenever $\mathbf{r} \notin [0, \overline{r}]^c$.

Proof: Given $\{\widetilde{\mathbf{R}}^-_0 = \mathbf{r}\}$, either $\wedge(\widetilde{\mathbf{R}}^-_1) = \wedge(\mathbf{r}) - A_1$, or for some $i \in \{1, .., c\}$ we have $r_i \le A_1$ and $\wedge(\widetilde{\mathbf{R}}^-_1) \le B_{i,j}$ where $j \le \widehat{J}_i(A_1)$. Put $\overline{B}(\mathbf{r}) = \max_{i=1,..,c}\, \max_{j \le \widehat{J}_i(A_1)} B_{i,j}$ on the event $\{\widetilde{\mathbf{R}}^-_0 = \mathbf{r}\}$.

Then

$$\mathrm{E}[\,\wedge(\widetilde{\mathbf{R}}_1^-)\mid \widetilde{\mathbf{R}}_0^- = \mathbf{r}\,] \quad\le\quad \wedge(\mathbf{r}) \quad-\quad \mathrm{E}[\,A_1\,;\,A_1 \le \wedge(\mathbf{r})\,]$$
$$+\;\mathrm{E}[\,\overline{B}(\mathbf{r})\,;\,\overline{B}(\mathbf{r}) \ge \wedge(\mathbf{r}) - A_1\,]. \tag{53}$$

For $\mathbf{r} \notin [0,\overline{r}]^c$, we have $\{A_1 \le \overline{r}\} \subset \{A_1 \le \wedge(\mathbf{r})\}$, and hence, $\mathrm{E}[A_1; A_1 \le \wedge(\mathbf{r})] \ge \mathrm{E}[A_1; A_1 \le \overline{r}] \uparrow \mathrm{E}[A_1]$ as $\overline{r} \uparrow \infty$. Next, we have $\overline{B}(\mathbf{r}) \le \overline{B} = \max_i \max_{j=1,..,N_0} B_{i,j}$, and hence, for $\mathbf{r} \notin [0,\overline{r}]^c$ the last term in the above display is upper bounded by $\mathrm{E}[\overline{B}; \overline{B} \ge \overline{r} - A_1] \downarrow 0$ as $\overline{r} \uparrow \infty$. ■

Lemma C.3: Assume each $F_i(\cdot)$ is spread out. $\{\widetilde{\mathbf{R}}_k^-\}$ is a positive Harris recurrent Markov chain, and for any $\overline{r} > 0$ the set $[0,\overline{r}]^c$ is a small set.

Proof: Lemma C.3 follows form Lemmas C.1 and C.2. The proof is a direct parallel to the proof of Lemma B.4. ■

Lemma C.4: Assume each $F_i(\cdot)$ is spread out and $\lambda > \mu$. Then the Markov additive process $\{(\widetilde{\mathbf{R}}_k^-, \widetilde{Q}_k^-)\}$ satisfies conditions (i) through (iii) in Appendix A.

Proof: Put $\overline{f}(z) = \mathcal{P}\Big(N_0 + \sum_{i=1}^{c} \widehat{J}_i(A_1) \ge z \,\Big|\, \widetilde{\mathbf{R}}_0^- = \mathbf{0}\Big)$ and apply Lemmas C.1, C.2 and C.3. ■

Lemma 5.2 follows from Lemma A.3, the hypothesis having been verified by Lemmas C.1 – C.4. The proof follows exactly the same lines as the proof of Lemma 4.2 given in Appendix B.

Acknowledgment

The authors would like to express their sincere thanks to Sean Meyn (University of Illinois at Urbana-Champaign), Richard Tweedie (Colorado State University) and Steve Lalley (Purdue University). Their valuable advice help proved indispensable in the course of this work.

References

[1] Aldous, D., Hofri, M. and Szpankowski W. (1992), Maximum size of a dynamic data structure: hashing with lazy deletion revisited, *SIAM J. Computing*, **21**, 713-732.

[2] Arjas, E., Nummelin, E. and Tweedie, R. L. (1978), Uniform limit theorems for non-singular renewal and Markov renewal processes, *J. Appl. Prob.* **15**, 112-125.

[3] Athreya, K. B., McDonald, D. and Ney, P. (1978), Limit theorems for simi-Markov processes and renewal theory for Markov chains, *Ann. Probab.* **6**, 788-797.

[4] Bucklew, J. A., Ney, P. and Sadowsky, J. S. (1990), Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains, *J. Appl. Prob.* **27**, 44-59.

[5] Feller, W. (1971), *An Introduction to Probability and its Applications, Vol. II*, John Wiley & Sons, New York.

[6] Iglehart, D. L. (1989), Extreme values in the GI/G/1 queue, *Ann. Math. Statis.* **43**, 627-635.

[7] Meyn, S. P. and Tweedie, R. L. (1992), Stability of Markovian processes I: criteria for discrete time chains, *Adv. Appl. Probab.* **24**, 542-574.

[8] Meyn, S. P. and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer-Verlag, New York.

[9] Neuts, M. and Takahashi, Y. (1981), Asymptotic behavior of the stationary distribution in the $GI/PH/c$ queue with heterogeneous servers, *Z. Wahr.* **57**, 441-452.

[10] Ney, P. and Nummelin, E. (1984), Some limit theorems for Markov additive processes, *Proc. of Symp. on Semi-Markov Models*, Brussels.

[11] Nummelin, E. (1984), *General Irreducible Markov Chains and Non-negative Operators,* Cambridge University Press, Cambridge.

[12] Sadowsky, J. S. (1991), Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/m queue, *IEEE Trans. Auto. Control* **AC-36**, 1383-1394.

[13] Sadowsky, J. S. (1993), On the optimality and stability of exponential twisting in Monte Carlo estimation, *IEEE Trans. Inform. Theory* **IT-39**, 119-128.

[14] Sadowsky, J. S. and Szpankowski, W. (1992), Maximum queue length and waiting time revisited: G/G/c queue, *Probab. in Eng. & Inform. Sci.*, **6**, 157-170.

[15] Takahashi, Y. (1981), Asymptotic exponentiality of the tail of the waiting-time distribution in a $PH/PH/c$ queue, *Adv. Appl. Probab.* **25**, 391-403.