DUE-DATE SCHEDULING: ASYMPTOTIC OPTIMALITY OF GENERALIZED LONGEST QUEUE AND GENERALIZED LARGEST DELAY RULES

JAN A. VAN MIEGHEM

Kellogg School of Management, Northwestern University, Evanston, Illinois 60208-2009, vanmieghem@kellogg.northwestern.edu

Consider the following due-date scheduling problem in a multiclass, acyclic, single-station service system: Any class k job arriving at time t must be served by its due date $t+D_k$. Equivalently, its delay τ_k must not exceed a given delay or lead-time D_k. In a stochastic system, the constraint $\tau_k \leqslant D_k$ must be interpreted in a probabilistic sense. Regardless of the precise probabilistic formulation, however, the associated optimal control problem is intractable with exact analysis. This article proposes a new formulation which incorporates the constraint through a sequence of convex-increasing delay cost functions. This formulation reduces the intractable optimal scheduling problem into one for which the Generalized $c\mu$ (G$c\mu$) scheduling rule is known to be asymptotically optimal. The G$c\mu$ rule simplifies here to a generalized longest queue (GLQ) or generalized largest delay (GLD) rule, which are defined as follows. Let N_k be the number of class k jobs in system, λ_k their arrival rate, and a_k the age of their oldest job in the system. GLQ and GLD are dynamic priority rules, parameterized by θ: GLQ(θ) serves FIFO within class and prioritizes the class with highest index $\theta_k N_k$, whereas GLD(θ) uses index $\theta_k \lambda_k a_k$.

The argument is presented first intuitively, but is followed by a limit analysis that expresses the cost objective in terms of the maximal due-date violation probability. This proves that GLQ(θ_*) and GLD(θ_*), where $\theta_{*,k} = 1/\lambda_k D_k$, asymptotically minimize the probability of maximal due-date violation in heavy traffic. Specifically, they minimize $\liminf_{n\to\infty} \Pr\{\max_k \sup_{s\in[0,t]} \frac{\tau_k(ns)}{n^{1/2}D_k} \geqslant x\}$ for all positive t and x, where $\tau_k(s)$ is the delay of the most recent class k job that arrived before time s. GLQ with appropriate parameter θ_α also reduces "total variability" because it asymptotically minimizes a weighted sum of αth delay moments. Properties of GLQ and GLD, including an expression for their asymptotic delay distributions, are presented.

Received February 2001; revisions received August 2001, September 2001; accepted October 2001.

Subject classifications: Queues, optimization: optimal control to meet due-dates or lead-times. Production/scheduling, sequencing, stochastic: lead-time constrained scheduling. Inventory/production, policies, review/lead-times: production policies to guarantee lead-times.

Area of review: Stochastic Models.

1. INTRODUCTION AND INTUITION

Completing services or manufacturing products by preestablished "due dates" has been—and continues to be—a major concern in many practical settings. This traditional scheduling problem has received much attention in a static or deterministic setting. When services or products are processed over time in a more realistic environment where their processing times, or the times when their requests are received, exhibit some uncertainty, this "simple" scheduling problem becomes intractable with exact analysis. This article solves this important stochastic scheduling problem in the asymptotic regime where scheduling has its highest impact: when the processing system is heavily loaded.

Consider a multiclass, single-station service system without feedback routing. The objective is to schedule the servicing of jobs in this system such that each class k's delay (also called sojourn, throughput, or flow time) τ_k does not exceed a given deterministic delay or "lead-time" D_k. This means that a class k job arriving at time t must be served by its due date $t+D_k$. Henceforth, we will refer to this loosely as the *due-date scheduling problem*. In a stochastic system the delay τ_k is a random variable. Hence, the due-date scheduling objective must be interpreted in a probabilistic sense that specifies the meaning of the constraint violation $\tau_k > D_k$ in a stochastic setting. Such probabilistic interpretations may include enforcing upper bounds on the *violation probabilities* $\Pr\{\tau_k > D_k\}$ or fraction of late jobs, or minimizing some cost functional on $\cup_k\{\tau_k > D_k\}$. For example, minimize a weighted sum of the violation probabilities or a weighted sum of average "tardiness" $(\tau_k - D_k)^+$, etc. Regardless of the precise probabilistic formulation, however, the associated optimal control problem is intractable with exact analysis.

This article proposes a new formulation that incorporates the constraint $\tau_k \leqslant D_k$ through a sequence of convex-increasing delay cost functions and that relates to minimizing the maximal violation probability. This formulation reduces the intractable optimal scheduling problem into one for which the Generalized $c\mu$ (G$c\mu$) scheduling rule introduced by Van Mieghem (1995) is known to be asymptotically optimal. The G$c\mu$ rule for this formulation turns out to be a *generalized longest queue* (GLQ) or *generalized largest delay* (GLD) scheduling rule, which are defined as follows. Let $N_k(t)$ denote the number of class k jobs in the system at time t, λ_k their average arrival rate, and $a_k(t)$ the age of their oldest job in the system at time t. GLQ and GLD are dynamic priority rules, parameterized by a nonnegative vector θ. GLQ(θ) serves FIFO within class and gives priority to the class with highest index $\theta_k N_k(t)$, whereas GLD(θ) uses the dynamic priority index $\theta_k \lambda_k a_k(t)$. The formulation results in a short proof that,

0030-364X/03/5101-0113 $05.00
1526-5463 electronic ISSN

among all work-conserving scheduling rules, GLQ(θ_*) and GLD(θ_*) with parameter

$$\theta_{*,k} = \frac{1}{\lambda_k D_k} \tag{1}$$

asymptotically minimize the maximal violation probability. Note that GLD(θ_*) serves the class with earliest relative deadline $a_k(t)/D_k$.

1.1. Intuition

The intuition behind the approach is rather simple and can be summarized as follow. Consider a new formulation of the due-date problem in terms of a sequence of relaxations that replace the hard constraint $\tau_k \leqslant D_k$ by a minimization of convex-increasing delay cost functions $C_{\alpha,k}$ indexed by $\alpha \in [1, \infty)$:

$$C_{\alpha,k}(\tau_k) = \beta_k \left(\frac{\tau_k}{D_k} \right)^{\alpha}, \tag{2}$$

where β is a positive weight vector. Each convex cost relaxation approximates the due-date problem with accuracy increasing in α, in a sense to be made precise later. The limiting functions $\lim_{\alpha\to\infty} C_{\alpha,k} = C_{*,k}$ represent the "ideal cost formulation" of the due-date problem: The ith class k job with flow time $\tau_{k,i}$ incurs cost $C_{*,k}(\tau_{k,i})$ equal to 0 if $\tau_{k,i} < D_k$; β_k if $\tau_{k,i} = D_k$; and ∞ elsewhere. For finite α, the formulation is well posed and it is then natural to determine a scheduling policy that minimizes the cumulative delay cost $J_\alpha(t)$ over an arbitrary time interval $[0, t]$, in expectation or in distribution, which is the stronger criterion used here. Let $A_k(t)$ denote the number of class k arrivals during $[0, t]$. Then

$$J_\alpha(t) = \sum_{\text{class } k} \sum_{\text{job} i=1}^{A_k(t)} C_{\alpha,k}(\tau_{k,i}) \tag{3}$$

and $J_\alpha(t)$ is asymptotically minimized in distribution at every point in time t by a Generalized $c\mu$ scheduling rule as shown in Van Mieghem (1995). (The precise asymptotic statements will be given later.) Let $1/\mu_k$ denote the average service time of a class k job and $c_{\alpha,k} = \dot{C}_{\alpha,k}$ the marginal delay cost function. The G$c\mu$ rule is a dynamic priority rule that serves FIFO within class and serves the class with highest index $\mu_k c_{\alpha,k}(a_k(t))$. This is precisely the familiar $c\mu$ rule, except that "c" now depends on the system state through the marginal delay cost *function*. The index is easily calculated as $\frac{\mu_k \alpha \beta_k}{D_k}(\frac{a_k(t)}{D_k})^{\alpha-1}$ and is equivalent to using the index $(\frac{\mu_k \beta_k}{D_k})^{1/(\alpha-1)} \frac{a_k(t)}{D_k}$. Hence, the asymptotically optimal rule for the α-relaxation of the due-date problem is GLD with parameter

$$\theta_{\alpha,k} = \left(\frac{\mu_k \beta_k}{D_k} \right)^{1/(\alpha-1)} \theta_{*,k}. \tag{4}$$

In the asymptotic regime, Little's law suggests that the ages a_k and scaled queue count N_k/λ_k have the same distribution (as will be shown rigorously) so that GLQ(θ_α) is also asymptotically optimal. The intuitive argument ends by noting that both policies are well behaved in α and $\lim_{\alpha\to\infty} \theta_\alpha = \theta_*$.

