PARALLEL SCHEDULING OF MULTICLASS $M/M/m$ QUEUES: APPROXIMATE AND HEAVY-TRAFFIC OPTIMIZATION OF ACHIEVABLE PERFORMANCE

KEVIN D. GLAZEBROOK

Department of Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK; kevin.glazebrook@newcastle.ac.uk

JOSÉ NIÑO-MORA

Department of Economics and Business, Universitat Pompeu Fabra, E-08005, Barcelona, Spain; jose.nino-mora@econ.upf.es

(Received December 1996; revisions received January 1999, November 1999; accepted April 2000)

We address the problem of scheduling a multiclass $M/M/m$ queue with Bernoulli feedback on m parallel servers to minimize time-average linear holding costs. We analyze the performance of a heuristic priority-index rule, which extends Klimov's optimal solution to the single-server case: servers select preemptively customers with larger Klimov indices. We present closed-form suboptimality bounds (*approximate optimality*) for Klimov's rule, which imply that its suboptimality gap is uniformly bounded above with respect to (i) external arrival rates, as long as they stay within system capacity; and (ii) the number of servers. It follows that its *relative* suboptimality gap vanishes in a heavy-traffic limit, as external arrival rates approach system capacity (*heavy-traffic optimality*). We obtain simpler expressions for the special no-feedback case, where the heuristic reduces to the classical $c\mu$ rule. Our analysis is based on comparing the expected cost of Klimov's rule to the value of a strong linear programming (LP) relaxation of the system's region of achievable performance of mean queue lengths. In order to obtain this relaxation, we derive and exploit a new set of *work decomposition laws* for the parallel-server system. We further report on the results of a computational study on the quality of the $c\mu$ rule for parallel scheduling.

1. INTRODUCTION

Can we match the performance of a *fast* processor (with speed m) with a set of m *slow* parallel processors (with speed 1)? Clearly not, because of the inefficiencies inherent in parallel processing: The parallel system's total processing rate will fall below m when there are fewer than m jobs available. How close, then, can we get to matching the performance of the fast processor with the corresponding set of slow processors, and how should we schedule the parallel system to achieve its best performance? These issues are significant in the design and operation of complex service systems, such as flexible manufacturing systems and computer communication networks. In this paper we address such problems in the idealized setting of a versatile service system model: a multiclass $M/M/m$ queue with Bernoulli feedback.

We shall thus consider the problem of allocating dynamically m identical servers to customers in an n-class $M/M/m$ queueing network to minimize a performance objective $c_1E_u[L_1]+\cdots+c_nE_u[L_n]$ of expected linear holding costs, where $E_u[L_j]$ represents the steady-state expected number of class j customers in the system under policy u, and $c_j \geqslant 0$ their holding cost rate. Admissible scheduling policies make history-dependent decisions, allow customer preemptions, and are nonidling (no server can lie idle when there are customers waiting). Consider now the corresponding problem in which the m *slow* parallel servers (when $m \geqslant 2$) are replaced by a *pooled resource* consisting of one *fast* m-fold speed single server. While the parallel-server optimal scheduling problem is likely to be computationally intractable, the solution for the pooled resource constitutes a classical result in the field of stochastic scheduling: Klimov (1974, 1978) showed that the optimal policy is characterized by class-dependent priority-indices $\gamma_1, \ldots \gamma_n$, efficiently computed by an *adaptive greedy algorithm*, so it is optimal to give at each decision epoch higher service priority to a customer with larger index. Clearly, Klimov's rule extends naturally into a simple heuristic for the parallel-server system: At each decision epoch, let servers select preemptively available customers with larger indices. The current paper investigates the performance of this heuristic.

In related work Weiss (1990, 1992, 1995) has analyzed the performance of index-based heuristics in several models for the optimal scheduling of a *batch* of stochastic jobs on parallel machines. He has argued that the index rules considered, which may be thought of as policies whose aim is to drive down fastest the cost rate of waiting jobs, are suboptimal because of an *end effect* caused by the loss of processing efficiency when the number of machines exceeds that of jobs present. He was able to bound the magnitude of this effect by deriving and applying certain decomposition formulae for the system's total expected workload. He thus obtained suboptimality bounds, independent of the batch size, for the index rules considered. Asymptotic optimality as the batch size grows to infinity follows. Weiss further argued the importance of proceeding to analyze index rules

Subject classifications: Queues/optimization: multiclass queues, parallel servers. Dynamic programming: performance guarantees for heuristic policies.
Area of review: Stochastic Models.