1.2. Motivation

The prime motivation behind this article is to demonstrate the ease and power of the intuitive use of G$c\mu$ rules to optimize nonlinear criteria specified in terms of delays τ and/or queue count N. G$c\mu$ provides a simple and effective tool that yields realistic policies in settings where classical queuing theory is of little help. While G$c\mu$ methodology is grounded in heavy-traffic theory, its greatest strength is that near-optimal rules and performance estimates can be obtained without knowledge of the underlying sophisticated theory. An earlier example was given in Van Mieghem (2000), which embeds G$c\mu$ scheduling in an economic setting to derive optimal quality-of-service offerings and incentive-compatible pricing. Other G$c\mu$-based scheduling rules are derived in Ayhan and Olsen (2000), which also reviews the few papers on scheduling with an objective different from minimizing average holding or delay costs. The lack of such research is surprising, given that service specifications in terms of 95th or 99th percentiles on delay are much more relevant in practice than minimizing the traditional average "holding costs" that are devoid of meaning in a service setting.

While an intuitive argument suffices to advocate the use of GLQ and GLD in practical due-date scheduling settings, a rigorous treatment of the limiting argument for $\alpha \to \infty$ is necessary for a precise characterization of the optimality results. The secondary motivation of this article, then, is to show how G$c\mu$ methodology can deal with various technical complications. Our interest lies in a double limiting regime: a heavy traffic limit and the α-limit. To be robust and meaningful, the results must be valid regardless of the order in which these limits are taken. While establishing such interchange of limits is usually extremely difficult technically, a careful insertion of the α-limit in the main proof of Van Mieghem (1995) suffices. Finally, the limiting argument yields a simpler equivalent characterization of the cost optimality criterion directly in terms of violation probabilities. By showing the asymptotic equivalence of $\lim_\alpha J_\alpha$ and the maximal violation probability, it proves that GLQ(θ_*) and GLD(θ_*) asymptotically minimize the maximal violation probability in the due-date scheduling problem.

The next section in this article details and supports the earlier arguments. In addition, the complete tractability of the asymptotic system allows explicit calculation of all quantities of interest, including asymptotic delay distributions and thus violation probabilities under GLQ and GLD. The third and last section discusses implications and limitations of this approach and reviews related literature.

2. RIGOROUS STATEMENTS AND PROOFS

2.1. Preliminaries on Notation

Notation of Queuing System Primitives. For ease of reference, adopt the notation of Van Mieghem (1995) for a wide class of single-station service systems with d classes that includes the multiclass G/G/1 queuing system. As usual, we are given a d-dimensional arrival process A and an independent d-dimensional service process S. $A_k(t)$ represents the number of class k jobs that have arrived during $[0,t]$ and $S_k(t)$ is the number of class k jobs that are served during the first t time units that the server devotes to class k. Construct d sequences of interarrival times $\{u_{k,i} : i \in \mathbb{N}\}$ for $k = 1, \dots, d$ and a corresponding partial sums process U such that

$$U_k(j) = \sum_{i=1}^{\lfloor j \rfloor} u_{k,i} \quad \text{with} \quad U_k(0) = 0,$$

$$A_k(t) = \max\{j \in \mathbb{N} : U_k(j) \leqslant t\}.$$

$U_k(j)$ is the arrival time of the jth class k job. Similarly, one can construct d sequences of service times $\{v_{k,i} : i \in \mathbb{N}\}$ for $k = 1, \dots, d$ and a corresponding cumulative service process V, where $V_k(j)$ is the total service requirement of the first j class k jobs. As usual, λ_k denotes the average arrival rate of class k and $1/\mu_k$ denotes the average service time of a class k job so that $\rho_k = \lambda_k/\mu_k$ and $\rho = \sum_k \rho_k$ denote the traffic intensity of class k and the system.

A scheduling rule r can be expressed as a vector allocation process T, where $T_k(t)$ represents the total amount of time during $[0,t]$ that the server allocates to class k. Let $N_k(t)$ denote the total number of class k jobs present in the system at time t, and define the vector headcount process N in the obvious way. We have the fundamental flow identity

$$N_k(t) = A_k(t) - S_k(T_k(t)).$$

The total amount of work (expressed in units of time) requested by the class k jobs that are in the system at time t is called the *class k workload* $W_k(t)$ and is defined as

$$W_k(t) = V_k(A_k(t)) - T_k(t). \tag{5}$$

A work-conserving policy is a scheduling rule that provides service whenever the system is not empty. It follows directly that the *total workload* $W_+ = \sum_k W_k$ is independent of the work-conserving scheduling policies. To emphasize the dependence of a quantity on the scheduling rule r, we may add a superscript r. Using that notation, the fact that $\sum_k W_k^r = W_+$ for any work-conserving policy r is called *workload conservation*. For any delay minimization problem that allows preemptive scheduling, it is natural to restrict attention to work-conserving rules because voluntary insertion of idleness increases delays. (In addition, our mode of asymptotic analysis is too crude to differentiate between preemptive and nonpreemptive rules.)

Finally, let $\tau_{k,i}^r$ denote the delay (or flow time or throughput time) of the ith class k job when using scheduling control rule r. The continuous-time process $\tau_k^r(t)$ denotes the delay of the most recent class k job that arrived before time t under control r: $\tau_k^r(t) = \tau_{k,A_k(t)}^r$. Under convex-increasing delay costs, it is optimal to serve FIFO within each class (Van Mieghem 1995, Proposition 1). In that case, the delay process is defined as

$$W_k^r(t) = T_k^r(t + \tau_k^r(t)) - T_k^r(t).$$

Notation of Heavy Traffic. Consider a sequence of queuing systems, parameterized by n, under a policy r. Queuing primitives indexed by superscript n denote quantities in the nth system so that, for example, τ^n and ρ^n denote the delays and the utilization in system n. The usual "heavy-traffic condition" requires that $\lim_{n\to\infty} n^{1/2}(1-\rho^n)$ is finite (so that $\rho^n \to_n 1$). Given that delays τ^n grow unbounded as $\rho^n \to 1$, heavy traffic requires scaling. In addition, to keep cumulative cost finite as $\alpha \to \infty$, denote the scaled cost

$$J_\alpha^{n,r}(t) = \left[\frac{1}{n}\sum_k \sum_{i=1}^{A_k^n(nt)} C_{\alpha,k}\left(\frac{\tau_{k,i}^n}{n^{1/2}}\right)\right]^{1/\alpha}$$
$$= \left[\frac{1}{n}\sum_k \int_0^{nt} C_{\alpha,k}\left(\frac{\tau_k^n(s)}{n^{1/2}}\right) dA_k^n(s)\right]^{1/\alpha}, \tag{6}$$

where the cost functions $C_{\alpha,k}$ are defined by Equation (2). As usual, $X^n \Rightarrow X$ denotes weak convergence of X^n to X in the space $\mathcal{D}$ of simply discontinuous functions under the Skorohod topology. Given that all our limiting processes will be continuous, convergence under the Skorohod metric is equivalent to convergence under the uniform norm $||x||_t = \sup_{s\in[0,t]} |x(s)|$. We will also use the α-norm, denoted by $||x||_{t,\alpha} = [\int_0^t |x(s)|^\alpha ds]^{1/\alpha}$. Let $X^n \simeq Y^n$ stand for $(X^n, Y^n) \Rightarrow (X, X)$. A sequencing rule r^* is said to be *asymptotically optimal* for the α-relaxation of the due-date scheduling problem if its asymptotic cost over an arbitrary time horizon t is stochastically smaller than the cost under any other work-conserving rule r. Thus, r^* minimizes $\liminf_{n\to\infty} \Pr\{J_\alpha^{n,r}(t) \geqslant x\}$ for all $x, t \geqslant 0$. (This is a strong notion of asymptotic optimality that relates to the notion of "pathwise optimality" in heavy traffic.) Its associated cost $J_\alpha^{n,r^*}(t) \Rightarrow J_\alpha^*(t)$, where J_α^* is the finite, tight lower bound on cumulative cost given in Van Mieghem (1995, Proposition 6), so that $\lim_{n\to\infty} \Pr\{J_\alpha^{n,r^*}(t) > x\} = \Pr\{J_\alpha^*(t) > x\}$. Finally, the hallmark heavy-traffic result is that the scaled total workload process converges $n^{-1/2} W_+^n \Rightarrow \widetilde{W}_+^*$. Here, $\widetilde{W}_+^*$ is a reflected Brownian motion that is independent of the work-conserving sequencing rule and whose stationary distribution is exponential and denoted by

$$F_W(x) = \lim_{t\to\infty} \Pr\{\widetilde{W}_+^*(t) \leqslant x\} = 1 - \exp(-\gamma x).$$

For example, Van Mieghem (2000, Proposition 3) gives an expression for γ. In addition, in that paper it is argued that the mixed distribution $1 - \rho\exp(-\gamma x)$, which is equivalent to $F_W(x)$ in heavy traffic, may be a better approximation for the total workload process of a multiclass GI/G/1 queue in moderate traffic.

2.2. Preliminary Properties of GLQ and GLD Policies

Using the G$c\mu$ methodology, we first derive some general performance properties of GLQ and GLD policies that will be useful later for the optimal control problem.

Proposition 1. *Under GLQ(θ) and GLD(θ), the class count process "hugs the curve" $\theta_i N_i = \theta_j N_j$: $\forall\, i, j$:*

$$\sup_{s\in[0,\,t]} n^{-1/2}|\theta_i N_i^n(ns) - \theta_j N_j^n(ns)| \Longrightarrow 0.$$

In addition, under GLD(θ) ages and delays are asymptotically equivalent: $n^{-1/2}\tau_i^{n,GLD} \simeq n^{-1/2} a_i^{n,GLD}$.

The proof is relegated to the Appendix.

Intuitively, Proposition 1 results from the fact that class counts and class workloads "live on a faster time scale" than total workload. Roughly speaking, if class i is being served at time s, its workload changes at rate $n^{-1/2}\dot{W}_i^n(ns) \simeq n^{-1/2}\mu_i^{-1}\dot{N}_i^n(ns) \simeq -n^{1/2}(1-\rho_i^n)$ while other workloads change at rate $n^{-1/2}\dot{W}_{k\neq i}^n \simeq n^{1/2}\rho_k^n$. Thus, in the heavy-traffic limit, class counts can be changed instantaneously while the total workload change $n^{-1/2}\dot{W}_+^n(ns) \simeq -n^{1/2}\ (1-\rho^n)$ remains finite. This allows scheduling rules to distribute the total workload into an arbitrary class workload configuration.

The law of large numbers shows that $\frac{W_i^{n,r}}{n^{1/2}} \simeq \frac{N_i^{n,r}}{n^{1/2}\mu_i}$ (Van Mieghem 1995, Proposition 3) so that the proposition can also be expressed as $\sup_{s\in[0,t]} n^{-1/2}|\theta_i\mu_i W_i^n(ns) - \theta_j\mu_j W_j^n(ns)| \Rightarrow 0$. Workload conservation $\sum_k n^{-1/2} W_k^n \simeq \widetilde{W}_+^*$ now directly shows that $\mu_k\theta_k n^{-1/2} W_k^n \simeq \Theta(\theta)\widetilde{W}_+^*$, where

$$\Theta(\theta) = \left(\sum_k (\mu_k\theta_k)^{-1}\right)^{-1}. \tag{7}$$

Three insights follow. First, the GLQ(θ) and GLD(θ) policies are converging; i.e., each scaled-class workload $n^{-1/2}W_k^n$ has a weak limit: $n^{-1/2}W_k^n \Rightarrow (\mu_k\theta_k)^{-1}\Theta(\theta)\widetilde{W}_+^*$. For such converging policies, an application of Little's law shows that the scaled delay and headcount processes are asymptotically proportional (Van Mieghem 1995, Proposition 5). Together with the earlier workload-headcount equivalence, they satisfy $\frac{W_i^{n,r}}{n^{1/2}\rho_i} \simeq \frac{\tau_i^{n,r}}{n^{1/2}} \simeq \frac{N_i^{n,r}}{n^{1/2}\lambda_i}$. Together with the age and delay equivalence of Proposition 1, this shows that the GLQ(θ) and GLD(θ) are asymptotically equivalent policies in the sense that the scaled performance measures such as W, τ, N, and a, have asymptotically identical distributions under both policies.

Second, the GLQ(θ) and GLD(θ) policies exhibit state space collapse, i.e., class workloads, delays, and headcounts are a deterministic function of total workload only. More specifically, GLQ(θ) and GLD(θ) strive for a *proportional* configuration of the class workload.

Third, Proposition 5 in Van Mieghem (1995) also shows that for converging policies

$$\begin{aligned}[J_\alpha^n(t)]^\alpha &\Longrightarrow \int_0^t \sum_k \lambda_k C_{\alpha,k}\left(\frac{1}{\rho_k\mu_k\theta_k}\Theta(\theta)\widetilde{W}_+^*(s)\right)ds \\ &= \int_0^t \sum_k \lambda_k\beta_k\left(\frac{1}{\lambda_k D_k\theta_k}\Theta(\theta)\widetilde{W}_+^*(s)\right)^\alpha ds.\end{aligned}$$

Summarizing:

Corollary 1. *GLQ(θ) and GLD(θ) are converging policies and exhibit state space collapse, meaning that, under either policy,*

$$n^{-1/2}\mu_k\theta_k W_k^n \simeq n^{-1/2}\theta_k N_k^n \simeq n^{-1/2}\lambda_k\theta_k\tau_k^n \Longrightarrow \Theta(\theta)\widetilde{W}_+^*.$$

In addition, their stationary asymptotic delay distributions are $\lim_{n\to\infty}\Pr\{n^{-1/2}\tau_k^n \leqslant x\} = F_W(\lambda_k\theta_k\Theta^{-1}(\theta)x)$, *and for any* $\alpha \geqslant 1$, *their cost functions also converge*:

$$J_\alpha^n(t) \Longrightarrow \left(\sum_k \frac{\lambda_k\beta_k}{(\lambda_k D_k\theta_k)^\alpha}\right)^{1/\alpha}\Theta(\theta)\left\|\widetilde{W}_+^*\right\|_{t,\,\alpha}. \tag{8}$$

2.3. Asymptotic Optimality of GLQ(θ_α) and GLD(θ_α) for the α-Relaxation of the Due-Date Problem

For any $\alpha \geqslant 1$, minimizing $J_\alpha^{n,r}(t)$ is equivalent to minimizing $(J_\alpha^{n,r}(t))^\alpha$. According to Proposition 8 in Van Mieghem (1995), the latter is accomplished asymptotically by any policy r that satisfies the G$c\mu$ condition, $\forall\, i, j$:

$$\sup_{s\in[0,\,t]}\left|\mu_i c_{\alpha,i}\left(\frac{W_i^{n,r}(ns)}{n^{1/2}\rho_i}\right) - \mu_j c_{\alpha,\,j}\left(\frac{W_j^{n,r}(ns)}{n^{1/2}\rho_j}\right)\right| \Longrightarrow 0. \tag{9}$$

Any such policy r is necessarily converging according to Van Mieghem (1995, Proposition 7). Given that $\frac{W_i^n}{n^{1/2}\rho_i} \simeq \frac{a_i^{n,GLD}}{n^{1/2}} \simeq \frac{N_i^{n,GLQ}}{n^{1/2}\lambda_i}$, the policies which serve the class with highest index $\mu_k c_{\alpha,k}(\frac{N_k}{\lambda_k})$ or $\mu_k c_{\alpha,k}(a_k)$ are not only equivalent in heavy traffic, but also the natural candidates to implement the G$c\mu$ condition (9). Proposition 1 and its corollary show that GLQ(θ_α) and GLD(θ_α) indeed do satisfy that condition. (Recall that the G$c\mu$ index $\mu_k c_{\alpha,k}(\frac{N_k}{\lambda_k}) = \frac{\mu_k\alpha\beta_k}{D_k}(\frac{N_k}{\lambda_k D_k})^{\alpha-1}$ is equivalent to using the GLQ(θ_α) index $(\frac{\mu_k\beta_k}{D_k})^{1/(\alpha-1)}\frac{N_k}{\lambda_k D_k} = \theta_{\alpha,k}N_k$, where $\theta_{\alpha,\,k}$ is defined in Equation (4).)