0030-364X/01/4904-0609 $05.00
1526-5463 electronic ISSN

Operations Research © 2001 INFORMS
Vol. 49, No. 4, July–August 2001, pp. 609–623

in more complex models incorporating job arrivals, such as queueing networks. This is the task we undertake in the present paper.

In our analysis of the performance of Klimov's rule in the above multiclass $M/M/m$ system we shall focus on the following issues.

1. Approximate optimality. How far from the optimal cost can the expected cost under Klimov's rule be? How large can the gap be between the expected cost achieved by Klimov's rule in the parallel and in the pooled systems? Can one obtain simple bounds for the corresponding gaps?

2. Heavy traffic optimality. Does the relative suboptimality gap for Klimov's rule vanish in heavy traffic, as arrival rates approach system capacity?

Our findings support the claim that Klimov's rule is a good heuristic for the parallel-server system: We show that both its suboptimality gap and the gap between its expected cost in the parallel and in the pooled systems are uniformly bounded with respect to (i) external arrival rates, as long as they stay within system capacity; and (ii) the number of servers. The first such uniform boundness result implies its *heavy-traffic optimality*, in the following sense: The *relative* suboptimality gap of Klimov's rule vanishes as external arrival rates approach system capacity. We note that this notion of heavy-traffic optimality is not the standard one in the literature on queueing systems control (cf., Harrison 1998), where one typically considers the asymptotic behaviour of a sequence of systems appropriately scaled in time and space. The form of heavy-traffic optimality established in this paper is technically simpler, yet we believe it has the advantage of being intuitive.

In fact, we establish a stronger result, namely that the relative gap between the expected performance of Klimov's rule in the parallel and in the pooled systems vanishes in heavy traffic, in the sense stated above. The fact that intelligent dynamic scheduling of a queueing network may lead to an effective pooling of processing resources in heavy traffic has been studied in a variety of models (see, e.g., the review paper by Kelly and Laws 1993). However, as pointed out by Harrison (1998), "studies of resource pooling have been largely heuristic to date." Harrison proves a resource pooling result, and establishes a strong form of heavy-traffic optimality for a specific policy in the context of a model different from the one discussed here.

The approach in this paper to a rigorous development of a resource pooling/heavy-traffic optimality result is radically different and is based on an analysis of the system's *region of achievable mean queue lengths* (see below). We believe that this approach has the potential to be extended to more complex systems.

Our mode of analysis is the so-called *achievable region approach to stochastic optimisation*. In outline, this approach proceeds as follows. With each admissible control u for a stochastic system of interest, a performance vector $\boldsymbol{x}^u$ is associated, which in our analyses will always be a vector of mean queue lengths. A cost $c(\boldsymbol{x}^u)$ is incurred when control u is applied, which depends upon u only through performance vector $\boldsymbol{x}^u$. The stochastic optimisation problem seeks a cost minimising control u^{OPT}. We write

$$Z^{\min} = \inf\{c(\boldsymbol{x}^u) : u \in \mathcal{U}\}, \tag{1}$$

where $\mathcal{U}$ is a set of *admissible* controls. An achievable region approach to such a problem will seek to obtain or characterize the set of all possible performance vectors (the *achievable region*) of the system, given by

$$\mathcal{X} = \{\boldsymbol{x}^u : u \in \mathcal{U}\}. \tag{2}$$

The approach will then identify a cost-minimising performance $\boldsymbol{x}^{\mathrm{OPT}}$, which attains the infimum in the equation

$$Z^{\min} = \inf\{c(\boldsymbol{x}) : x \in \mathcal{X}\}. \tag{3}$$

Plainly, any control which realizes $\boldsymbol{x}^{\mathrm{OPT}}$ solves the stochastic optimisation problem. The final step is to identify such controls.

This approach was introduced in a seminal paper by Coffman and Mitrani (1980) and has since been extended to ever more encompassing frameworks in Gelenbe and Mitrani (1980), Federgruen and Groenevelt (1988), Ross and Yao (1989), Shanthikumar and Yao (1992), and Bertsimas and Niño-Mora (1996). In all these analyses, the agenda outlined in Equations (1)–(3) is carried through in full. Bertsimas and Niño-Mora (1996) use the achievable region approach to unify classical priority index optimality results in a variety of problem domains, including deterministic machine scheduling (Smith's rule; see Smith 1956), multi-armed bandits (Gittins' rule; see Gittins and Jones 1974), and multiclass queueing networks (Klimov's rule; see Klimov 1974, 1978). The technical challenge posed by the parallel server system studied here lies in the fact that when $m \geqslant 2$, we cannot identify the achievable region $\mathcal{X}$. However, a new *work decomposition* result (see Step 1 below) enables us to identify a polyhedron $\mathcal{P}$ which contains $\mathcal{X}$. It is this which facilitates the analysis.