Instead of using the index representation, one can also use the equivalent workload formulation in Van Mieghem (1995, Equation (43)) because it allows us to calculate the optimal cost. Earlier we said that class workloads live on a faster time scale than total workload, which converges to reflected Brownian motion $\widetilde{W}_+^*$. Thus, in the heavy-traffic limit, sequencing can distribute the total workload into a class workload configuration that minimizes instantaneous cost. This optimal configuration is obtained by a G$c\mu$ rule and is defined by the mapping $g_\alpha : \mathscr{D} \to \mathscr{D}^d : \widetilde{W}_+^* \to \widetilde{W}^* = g_\alpha \circ \widetilde{W}_+^*$ where, for any $s \in [0, t]$:

$$\widetilde{W}^*(s) = \arg\min\left\{\sum_k \lambda_k C_{\alpha,\,k}\left(\frac{x_k}{\rho_k}\right) : x \geqslant 0 \right. \\ \left. \text{and } \sum_k x_k = \widetilde{W}_+^*(s)\right\}.$$

Thus, $\widetilde{W}^*$ is the vector of class workloads that holds $\widetilde{W}_+^*$ in minimal cost fashion at each point in time. Given that g_α is a deterministic function, a G$c\mu$ rule thus always exhibits "optimal" state space collapse. The solution to this convex minimization problem is very simple:

$$\widetilde{W}^*_k(s) = \big[g_\alpha \circ \widetilde{W}^*_+\big]_k(s) = \frac{\Theta(\theta_\alpha)}{\mu_k \theta_{\alpha,k}} \widetilde{W}^*_+(s), \tag{10}$$

and the associated lower bound on cost in Van Mieghem (1995, Proposition 6) simplifies to

$$\begin{aligned}[J^*_\alpha(t)]^\alpha &= \int_0^t \sum_k \lambda_k C_{\alpha,k}\big(\rho_k^{-1}\big[g_\alpha \circ \widetilde{W}^*_+\big]_k(s)\big)\,ds \\ &= \int_0^t \sum_k \lambda_k \beta_k \left(\frac{1}{\lambda_k D_k \theta_{\alpha,k}} \Theta(\theta_\alpha) \widetilde{W}^*_+(s)\right)^\alpha ds,\end{aligned}$$

which is exactly $\lim_{n\to\infty} J^{n,r}_\alpha(t)$, where r is either GLQ(θ_α) and GLD(θ_α). Indeed, the constant factor simplifies to the one in Expression (8):

$$\begin{aligned}\sum_k \frac{\lambda_k \beta_k}{(\lambda_k D_k \theta_{\alpha,k})^\alpha} &= \sum_k \lambda_k \beta_k (D_k/\mu_k\beta_k)^{\alpha/(\alpha-1)} \\ &= \sum_k \mu_k^{-1} \lambda_k D_k (D_k/\mu_k\beta_k)^{1/(\alpha-1)} \\ &= 1/\Theta(\theta_\alpha).\end{aligned}$$

Hence, we have two equivalent arguments for the asymptotic optimality of GLQ(θ_α) and GLD(θ_α). First, Proposition 1 and its corollary show that they satisfy the sufficient G$c\mu$ condition (9) for optimality. Second, explicit calculation of costs show that their asymptotic cost attains the lower bound $J^*_\alpha(t)$. Fix any nonnegative x and t a priori, which we denote throughout this paper by $\forall\, x, t \geqslant 0$. Summarizing then:

Theorem 1. *For any $\alpha \geqslant 1$, GLQ(θ_α) and GLD(θ_α) are asymptotically optimal for the α-relaxation of the due-date scheduling problem*: $\forall\, x, t \geqslant 0$ *they minimize* $\liminf_{n\to\infty} \Pr\{J^{n,r}_\alpha(t) \geqslant x\}$ *over all work-conserving scheduling rules r. The associated asymptotically optimal cost is equal in distribution to*

$$J^*_\alpha(t) = (\Theta(\theta_\alpha))^{\frac{\alpha-1}{\alpha}} \|\widetilde{W}^*_+\|_{t,\alpha}. \tag{11}$$

Remark. The sequence $J^*_\alpha(t)$ is bounded pathwise as a function of $\alpha \geqslant 1$ for any fixed $t \geqslant 0$. Indeed, for any continuous function x, $0 \leqslant ||x||_{t,\alpha} \leqslant t^{1/\alpha}||x||_t \leqslant \max(1, t)||x||_t$. In addition, $0 \leqslant \Theta(\theta_\alpha) \leqslant \min_k(\frac{\mu_k\beta_k}{D_k})^{1/(\alpha-1)}\Theta(\theta_*)$, so that

$$\begin{aligned}0 \leqslant (\Theta(\theta_\alpha))^{\frac{\alpha-1}{\alpha}} &\leqslant \min_k \left(\Theta(\theta_*)\frac{\mu_k\beta_k}{D_k}\right)^{1/\alpha} \Theta(\theta_*) \\ &\leqslant \min_k \left(\max(1, \Theta(\theta_*)\frac{\mu_k\beta_k}{D_k})\right)\Theta(\theta_*) \\ &\overset{\text{def}}{=} M_t/\max(1,\, t).\end{aligned}$$

Hence, $0 \leqslant J^*_\alpha(t) \leqslant M_t \|\widetilde{W}^*_+\|_t$ on each path.

2.4. Asymptotic Optimality of GLQ(θ_*) and GLD(θ_*) for the Due-Date Problem

A first indication that GLQ(θ_*) and GLD(θ_*) are "smart" policies for the due-date problem stems from letting α approach ∞ in Theorem 1. This shows that the sequence of policies r^*_α, where r^*_α is either GLQ(θ_α) or GLD(θ_α), minimize $\lim_\alpha \liminf_n \Pr\{J^{n,r}_\alpha(t) \geqslant x\}$ and for any policy r and $\forall\, x, t \geqslant 0$:

$$\begin{aligned}\lim_{\alpha\to\infty} \liminf_{n\to\infty} \Pr\{J^{n,r}_\alpha(t) \geqslant x\} &\geqslant \lim_{\alpha\to\infty} \lim_{n\to\infty} \Pr\big\{J^{n,\, r^*_\alpha}_\alpha(t) \geqslant x\big\} \\ &= \lim_{\alpha\to\infty} \Pr\{J^*_\alpha(t) \geqslant x\}.\end{aligned} \tag{12}$$

Given that $\lim_{\alpha\to\infty} ||x||_{t,\,\alpha} = ||x||_t$ for continuous $x(\cdot)$ and that $\theta_\alpha \to \theta_*$, so that

$$\lim_{\alpha\to\infty} (\Theta(\theta_\alpha))^{\frac{\alpha-1}{\alpha}} = \Theta(\theta_*) = 1/\sum_k \rho_k D_k, \tag{13}$$

the dominated convergence theorem yields

$$\begin{aligned}\lim_{\alpha\to\infty} \Pr\{J^*_\alpha(t) \geqslant x\} &= \Pr\Big\{\lim_{\alpha\to\infty} J^*_\alpha(t) \geqslant x\Big\} \\ &= \Pr\{J^*_*(t) \geqslant x\},\end{aligned} \tag{14}$$

where we define

$$J^*_*(t) \overset{\text{def}}{=} \Theta(\theta_*) \sup_{s\in[0,\,t]} \widetilde{W}^*_+(s). \tag{15}$$

To show the asymptotic heavy-traffic optimality of the policies r^*_*, where r^*_* is either GLQ(θ_*) or GLD(θ_*), however, we must show that r^*_* also minimizes $\liminf_{n\to\infty} \Pr\{\lim_{\alpha\to\infty} J^{n,r}_\alpha(t) \geqslant x\}$. This is equivalent to asymptotically minimizing maximal violation probabilities because for any rule r, for any sample path, and $\forall\, n$,

$$\begin{aligned}J^{n,\,r}_*(t) &\overset{\text{def}}{=} \lim_{\alpha\to\infty} J^{n,\,r}_\alpha(t) \\ &= \lim_{\alpha\to\infty} \left[n^{-1} \sum_{\text{class}\,k} \sum_{j=1}^{A^n_k(nt)} \beta_k \left(\frac{\tau^{n,r}_{k,j}}{n^{1/2} D_k}\right)^\alpha\right]^{1/\alpha} \\ &= \max_k \sup_{1\leqslant j\leqslant A^n_k(nt)} \left(\frac{\tau^{n,r}_{k,j}}{n^{1/2} D_k}\right) \\ &= \max_k \sup_{s\in[0,\,t]} \frac{\tau^{n,\,r}_k(ns)}{n^{1/2} D_k}.\end{aligned}$$

The proof follows two steps. First, a careful insertion of the α-limit in the main proof of Van Mieghem (1995) establishes:

Proposition 2. *The asymptotic cost $J^{n,\,r}_*$ is stochastically bounded from below by J^*_*: for any rule r and $\forall\, x, t \geqslant 0$:*

$$\begin{aligned}&\liminf_{n\to\infty} \Pr\Big\{\lim_{\alpha\to\infty} J^{n,\,r}_\alpha(t) \geqslant x\Big\} \\ &\quad = \liminf_{n\to\infty} \Pr\left\{\max_k \sup_{s\in[0,\,t]} \frac{\tau^{n,r}_k(ns)}{n^{1/2} D_k} \geqslant x\right\} \geqslant \Pr\{J^*_*(t) \geqslant x\}.\end{aligned}$$

The proof is relegated to the Appendix.

Second, it is easy to show that GLQ(θ_*) and GLD(θ_*) asymptotically attain the lower bound J_*^*. Indeed, the corollary shows that $n^{-1/2}\lambda_k\theta_{*,k}\tau_k^{n,r_*^*} \Rightarrow \Theta(\theta_*)\widetilde{W}_+^*$. Given that $\lambda_k\theta_{*,k} = D_k^{-1}$, the continuous mapping theorem then directly shows that for any class k: $f_k^n(t) \overset{\text{def.}}{=} \sup_{s\in[0,t]} n^{-1/2}D_k^{-1}\tau_k^{n,r_*^*}(ns) \Rightarrow f^*(t) \overset{\text{def.}}{=} \sup_{s\in[0,t]}\Theta(\theta_*)\widetilde{W}_+^*(s)$. Given that for any class k the process $f_k^n(t)$ weakly converges (i.e., under the uniform norm here) to the common limit $f^*(t)$, the process $\max_k f_k^n(t)$ also weakly converges to $f^*(t)$:

$$\lim_{n\to\infty}\Pr\left\{\lim_{\alpha\to\infty} J_\alpha^{n,r_*^*}(t)\geqslant x\right\} = \lim_{n\to\infty}\Pr\left\{\max_k \sup_{s\in[0,t]} \frac{\tau_k^{n,r_*^*}(ns)}{n^{1/2}D_k}\geqslant x\right\}$$
$$=\Pr\{J_*^*(t)\geqslant x\}. \qquad (16)$$

This shows that the two limits can be interchanged and proves:

Theorem 2. *GLQ(θ_*) and GLD(θ_*) are asymptotically optimal for the due-date scheduling problem*: $\forall\, x, t \geqslant 0$ *they minimize, over all work-conserving scheduling rules r,*

$$\liminf_{n\longrightarrow\infty}\Pr\left\{\max_k \sup_{s\in[0,t]} \frac{\tau_k^{n,r}(ns)}{n^{1/2}D_k} \geqslant x\right\}$$
$$= \liminf_{n\longrightarrow\infty}\Pr\left\{\lim_{\alpha\longrightarrow\infty} J_\alpha^{n,r}(t) \geqslant x\right\},$$

and the associated asymptotically optimal maximal due-date violation probability during $[0, t]$ *is*

$$\Pr\left\{\Theta(\theta_*)\sup_{s\in[0,t]}\widetilde{W}_+^*(s)\geqslant x\right\}.$$

Remark. Consistent with a min-max criterion, GLQ(θ_*) and GLD(θ_*) asymptotically minimize the maximal violation probability by equalizing them to $F_W(\sum_k \rho_k D_k)$ in steady state.

3. REVIEW OF RELATED LITERATURE AND DISCUSSION

3.1. Related Literature

In various recent heavy-traffic and large-deviation analyses, a GLQ or GLD rule, or a closely related rule, has emerged. Stolyar and Ramanan (2001) show that GLD, also called *largest weighted delay first*, is asymptotically optimal in a large deviation sense. Their model does not explicitly address bounds D, but shows that GLD with parameter $\theta_k = 1/\beta_k\lambda_k$ asymptotically maximizes the following cost functional for violation probabilities:

$$\min_k\left[\beta_k \lim_{n\longrightarrow\infty}\frac{-\log\Pr\{\tau_k > n\}}{n}\right].$$

Stolyar (2000) gives an impressive generalization of this result to a network setting.

Doytchinov et al. (2001) show that a GLD variant, called *earliest deadline first* (EDF), is asymptotically optimal in heavy traffic in a single-class system with more general stochastic due-date structure. Specifically, each job j has an individual deadline D_j that is drawn from a known distribution function and observed upon its arrival at time t_j. EDF gives dynamic priority to the job with earliest deadline or smallest index $t_j + D_j - t$, which equals $D_j - a_j(t)$ in terms of that job's age $a_j(t)$. In contrast to the multiclass system discussed in this paper, their model restricts attention to a setting where all jobs share a common interarrival and service distribution. Their more general due-date structure leads to EDF, which is related, but not equivalent, to GLQ and GLD. Indeed, EDF prioritizes the class with earliest deadline measured in absolute units by $D_k - a_k$, whereas GLD prioritizes the class with earliest relative deadline a_k/D_k. In heavy traffic, EDF is equivalent to prioritizing the class with largest index $D_k(\theta_{*,k}N_k - 1)$, which is affine, instead of linear, in N_k. (This "shifting of switching curves" for cyclic scheduling rules in the presence of due dates and setup costs is identified and explained by Markowitz and Wein 2001.)

Two recent articles consider admission control in addition to sequencing. Plambeck et al. (2001) show that a sequencing rule, which is exactly GLQ(θ_*), together with dynamic admission control asymptotically minimizes penalties associated with jobs that are rejected when their delay is expected to exceed their delay bound. Specifically, they show that asymptotic violation probabilities essentially vanish under their control, denoted by PKH, in that for any class k and any positive ε and t,

$$\lim_{n\to\infty}\Pr\left\{\sup_{s\in[0,t]} n^{-1/2}\tau_k^{n,\mathrm{PKH}}(ns) > D_k+\varepsilon\right\} = 0.$$

The analysis here shows the asymptotic min-max violation optimality of GLQ by simple G$c\mu$ reasoning without requiring admission control. Theorem 2 is also "intuitively consistent" with the PKH admission rule: It suggests that violation probabilities approach zero if one denies admission to an arbitrary class whenever total workload exceeds $\sum_k \rho_k D_k$ because that signals that a class may violate its due date. Maglaras and Van Mieghem (2001) show asymptotic fluid optimality of GLQ(θ_*) in the sense that the admission region associated with GLQ (i.e., the region of initial conditions for which GLQ will guarantee the delay constraints in fluid scale) is maximized in heavy traffic when $\theta = \theta_*$.

Without attempting an exhaustive literature overview, we mention that Cohen (1987), Sethuraman (1999), and Zipkin (1995) derive exact results for *longest-queue* scheduling, which is GLQ with parameter $\theta_i = 1$. Bertsimas et al. (1998) show that GLQ outperforms generalized processor sharing in a finite buffer system in the sense that GLQ yields lower buffer overflow probabilities under a large deviations criterion. Finally, while all this literature assumes an exogenous due-date structure, it is important not to forget the big message of Wein (1991), that endogenous dynamic due-date setting has a larger impact on performance than due-date-based sequencing policies.

3.2. Discussion

GLQ and GLD are different scheduling rules, as is evident by their different information requirements. They will yield different performance in moderate traffic, including in large-deviation regimes under moderate traffic. There, Stolyar and Ramanan (2000) have showed the necessity of age information so that GLD remains optimal in a large deviations sense, whereas GLQ does not. This article, however, shows that in heavy traffic both rules are *in essence equivalent*. (It also is very likely that in heavy traffic GLQ is also optimal in a large-deviations sense.) The equivalence stems from the fact that in heavy-traffic class delays τ_k and ages a_k have the same distribution as the relative queue lengths N_k/λ_k, so that age and queue-count formulations are equivalent.

The fact that GLQ does *not* require age information is very attractive because of its simplicity and scalability. An additional noteworthy feature of GLQ(θ_*) is *parsimony*: It is independent of the service time distributions and only depends on the first moment of the interarrival time distributions. This parsimony reflects the fact that GLQ is optimal in an "asymptotic" or "first-order" sense that masks fine structural details. In addition, recall that our G$c\mu$-based analysis showed GLQ and GLD optimality in terms of a min-max criterion on violation probabilities. This also shows some limitations of the approach in that it does not apply to all optimality criteria. For example, the G$c\mu$ approach does not apply directly to the minimization of a weighted sum of violation probabilities, which requires non-convex functions $\beta_k 1_{\{\tau_k > D_k\}}$.

Finally, it is worthy to note that the α-relaxation also shows that GLQ(θ_α) reduces *total variability* because it minimizes a weighted sum of αth delay moments. Related variability-minimization issues are also discussed in Ayhan and Olsen (2000).

APPENDIX A. SUMMARY OF KEY RESULTS IN VAN MIEGHEM (1995) THAT WILL BE USED IN THE PROOFS OF PROPOSITION 1 AND 2

As usual in heavy-traffic analysis, we consider a sequence of systems, indexed by n. For this sequence, consider the following expansions:

$$A^n(nt) = n\overline{A}^n(t) + n^{1/2}\widetilde{A}^n(t) + o(n^{1/2}), \tag{17}$$

$$S^n(nt) = n\overline{S}^n(t) + n^{1/2}\widetilde{S}^n(t) + o(n^{1/2}). \tag{18}$$

One may think of the first-and second-order terms as the long-term trend and the variation around this trend, respectively. Because A^n and S^n are nondecreasing, we can always require same of their continuous first-order terms $\overline{A}^n$ and $\overline{S}^n$ so that the inverse functions $\overline{A}^{n^{-1}}$ and $\overline{S}^{n^{-1}}$ exist. Introduce the following functions,

$$R_k^n = \overline{S}_k^{n^{-1}} \circ \overline{A}_k^n \quad \text{and} \quad R_+^n = \sum_k R_k^n.$$

The function R_k^n is the first-order approximation of the work input process $V^n \circ A^n$, so that the nth system operates near full capacity if R_+^n is close to the identity function, which we denote by e.