The paper proceeds as follows. The parallel-server system that is our prime object of study is described in §2. To assist the reader we also give a brief account of an achievable region analysis of this system in the single-server (or pooled) case, when Klimov's rule is optimal. In §§3–5 we analyse the parallel server system according to the following three-step plan:

Step 1. **Formulate a family of work decomposition laws for the parallel-server system.** This is the subject matter of §3. In a system whose n customer classes are labelled $\{1, 2, \ldots, n\} \equiv \mathcal{N}$ we obtain, for each $S \subseteq \mathcal{N}$, an expression for the mean workload over classes in S (*S-workload*). The S-workload under control u is given by $\sum_{j \in S} V_j^S x_j^u$, where $\boldsymbol{V} = (V_j^S)_{j \in \mathcal{N},\, S \subseteq \mathcal{N}}$ is a matrix whose nonnegative entries have a workload interpretation and x_j^u is written for $E_u[L_j]$, the mean queue length of class j under u. Theorem 1 gives an expression for this quantity for our model, which decomposes it into interpretable components. We describe how this new result relates to previous work decomposition results in the literature.

Step 2. **Use the work decomposition laws to demonstrate approximate optimality of Klimov's rule**. Refer to our brief description of the achievable region approach in Equations (1)–(3) above. In our analysis, $\boldsymbol{x}^u$ is the vector of mean queue lengths under u and $c(\boldsymbol{x}^u)$ is linear and given by

$$c(\boldsymbol{x}^u) = \sum_{j=1}^{n} c_j x_j^u.$$

In the parallel server case ($m \geqslant 2$), the achievable region $\mathscr{X}$ is not available. However, we utilise the work decomposition laws in Theorem 1 to show that a polyhedron $\mathscr{P}$ of the form

$$\mathscr{P} = \left\{ \boldsymbol{x} \in \Re_+^n : \sum_{j \in S} V_j^S x_j \geqslant \beta(S),\ S \subseteq \mathscr{N} \right\} \tag{4}$$

contains $\mathscr{X}$. In this situation, Equation (4) now extends to

$$Z^{\min} = \inf \left\{ \sum_{j=1}^{n} c_j x_j :\ x \in \mathscr{X} \right\} \geqslant \min \left\{ \sum_{j=1}^{n} c_j x_j :\ x \in \mathscr{P} \right\}, \tag{5}$$

where the last term in (5) is the value of a linear program (LP) and is denoted by Z^{LP}. We are able to identify a feasible solution to the dual of this LP with associated value Z^{D}. Writing Z^{KR} for the cost associated with Klimov's rule, we invoke weak LP duality to infer that

$$Z^{\mathrm{D}} \leqslant Z^{\mathrm{LP}} \leqslant Z^{\min} \leqslant Z^{\mathrm{KR}}, \tag{6}$$

which immediately yields $Z^{\mathrm{KR}} - Z^{\mathrm{D}}$ as a bound on the suboptimality gap $Z^{\mathrm{KR}} - Z^{\min}$. These ideas are presented in the context of general service systems in §4, which extends and develops earlier work by Glazebrook and Garbe (1999). This general theory is applied to the parallel server queueing network in §5. A simple bound on the suboptimality gap $Z^{\mathrm{KR}} - Z^{\min}$ is given in Theorem 3, which is our principal approximate optimality result. The uniform boundedness results mentioned above follow simply. See Corollary 2.

Step 3. **Use the approximate optimality results to infer heavy-traffic optimality**. From the approximate optimality result in Theorem 3 it is a relatively straightforward matter to establish that the relative suboptimality gap of Klimov's rule, namely $(Z^{\mathrm{KR}} - Z^{\min})/Z^{\min}$, vanishes in a suitably defined heavy-traffic limit. The same is true of a related quantity which measures the relative performance degradation of Klimov's rule due to parallelism. The details are given in Corollary 3.

§6 ends the paper with some concluding remarks and directions for further research.