For a wide class of systems, which includes not only the GI/G/1 system but also some systems with correlated arrival and service processes, there exist processes $\widetilde{A}^*, \widetilde{S}^*, \tilde{c}^*$, with a.s. continuous sample paths on $[0, 1]$ and processes $\overline{A}^*, \overline{S}^*$ with a.s. continuously differentiable increasing sample paths on $[0, 1]$, such that:

$$\left(\overline{A}^n, \widetilde{A}^n, \overline{S}^n, \widetilde{S}^n, n^{1/2}(R_+^n - e)\right) \Longrightarrow \left(\overline{A}^*, \widetilde{A}^*, \overline{S}^*, \widetilde{S}^*, \tilde{c}^*\right).$$

Indeed, weak convergence of counting processes shows that the variation limiting processes $\widetilde{A}^*$ and $\widetilde{S}^*$ are Brownian motions, and the strong law shows that

$$\overline{A}^*(t) = \lambda t, \overline{S}^*(t) = \mu t, R^*(t) = \rho t.$$

Proposition 2 in Van Mieghem (1995) then shows that for any scheduling policy,

$$N^n(nt) = n^{1/2}\widetilde{N}^n(t) + o(n^{1/2}), \tag{19}$$

$$T^n(nt) = n\overline{T}^n(t) + n^{1/2}\widetilde{T}^n(t) + o(n^{1/2}), \tag{20}$$

$$U^n(nt) = n\overline{U}^n(t) + n^{1/2}\widetilde{U}^n(t) + o(n^{1/2}), \tag{21}$$

$$V^n(nt) = n\overline{V}^n(t) + n^{1/2}\widetilde{V}^n(t) + o(n^{1/2}), \tag{22}$$

$$W^n(nt) = n^{1/2}\widetilde{W}^n(t) + o(n^{1/2}), \tag{23}$$

and for FIFO sequencing in each class,

$$\tau^n(nt) = n^{1/2}\tilde{\tau}^n(t) + o(n^{1/2}), \tag{24}$$

with the following convergence relationships:

$$\overline{T}^n \longrightarrow R^* \in \mathscr{C}^1, \tag{25}$$

$$\overline{U}^n \longrightarrow \overline{U}^* = (\overline{A}^*)^{-1} \in \mathscr{C}^1,$$

$$\overline{V}^n \longrightarrow \overline{V}^* = (\overline{S}^*)^{-1} \in \mathscr{C}^1,$$

$$\widetilde{U}^n \longrightarrow \widetilde{U}^* \in \mathscr{C},$$

$$\widetilde{V}^n \longrightarrow \widetilde{V}^* \in \mathscr{C},$$

$$\widetilde{W}_+^n \longrightarrow \widetilde{W}_+^* \in \mathscr{C},$$

$$\widetilde{W}^n \text{ converges } \Longleftrightarrow \widetilde{T}^n \text{converges} \Longleftrightarrow \widetilde{N}^n \text{ converges}$$
$$\Longleftrightarrow \tilde{\tau}^n \text{converges.}$$

The proof of that proposition also shows that a counting process and its associated partial-sums process are (asymptotically) inverse processes:

$$n^{-1}U^n \circ A^n \circ n\, e \longrightarrow e, \tag{26}$$

$$n^{-1}V^n \circ S^n \circ n\, e \longrightarrow e. \tag{27}$$

In addition, it shows that for any class i,

$$\widetilde{N}_i^n(s) = \widetilde{A}_i^*(s) - \widetilde{S}_i^*(\rho_i s) - \lambda_i \widetilde{T}_i^n(s) + o_n(1). \tag{28}$$

It is important to stress that the convergence of the error terms $o_n(1) \to 0$ is *uniform* over $s \in [0, t]$ in all expressions in this appendix.

APPENDIX B. PROOF OF PROPOSITION 1

We give a direct proof using the original G$c\mu$ setup for both GLQ and GLD. Proposition 1 can also be shown via fluid analysis, using recent results by Bramson (1998) and Williams (1998), a track pursued in Plambeck et al. (2001).

B.1. Proof for GLQ

First assume preemptive GLQ(θ). Fix an arbitrary class i and denote

$$\begin{aligned}\delta_i^n(s) &= n^{-1/2}\big(\max_k \theta_k N_k^n(ns) - \theta_i N_i^n(ns)\big)\\ &= \max_k \theta_k \widetilde{N}_k^n(s) - \theta_i \widetilde{N}_i^n(s) \geqslant 0.\end{aligned}$$

Over time, the system fluctuates over three states: Either the system is empty, class i is being served or another class k is being served. During idle periods, δ_i^n is clearly zero. Then, class i being served means $\theta_i N_i^n(s) = \max_k \theta_k N_k^n(s)$, so that again $\delta_i^n = 0$. Now consider a period $(t_1^n, t_2^n) \subset [0, t]$ so that $\delta_i^n(nt_1^n) = \delta_i^n(nt_2^n) = 0$ and $\delta_i^n(n\xi) > 0 \, \forall \, \xi \in (t_1^n, t_2^n)$. Under preemptive GLQ($\theta$), this is a period during which the system processes work of classes other than i. Using a sample path analysis, we now show that $\sup_{\xi \in [t_1^n, t_2^n]} \delta_i^n(n\xi) \to_n 0$ by investigating the dynamics of δ_i^n.

Let the first class that is served after t_1^n be called k_1 and let $t_{k_1}^n$ denote the end of the period that k_1 is being served uninterruptedly. Thus, $\forall \, \xi \in (t_1^n, t_{k_1}^n)$ we have that the time allocation has $\dot{T}_{k_1}^n(n\xi) = 1$ while $\dot{T}_i^n(n\xi) = 0$. Given (20), (19), and (25), we have that

$$\begin{aligned}\widetilde{T}_{k_1}^n(\xi) - \widetilde{T}_{k_1}^n(t_1^n) &= n^{1/2}(\xi - t_1^n) - n^{1/2}\rho_{k_1}(\xi - t_1^n) + o_n(1)\\ &= n^{1/2}(1 - \rho_{k_1})(\xi - t_1^n) + o_n(1),\\ \widetilde{T}_i^n(\xi) - \widetilde{T}_i^n(t_1^n) &= -n^{1/2}\rho_k(\xi - t_1^n) + o_n(1).\end{aligned}$$

Using (19) and denoting the continuous function $\theta_j(\widetilde{A}_j^*(s) - \widetilde{S}_j^*(\rho_j s))$ by $B_j(s)$, we have that (recall that $\delta_i^n(t_1^n) = 0$)

$$\begin{aligned}\delta_i^n(\xi) &= \delta_i^n(\xi) - \delta_i^n(t_1^n),\\ &= \big[B_{k_1}(\xi) - B_{k_1}(t_1^n) - n^{1/2}\lambda_{k_1}\theta_{k_1}(1 - \rho_{k_1})(\xi - t_1^n)\big]\\ &\quad - \big[B_i(\xi) - B_i(t_1^n) + n^{1/2}\lambda_i\theta_i\rho_i(\xi - t_1^n)\big] + o_n(1),\\ &= \big[B_{k_1}(\xi) - B_{k_1}(t_1^n) - (B_i(\xi) - B_i(t_1^n))\big]\\ &\quad - n^{1/2}\big[\lambda_{k_1}\theta_{k_1}(1 - \rho_{k_1}) + \lambda_i\theta_i\rho_i\big](\xi - t_1^n) + o_n(1)\\ &= O(\xi - t_1^n) - n^{1/2}\big[\lambda_{k_1}\theta_{k_1}(1 - \rho_{k_1}) + \lambda_i\theta_i\rho_i\big]\\ &\quad \cdot (\xi - t_1^n) + o_n(1).\end{aligned}$$

Class k_1 being served and $\delta_i^n > 0$ implies $\theta_{k_1} > 0$. Hence, for $\delta_i^n(\xi)$ to be positive, it must be that $(\xi - t_1^n) \leqslant t_{k_1}^n - t_1^n \leqslant o(n^{-1/2})$, so by continuity of the B_j,

$$\delta_i^n(\xi) \leqslant O(t_{k_1}^n - t_1^n) - n^{1/2} o(n^{-1/2}) + o_n(1) = o_n(1).$$

The argument can now be repeated for the mth class, denoted by k_m, that is served during $(t_{k_{m-1}}^n, t_{k_m}^n)$. Stitching all m periods together with the uniform bound $o_n(1)$ yields $\forall \, \xi \in (t_{k_{m-1}}^n, t_{k_m}^n)$:

$$\begin{aligned}\delta_i^n(\xi) = &\Bigg[(B_{k_m}(\xi) - B_{k_m}(t_{k_{m-1}}^n)) + \sum_{j<m}(B_{k_j}(t_{k_j}^n) - B_{k_j}(t_{k_{j-1}}^n))\\ &\quad - (B_i(\xi) - B_i(t_1^n))\Bigg]\\ &\quad - n^{1/2}\Bigg[\lambda_{k_m}\theta_{k_m}(1 - \rho_{k_m})(\xi - t_{k_{m-1}}^n) + \sum_{j<m}\lambda_{k_j}\theta_{k_j}\\ &\quad \cdot (1 - \rho_{k_j})(t_{k_j}^n - t_{k_{j-1}}^n) + \lambda_i\theta_i\rho_i(\xi - t_1^n)\Bigg] + o_n(1).\end{aligned}$$

Again, for $\delta_i^n(\xi)$ to be positive, it must be that $(\xi - t_1^n) \leqslant t_{k_m}^n - t_1^n \leqslant o(n^{-1/2})$. Hence, all periods $(t_{k_j}^n - t_{k_{j-1}}^n) \leqslant \sum_{j<m}(t_{k_j}^n - t_{k_{j-1}}^n) \leqslant o(n^{-1/2})$. The continuity of the B_j then yields that

$$\delta_i^n(\xi) \leqslant O(t_{k_m}^n - t_1^n) - n^{1/2} o(n^{-1/2}) + o_n(1) = o_n(1).$$

In summary, any entire busy cycle (t_1^n, t_2^n) during which class i is not served has length $o(n^{-1/2})$. During such a cycle δ_i^n is $o_n(1)$ so that $\sup_{\xi \in [t_1^n, t_2^n]} \delta_i^n(\xi) \to_n 0$. Together with the fact that δ_i^n equals 0 during idle periods and periods during which i is served, this shows that for any class i : $\sup_{s \in [0,t]} \delta_i^n(ns) \to_n 0$. Finally, recognizing that $\sup_{s \in [0,t]} n^{-1/2}|\theta_i N_i^n(ns) - \theta_j N_j^n(ns)| \leqslant \sup_{s \in [0,t]}(\delta_i^n(s) + \delta_j^n(s))$ and invoking the Skorohod representation theorem as in Van Mieghem (1995, Proposition 8) shows that the convergence holds in distribution.

(It should be noted that a similar reasoning holds for nonpreemptive GLQ: While δ_i^n can now go negative during the last service time of a busy subcycle during which class k_1 is served, the number of class i arrivals during such generic service time v_{k_1} is $O(\lambda_i \mu_{k_1})$. Given the $n^{-1/2}$-scaling of δ_i^n, the potential increase in $N_i^n(s)$ is negligible as $n \to \infty$ and the argument above goes through.) □

B.2. Proof for GLD

First assume preemptive GLD(θ). The proof follows a similar reasoning as for GLQ; thus, we point out only the differences. Define the scaled age $\tilde{a}_j^n(\xi) = n^{-1/2} a_j^n(n\xi)$ and consider now

$$\begin{aligned}\delta_i^n(s) &= n^{-1/2}\big(\max_k \theta_k \lambda_k a_k^n(ns) - \theta_i \lambda_i a_i^n(ns)\big)\\ &= \max_k \theta_k \lambda_k \tilde{a}_k^n(s) - \theta_i \lambda_i \tilde{a}_i^n(s) \geqslant 0.\end{aligned} \tag{29}$$

Again, we will use a sample path analysis to show that $\sup_{\xi \in [t_1^n, t_2^n]} \delta_i^n(n\xi) \to_n 0$ by investigating the dynamics of

δ_i^n. Consider the time intervals as before, but now defined using GLD, where we first assume the nonpreemptive case. At any time, the age of any nonempty class increases linearly in time except for a discontinuity at the departure time for the age of the class that just finished service. Indeed, at such departure time, the age drops by the interarrival time of the new head-of-the-line job of that class. Thus, $\forall\, \xi \in (t_1^n, t_{k_1}^n)$ we have that

$$\begin{aligned}\tilde{a}_{k_1}^n(\xi) &= n^{-1/2}\left[a_{k_1}^n(nt_1^n) + n(\xi - t_1^n) - U_{k_1}^n(m_{k_1}^n)\right]\\ &= \tilde{a}_{k_1}^n(t_1^n) + n^{1/2}(\xi - t_1^n) - n^{-1/2}U_{k_1}^n(m_{k_1}^n), \qquad (30)\\ \tilde{a}_i^n(\xi) &= \tilde{a}_i^n(t_1^n) + n^{1/2}(\xi - t_1^n),\end{aligned}$$

where $U_{k_1}^n(m_{k_1}^n)$ are the interarrival times of the $m_{k_1}^n$ class k_1 jobs that have departed during $[nt_1^n, n\xi)$. Hence, the total service time of these $m_{k_1}^n$ is $n\xi - nt_1^n$ (plus or minus maximally two service times) so that $V_{k_1}^n(m_{k_1}^n) = n\xi - nt_1^n$. Given that the partial-sum process V^n and the counting process S^n are (asymptotically) inverse processes as shown in (27), we have that

$$\begin{aligned}m_{k_1}^n &= S_{k_1}^n(n\xi - nt_1^n) + o_n(1)\\ &= n\mu_{k_1}(\xi - t_1^n) + n^{1/2}\widetilde{S}_{k_1}^*(\xi - t_1^n) + o(n^{1/2}).\end{aligned}$$

And thus, because $\widetilde{U}_{k_1}^*$ is continuous,

$$\begin{aligned}U_{k_1}^n(m_{k_1}^n) &= n\overline{U}_{k_1}^*(\mu_{k_1}(\xi - t_1^n)) + n^{1/2}\widetilde{U}_{k_1}^*(\mu_{k_1}(\xi - t_1^n))\\ &\quad + o(n^{1/2})\\ &= n\frac{\mu_{k_1}}{\lambda_{k_1}}(\xi - t_1^n) + O(n^{1/2}((\xi - t_1^n)).\end{aligned}$$

Plugging into (30), we have that

$$\begin{aligned}\delta_i^n(\xi) = \big(&\theta_{k_1}\lambda_{k_1}\tilde{a}_{k_1}^n(t_1^n) - n^{1/2}\theta_{k_1}\lambda_{k_1}(\rho_{k_1}^{-1} - 1)(\xi - t_1^n)\\ &+ O(\xi - t_1^n) - \theta_i\lambda_i\tilde{a}_i^n(t_1^n) - \theta_i\lambda_i n^{1/2}(\xi - t_1^n)\big).\end{aligned}$$

Given that $\rho_{k_1}^{-1} - 1 \geqslant 0$, we have as before that nonnegativity of $\delta_i^n(\xi)$ requires that $(\xi - t_1^n) = o(n^{-1/2})$ and the same argument can be repeated so that under GLD, any entire busy cycle (t_1^n, t_2^n) during which class i is not served has length $o(n^{-1/2})$. As before, this shows that $\sup_{s\in[0,t]} n^{-1/2}(\theta_k\lambda_k a_k^n(ns) - \theta_i\lambda_i a_i^n(ns)) \Rightarrow 0$. (As earlier, the changes under a nonpreemptive GLD are negligible in heavy traffic.)

Now it only remains to translate the convergence of $\theta_k\lambda_k\tilde{a}_k^n - \theta_i\lambda_i\tilde{a}_i^n$ into a headcount convergence. This is accomplished in two steps:

First, translate convergence of age differences into convergence of flow time differences: During any of the intervals (t_1^n, t_2^n), the flow time of the head-of-the-line class i job equals its age $a_i(nt_1^n)$ plus the length of the entire busy cycle plus its service time. Thus, $n^{-1/2}(\tau_i(nt - a_i(nt)) - a_i(nt)) \leqslant n^{-1/2}o(n^{1/2})$, or, $\tilde{\tau}_i^n(t) - \tilde{a}_i^n(t) + o_n(1) \leqslant o_n(1)$, which formally proves that the scaled difference between flow times and age of head-of-the-line job is negligible: $\tilde{\tau}_i^n(t) - \tilde{a}_i^n(t) \Rightarrow 0$. Thus, $\sup_{s\in[0,t]} n^{-1/2}(\theta_k\lambda_k\tau_k^n(ns) - \theta_i\lambda_i\tau_i^n(ns)) \Rightarrow 0$.