2. THE MODEL

We consider a single-station Markovian multiclass queueing network populated by n customer classes which are serviced by m identical parallel servers. Customers of class $i \in \mathscr{N} = \{1, \ldots, n\}$ (or *i-customers*) arrive at the network from outside, according to a Poisson process with rate $\alpha_i \geqslant 0$. They may be processed by any server, and their service time is drawn from an exponential distribution with rate μ_i. Upon completion of his service, an i-customer is subject to Bernoulli routing, moving on to receive further service as a j-customer with probability p_{ij}, and leaving the network with probability $1 - \sum_{j \in \mathscr{N}} p_{ij}$. Routing probability matrix $\boldsymbol{P} = (p_{ij})_{i,\, j \in \mathscr{N}}$ is such that $\boldsymbol{I} - \boldsymbol{P}$ is invertible, which ensures that a single customer entering the network eventually exits. We further assume that all customer arrival processes, service times and routing events are mutually independent. This model is related to the multiclass $M/G/1$ queueing network studied by Klimov (1974). It is more general in that it incorporates parallel servers, and yet it is more restricted in requiring exponential service times rather than the general service times of Klimov. Research aimed at extending the results of the paper to a model with general service times and nonpreemptive scheduling policies is ongoing.

We next describe other quantities of interest for our system. The *total arrival rate of j-customers*, denoted by λ_j, is given by the solution of the *traffic equations*,

$$\lambda_j = \alpha_j + \sum_{i \in \mathscr{N}} \lambda_i p_{ij}, \qquad \text{for } j \in \mathscr{N},$$

and corresponds to the rate at which j-customer arrivals (external and internal) occur. The *traffic intensity of j-customers*, denoted by ρ_j, is given by

$$\rho_j = \frac{\lambda_j}{\mu_j}, \qquad \text{for } j \in \mathscr{N},$$

and represents the steady-state expected number of j-customers in service. The *total traffic intensity* ρ is given by

$$\rho = \sum_{j \in \mathscr{N}} \rho_j$$

and represents the steady-state expected number of busy servers. Given a subset of customer classes $S \subseteq \mathscr{N}$, we define similarly the *traffic intensity of S-customers* by

$$\rho(S) = \sum_{j \in S} \rho_j.$$

To develop more general notions of traffic intensity/system workload, we require the notion of the *mean S-workload of a j-customer*, for $j \in S$, denoted by V_j^S. We define this as the mean remaining service time a current j-customer receives until he leaves classes in subset S for the first time following completion of his current service. The V_j^Ss can be computed by solving the linear system

$$V_i^S = \frac{1}{\mu_i} + \sum_{j \in S} p_{ij} V_j^S, \qquad \text{for } i \in \mathscr{N},\ S \subseteq \mathscr{N}, \tag{7}$$

whose solution also defines parameters V_i^S, for $i \in S^c = \mathcal{N}\backslash S$. We further define the *external traffic intensity for S-customers* by

$$\rho^0(S) = \sum_{j \in S} \alpha_j V_j^S.$$

The network evolution is governed by a *scheduling policy*, which is a rule for dynamically allocating servers to available customers. We consider the space $\mathcal{U}$ of *admissible* scheduling policies to consist of all policies that are (1) *nonanticipative* (scheduling decisions are only based on system history up to and including the present time), (2) *preemptive* (the service of a customer may be interrupted at any time and resumed later), and (3) *nonidling* (a server is not allowed to stay idle when there are customers waiting). To guarantee that all such policies are *stable*, we shall assume the well-known condition

$$\rho < m$$

to hold.

We consider the following stochastic processes, which describe the system's evolution:

- $L_j(t)$: number of j-customers in the system at time t.
- $B_j^k(t)$: 1 if server k is busy with a j-customer at time t; 0 otherwise.
- $B_j(t)$: 1 if a j-customer is in service at time t; 0 otherwise.
- $B^k(t)$: 1 if server k is busy at time t; 0 otherwise.

We assume that the network operates in a steady-state regime, and we write L_j, B_j^k, B_j, and B^k to denote random variables with the steady-state distributions of the corresponding processes at an arbitrary time. It will simplify our notation considerably if we now introduce *performance vector* $\boldsymbol{x}^u$. This is the vector of mean queue lengths whose jth component is $x_j^u = E_u[L_j]$, where $E_u[\cdot]$ denotes a steady-state expectation taken under policy u. We now develop the *optimal scheduling problem* of interest by considering a cost structure in which j-customers incur linear holding costs at rate $c_j \geqslant 0$ per unit time in the system (waiting or in service). Our concern is with the problem of finding a scheduling policy to minimize the steady-state expected holding cost rate, and with evaluating the corresponding minimum cost, $Z^{\min}$. We write

$$Z^{\min} = \inf \left\{ \sum_{j \in \mathcal{N}} c_j x_j^u \colon u \in \mathcal{U} \right\}. \tag{8}$$

2.1. The Single-Server Case

An exact solution of the above problem is available in the special single-server case and is due to Klimov (1974, 1978). It will assist the reader if we sketch the main ideas involved in the achievable region approach to this special case, because our analysis of the parallel-server model is based on and extends them. To be precise, in this section we shall consider a system with a single server of speed m. This can be helpfully thought of as an approximation to the above parallel-server system with m servers each of speed 1.

Klimov (1974, 1978) showed that the optimal policy for such a single-server network is given by the following priority-index rule: Compute index vector $\gamma = (\gamma_j)_{j \in \mathcal{N}}$ by running Klimov's *adaptive greedy algorithm* (see Figure 1) on input $(\boldsymbol{c}, \boldsymbol{V})$, where $\boldsymbol{c} = (c_j)_{j \in \mathcal{N}}$ and $\mathbf{V} = (V_j^S)_{j \in \mathcal{N},\, S \subseteq \mathcal{N}}$ is the matrix with entries obtained from (7). Klimov's index rule operates by giving at each time higher preemptive priority to a customer with larger index. He interpreted the index γ_i as the maximum rate of decrease in expected holding cost per unit of expected processing time for a customer currently in class i.

Tsoucas (1991) extended Klimov's work by elucidating properties of performance vector $\boldsymbol{x}^u$, defined above as the vector of steady-state expected queue lengths. He demonstrated the existence of a non-negative set function $b(S)$, such that for any admissible scheduling policy u and $S \subset \mathcal{N}$,

$$\sum_{j \in S} V_j^S x_j^u \geqslant b(S), \tag{9}$$

with equality achieved in Equation (9), when policy u gives preemptive priority to classes in S. Moreover, for $S = \mathcal{N}$ we have for all admissible u that

$$\sum_{j \in \mathcal{N}} V_j^{\mathcal{N}} x_j^u = b(\mathcal{N}). \tag{10}$$

An explicit expression for $b(S)$ was given in Bertsimas et al. (1994), which, incorporating speedup factor m, simplifies to

$$b(S) = \frac{\sum_{j \in S} \rho_j V_j^S}{m - \rho^0(S)}, \qquad S \subseteq \mathcal{N}. \tag{11}$$

Consider now the *achievable performance region* $\mathcal{X} = \{\boldsymbol{x}^u : u \in \mathcal{U}\}$, spanned by vector $\boldsymbol{x}^u$ as u ranges over $\mathcal{U}$. Tsoucas (1991) showed that, in the single-server case, $\mathcal{X}$ is precisely the bounded polyhedron defined by linear constraints (9) and (10), namely,

$$\mathcal{P}_{\text{pooled}} = \left\{ \boldsymbol{x} \in \Re_+^n : \sum_{j \in S} V_j^S x_j \geqslant b(S), S \subset \mathcal{N}, \quad \text{and} \right.$$
$$\left. \sum_{j \in \mathcal{N}} V_j^{\mathcal{N}} x_j = b(\mathcal{N}) \right\}.$$

It thus follows that the minimum cost for the optimal scheduling problem, which we denote by $Z_{\text{pooled}}^{\min}$ (as it is achieved by Klimov's rule), can be computed as the optimal value of an LP problem as follows:

$$Z_{\text{pooled}}^{\min} = \min \left\{ \sum_{j \in \mathcal{N}} c_j x_j : \boldsymbol{x} \in \mathcal{P}_{\text{pooled}} \right\}. \tag{12}$$

The optimality of Klimov's rule is a consequence of the fact that his adaptive greedy algorithm produces an optimal solution $\{\bar{y}(S),\ S \subseteq \mathcal{N}\}$ to the dual of LP Problem (12).

Figure 1. Klimov's adaptive greedy algorithm.