Second and finally, translate convergence of flow time differences into convergence of headcount difference by applying the generalized Little's Law (see Van Mieghem 1995, Proposition 4) to any pair of classes k and i. This shows that $\sup_{s\in[0,t]} n^{-1/2}(\theta_k N_k^n(ns) - \theta_i N_i^n(ns)) \Rightarrow 0$. □

APPENDIX C. PROOF OF PROPOSITION 2

This is shown by inserting a $\lim_\alpha$ argument inside the proof of Proposition 6 of Van Mieghem (1995), which holds for any $\alpha \geqslant 1$. The argument goes as follows. Fix $\epsilon > 0$ and, for any $n \in \mathbb{N}$, consider the sequence of stopping times of $\widetilde{W}_+^*$, $\{t_i : i \in \mathbb{N}\}$, defined as follows:

$$\begin{aligned}t_1 &= \min\Big\{1, \inf\Big\{0 < t \leqslant 1 : \big|\widetilde{W}_+^*(t) - \lfloor\widetilde{W}_+^*(0)/\epsilon\rfloor\epsilon\big| \geqslant \epsilon\Big\}\Big\}\\ t_{i+1} &= \min\Big\{1, \inf\Big\{t_i < t \leqslant 1 : \big|\widetilde{W}_+^*(t) - \widetilde{W}_+^*(t_i)\big| \geqslant \epsilon\Big\}\Big\}.\end{aligned}$$

Thus t_{i+1} is the first time $\widetilde{W}_+^*$ changes by ϵ starting from $\widetilde{W}_+^*(t_i)$ at time t_i. Because $\widetilde{W}_+^*$ is continuous, $\sup_i(t_{i+1} - t_i) \to 0$ as $\epsilon \to 0$, so that $\sup_i(t_{i+1} - t_i) = O(\epsilon)$. Using the fact that $\widetilde{W}_+^n \to \widetilde{W}_+^*$ and the construction of the stopping times t_i, we have that

$$(t_{i+1} - t_i)^{-1}\int_{t_i}^{t_{i+1}} \widetilde{W}_+^n(t)\mathrm{d}t = \widetilde{W}_+^*(t_i) + O(\epsilon) + o_n(1). \qquad (31)$$

Pick up the proof at the end of Page 830, which shows that (recall that $\widetilde{J}_\alpha^n = (J_\alpha^{n,r})^\alpha$ and simplify notation by assuming that λ and μ are constants)

$$\begin{aligned}(J_\alpha^{n,r})^\alpha \geqslant \sum_k\sum_i &\lambda_k(t_{i+1} - t_i)\\ &\cdot C_{\alpha,k}\left(\rho_k^{-1}(t_{i+1} - t_i)^{-1}\int_{t_i}^{t_{i+1}} \widetilde{W}_k^n dt + \epsilon_i + \eta_{i,k}\right),\end{aligned}$$

where we use the more detailed error notation ϵ_i and $\eta_{i,k}$ to signify that the errors depend on the index i and/or k. All errors ϵ_i and $\eta_{i,k}$ are independent of α and are uniformly bounded by $O(\epsilon)$ and $o_n(1)$, respectively. Now invoke the mapping g_α and (31) to get:

$$\begin{aligned}(J_\alpha^{n,r})^\alpha &\geqslant \sum_k\sum_i \lambda_k(t_{i+1} - t_i)\\ &\qquad \cdot C_{\alpha,k}\big(\rho_k^{-1}[g_\alpha \circ \widetilde{W}_+^*]_k(t_i) + \epsilon_i + \eta_{i,k}\big)\\ &= \sum_k\sum_i \lambda_k(t_{i+1} - t_i)\beta_k\\ &\qquad \cdot \left(\rho_k^{-1}\frac{1}{\mu_k D_k\theta_{\alpha,k}}\Theta(\theta_\alpha)\widetilde{W}_+^*(t_i) + \frac{\epsilon_i + \eta_{i,k}}{D_k}\right)^\alpha.\end{aligned}$$

Now raise both sides to the power $1/\alpha$ and take the limit for $\alpha \to \infty$. Given that $\lim_\alpha(\lambda_k(t_{i+1} - t_i)\beta_k)^{1/\alpha} = 1$ and

$\lim_\alpha \theta_\alpha = \theta_*$ and $\lim_\alpha \Theta(\theta_\alpha) = \Theta(\theta_*)$, this yields

$$\begin{aligned} J_*^{n,r} &= \lim_{\alpha\to\infty} J_\alpha^{n,r} \\ &\geqslant \sup_i \max_k \left[\rho_k^{-1} \frac{1}{\mu_k D_k \theta_{*,k}} \Theta(\theta_*) \widetilde{W}_+^*(t_i) + \frac{\epsilon_i + \eta_{i,k}}{D_k} \right]. \end{aligned}$$

Take $\liminf_{n\to\infty}$ and recall that $\theta_{*k}^{-1} = \lambda_k D_k$ and all errors $\eta_{i,k}$ are bounded uniformly by $o_n(1)$:

$$\liminf_{n\to\infty} J_*^{n,r} \geqslant \sup_i [\Theta(\theta_*) \widetilde{W}_+^*(t_i) + \epsilon_i / D_k],$$

where the left-hand side is independent of ϵ. Therefore, letting $\epsilon \to 0$ so that $\sup \epsilon_i \to 0$ and $\sup_i (t_{i+1} - t_i) \to 0$ because the bounds $O(\epsilon)$ is uniform and $\widetilde{W}_+^*$ is continuous; this also implies that $\sup_i \widetilde{W}_+^*(t_i) \to \sup_{s\in[0,t]} \widetilde{W}_+^*(s)$ so:

$$\liminf_{n\longrightarrow\infty} J_*^{n,r} \geqslant \sup_{s\in[0,t]} \left[\Theta(\theta_*) \widetilde{W}_+^*(s) \right].$$

Finally, invoking the Skorohod representation theorem shows that this inequality holds in distribution. □

ACKNOWLEDGMENTS

The author is grateful to Costis Maglaras for constructive feedback. The author also benefited from the comments of Jim Dai, Erica Plambeck, Sasha Stolyar, and the anonymous reviewers on earlier versions.

REFERENCES

Ayhan H., T. L. Olsen. 2000. Scheduling of a multi-class single-server queue under non-traditional performance measures. *Oper. Res.* **48**(3) 482–289.

Bertsimas, D., I. C. Paschalidis, J. N. Tsitsiklis. 1998. Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach. *IEEE Trans. Automatic Control* **43**(3) 315–335.

Bramson, M. 1998. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* **30** 89–148.

Cohen, J. 1987. A two queue, one server model with priority for the longer queue. *Queueing Systems* **2** 261–284.

Doytchinov, B., J. Lehoczky, S. Shreve. 2001. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.* **11**(2) 332–378.

Maglaras, C., J. A. Van Mieghem. 2001. Queueing systems with leadtime constraints: A fluid-model approach for admission and sequencing control. Working paper, Northwestern University, Evanston, IL.

Markowitz, D. M., L. M. Wein. 2001. Heavy traffic analysis of dynamic cyclic policies: A unified treatment of the single machine scheduling problem. *Oper. Res.* **49**(2) 246–270.

Plambeck, E., S. Kumar, J. M. Harrison. 2001. Asymptotic optimality of a single server queueing system with constraints on throughput times. *Queueing Systems* **39** 23–54.

Sethuraman, J. 1999. Scheduling job shops and multiclass queueing networks using fluid and semidefinite relaxations. Ph.D. thesis, MIT, Cambridge, MA.

Stolyar, A. 2002. Control of end-to-end delay tails in a multiclass network: LWDF discipline optimality. *Ann. Appl. Probab.* Forthcoming.

——, K. Ramanan. 2001. Largest weighted delay first scheduling: Large deviations and optimality. *Ann. Appl. Probab.* **11**(1) 1–48.

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. Appl. Probab.* **5**(3) 809–833.

——. 2000. Price and service discrimination in queueing systems: Incentive compatibility of G$c\mu$ scheduling. *Management Sci.* **46**(9) 1249–1267.

Wein, L. M. 1991. Due-date setting and priority sequencing in a multiclass M/G/1 queue. *Management Sci.* **37**(7) 834–850.

Williams, R. 1998. Diffusion approximations for open multiclass queuing networks: Sufficient conditions involving state space collapse. *Queueing Systems* **30** 27–88.

Zipkin, P. H. 1995. Performance analysis of a multi-item production-inventory system. *Management. Sci.* **41**(4) 690–703.