Input: $(\boldsymbol{c}, \boldsymbol{V})$, where $\boldsymbol{c} = (c_j)_{j\in\mathcal{N}}$ and $\boldsymbol{V} = (V_j^S)_{j\in\mathcal{N}, S\subseteq\mathcal{N}}$.
Output: $(\pi, \bar{\boldsymbol{y}}, \gamma)$, where $\pi = (\pi_1, \ldots, \pi_n)$ is a permutation of $\mathcal{N}$, $\bar{\boldsymbol{y}} = (\bar{y}(S))_{S\subseteq\mathcal{N}}$ and $\gamma = (\gamma_1, \ldots, \gamma_n)$.
Step 1. Set $S_1 = \mathcal{N}$;
set $\bar{y}(S_1) = \min\{c_i/V_i^{S_1} : i \in S_1\}$;
pick $\pi_1 \in \operatorname{argmin}\{c_i/V_i^{S_1} : i \in S_1\}$;
set $\gamma_{\pi_1} = \bar{y}(S_1)$.
Step k. For $k = 2, \ldots, n$:
set $S_k = S_{k-1} \setminus \{\pi_{k-1}\}$; set $\bar{y}(S_k) = \min\{[c_i - \sum_{j=1}^{k-1} V_i^{S_j}\, \bar{y}(S_j)]/V_i^{S_k} : i \in S_k\}$;
pick $\pi_k \in \operatorname{argmin}\{[c_i - \sum_{j=1}^{k-1} V_i^{S_j}\, \bar{y}(S_j)]/V_i^{S_k} : i \in S_k\}$;
set $\gamma_{\pi_k} = \gamma_{\pi_{k-1}} + \bar{y}(S_k)$.
Step n+1. For $S \subseteq \mathcal{N}$:
set $\bar{y}(S) = 0$ if $S \notin \{S_1, \ldots, S_n\}$.

This result was actually the crux of Klimov's (1974) original analysis, based on an equivalent LP formulation, and has recently been extended by Bertsimas and Niño-Mora (1996) into a general framework. This approach yields the result that

$$Z_{\text{pooled}}^{\min} = \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\, b(\{j, \ldots, n\}), \tag{13}$$

where the customer classes are renumbered so that $\gamma_1 \leqslant \cdots \leqslant \gamma_n$, and we adopt the convention that $\gamma_0 = 0$. In §5.1, Identity (13) will allow us to compare the performance of Klimov's rule in the parallel and pooled networks.

3. WORK DECOMPOSITION LAWS

The properties of performance vector $\boldsymbol{x}^u$ enunciated in Equations (9)–(10) are central to the analysis of the single-server case. Bertsimas and Niño-Mora (1996) introduced the term *generalised conservation laws* (GCL) to describe this set of relations and showed that such laws are satisfied in a range of systems for suitably chosen $\boldsymbol{x}^u$, $\boldsymbol{V}$ and b. They further showed that, for a performance vector $\boldsymbol{x}^u$ that satisfies GCL, the problem of finding a scheduling policy that optimizes a linear performance objective is solved by a priority-index rule. It will emerge in our analysis that our parallel-server system does not satisfy GCL when $m \geqslant 2$, yet it comes close to doing so. Consequently, a suitably constructed priority-index rule comes close to being optimal for our linear objective. From Equations (9)–(10), we note that the key to developing such ideas lies in an ability to characterise the quantities $\sum_{j\in S} V_j^S\, x_j^u$ for any u and $S \subseteq \mathcal{N}$. The appropriate characterisation is given in the *work decomposition* result in Theorem 1, which is the main result of this section.

In a variety of single-server multiclass queueing systems, researchers have identified *work decomposition laws*, which describe a linear relation between the steady-state expected number in system from each customer class at an arbitrary time and at an arbitrary time during an interval when the server is idle. These laws have played a major role in the performance analysis of vacation and polling models (see, e.g., Boxma 1989 and references therein). Recently, Bertsimas and Niño-Mora (1999a, 1999b) have extended this work by identifying new work decomposition laws satisfied by (semi-) Markovian multiclass queueing networks with one or multiple single-server stations and have applied them to obtain improved performance bounds. We extend that line of research by obtaining the family of new work decomposition laws given in Theorem 1 below for the parallel-server model under study. These laws will play a central role in our analysis of Klimov's rule as it is developed in §§4 and 5. We remark that Weiss (1992) has established a similar work decomposition result for the batch case, which is also central to his approach. The reader should note that in Theorem 1, the matrix $\boldsymbol{V}$ and the performance vector $\boldsymbol{x}^u$ are as in §2.

THEOREM 1 (WORK DECOMPOSITION LAWS). *Under any nonanticipative, stable and preemptive scheduling policy u, and for any subset $S \subseteq \mathcal{N}$ of customer classes,*

$$\sum_{j\in S} V_j^S x_j^u = b(S) + \Delta_{\text{pr}}^u(S) + \Delta_{\text{id}}^u(S), \tag{14}$$

where

$$\begin{aligned} b(S) &= \frac{\sum_{j\in S} \rho_j\, V_j^S}{m - \rho^0(S)}, \\ \Delta_{\text{pr}}^u(S) &= \frac{\sum_{i\in S^c} \mu_i\, V_i^S \sum_{j\in S} V_j^S\, E_u\big[\big(\sum_{k=1}^m B_i^k\big)\, L_j\big]}{m - \rho^0(S)}, \\ \Delta_{\text{id}}^u(S) &= \frac{\sum_{j\in S} V_j^S\, E_u\big[\big(\sum_{k=1}^m (1 - B^k)\big)\, L_j\big]}{m - \rho^0(S)}. \end{aligned} \tag{15}$$

The quantity on the left-hand side of Equation (14) can be thought of as a measure of the steady-state mean amount of work in the system resulting from jobs in S (*S-workload*) under policy u. Close inspection of the terms in (14) will yield the conclusion that $\Delta_{\text{pr}}^u(S)$ is a *priority* term identifying a contribution to the mean S-workload under policy u when other classes (in S^c) are in service. Similarly, $\Delta_{\text{id}}^u(S)$ is an *idleness* term identifying such a contribution when some server is idle. Note that the priority term disappears when $S = \mathcal{N}$.

3.1. Derivation of the Work Decomposition Laws

To establish Theorem 1 we use the following two-step approach, introduced by Bertsimas and Niño-Mora (1999a, 1999b), in the setting of multiclass queueing networks with single-server stations.

Step 1. Utilise flow balance ideas to develop a set of linear equations involving the performance $\boldsymbol{x}^u$ and some auxiliary performance measures. This gives Lemma 2.

Step 2. Reformulate the resulting set of equations to derive the required work decomposition laws in Theorem 1.

STEP 1. FLOW BALANCE EQUATIONS. We now give a brief overview of the flow balance ideas we require before applying them to the parallel-server system of interest. A classical result of queueing theory states that, in a stable queueing system in which customers arrive and leave one at a time, the steady-state distribution of the number in system observed *just before* embedded arrival epochs, L^-, equals that *just after* departure epochs, L^+ (see, e.g., Burke 1956, and Finch 1959). This result, which follows from the system's flow balance equations, implies that

$$E^A[L^-] = E^D[L^+], \tag{16}$$

where $E^A[\cdot]$ (respectively $E^D[\cdot]$) denotes a steady-state expectation taken at embedded customer arrival (respectively, departure) epochs. It has been shown in Bertsimas and Niño-Mora (1999a, 1999b) that *event-average* identity (16) can be applied to a variety of multiclass queueing network models with single-server stations to formulate linear equations on performance measures representing steady-state expectations at an arbitrary time (*time averages*). The resulting equations are precisely those derived previously through the so-called *potential function method* in Bertsimas et al. (1994) and in Kumar and Kumar (1994), and thus reveal their fundamental physical interpretation. Note that a derivation of Theorem 1 via the potential function method would exploit the time independence of the second moment of the random quantity $\sum_{j\in S} V_j^S L_j(t)$ taken with respect to the steady-state distribution. We finally point out the fact that the flow-balance approach pursued here to derive linear equations on network performance measures was actually introduced by Klimov (1974) in his pioneering analysis of the single-server network; It thus predates by two decades recent derivations of his equations via the potential function method.

The basic idea for reformulating event-average identity (16) into an identity involving only time averages in a general queueing system is as follows: Let $\{L(t), t \geqslant 0\}$ be the number-in-system process, and suppose customer arrivals and departures are driven by nonanticipative *stochastic intensity* processes $\{\lambda(t), t \geqslant 0\}$ and $\{\mu(t), t \geqslant 0\}$, respectively, so that for any time $t \geqslant 0$, $E[\lambda(t)] = E[\mu(t)] = \lambda$, where $E[\cdot]$ denotes a steady-state time-stationary expectation. The key tool to relate event and time averages in the presence of stochastic intensities is *Papangelou's formula* of Palm calculus (see Papangelou 1972, Brémaud 1989), an extension of the well-known PASTA (Poisson Arrivals See Time Averages) property of queueing theory, which yields

$$\begin{aligned}\lambda E^A[L^-] &= E[\lambda(t)\, L(t)], \text{ and}\\ \lambda E^D[L^+] &= \lambda E^D[L^- - 1] = E[\mu(t)(L(t)-1)]\\ &= E[\mu(t)\, L(t)] - \lambda.\end{aligned}$$

The next result presents the corresponding reformulation of flow-balance identity (16) in terms of time averages.

LEMMA 1. *Under the above assumptions, for any* $t \geqslant 0$,

$$E[(\mu(t) - \lambda(t))\, L(t)] = \lambda. \tag{17}$$

We shall now apply these ideas to the parallel-server network under consideration. To do so we require some additional notation. We shall write, for a scheduling policy u,

$$x_{ij}^u = E_u\left[\left\{\sum_{k=1}^m B_i^k\right\} L_j\right], \quad \boldsymbol{X}^u = (x_{ij}^u)_{i,\, j\in\mathcal{N}}, \tag{18}$$

and

$$x_{0j}^u = E_u\left[\left\{\sum_{k=1}^m (1 - B^k)\right\} L_j\right].$$

We note in passing that the identity

$$B^k = \sum_{i\in\mathcal{N}} B_i^k$$

implies that

$$x_j^u = E_u[L_j] = \sum_{i\in\mathcal{N}} E_u[B_i^k\, L_j] + E_u[(1 - B^k)\, L_j], \tag{19}$$

whereupon summing both sides of (19) over k, $1 \leqslant k \leqslant m$, yields

$$m\, x_j^u = \sum_{i\in\mathcal{N}} x_{ij}^u + x_{0j}^u, \quad j \in \mathcal{N}. \tag{20}$$

We apply next the previous ideas to obtain a set of flow-balance equations on network performance measures in Lemma 2. As mentioned before, the corresponding set of equations for the special single-server case was first derived by Klimov (1974), also using flow-balance arguments. Let $\boldsymbol{\alpha} = (\alpha_j)_{j\in\mathcal{N}}$ and let $\boldsymbol{D}_\lambda$ (respectively, $\boldsymbol{D}_\mu$) denote the diagonal matrix corresponding to vector $\boldsymbol{\lambda} = (\lambda_j)_{j\in\mathcal{N}}$ (respectively, $\boldsymbol{\mu} = (\mu_j)_{j\in\mathcal{N}}$). We further denote by $\boldsymbol{I}$ the appropriate identity matrix.

LEMMA 2. *Under any nonanticipative and preemptive scheduling policy u, performance measures* $\boldsymbol{x}^u$ *and* $\boldsymbol{X}^u$ *satisfy the following set of linear equations*:

$$\begin{aligned}-\boldsymbol{\alpha}\boldsymbol{x}^{u'} - \boldsymbol{x}^u\boldsymbol{\alpha}' + (\boldsymbol{I}-\boldsymbol{P})'\boldsymbol{D}_{\boldsymbol{\mu}}\boldsymbol{X}^u + \boldsymbol{X}^{u'}\boldsymbol{D}_{\boldsymbol{\mu}}(\boldsymbol{I}-\boldsymbol{P})\\ = (\boldsymbol{I}-\boldsymbol{P})'\boldsymbol{D}_{\boldsymbol{\lambda}} + \boldsymbol{D}_{\boldsymbol{\lambda}}(\boldsymbol{I}-\boldsymbol{P}).\end{aligned} \tag{21}$$

PROOF. DIAGONAL EQUATION (j, j). The jth diagonal equation in (21) formulates flow-balance identity (16) as it applies to the queueing system consisting of j-customers only,

$$E^{A_j}[L_j^-] = E^{D_j}[L_j^+],$$

where $E^{A_j}[\cdot]$ (respectively $E^{D_j}[\cdot]$) denotes an expectation taken with respect to the steady-state distribution at *net* j-customer arrival (respectively departure) epochs, i.e., excluding feedback epochs from class j into itself. The stochastic intensity of net j-customer arrivals at time t is given by

$$\lambda^{A_j}(t) = \alpha_j + \sum_{l \in \mathcal{N}\setminus\{j\}} \mu_l p_{lj} \sum_{k=1}^{m} B_l^k(t), \tag{22}$$

whereas the stochastic intensity of net j-customer departures at t is

$$\mu^{D_j}(t) = \mu_j(1-p_{jj}) \sum_{k=1}^{m} B_j^k(t), \tag{23}$$

with $E[\lambda^{A_j}(t)] = E[\mu^{D_j}(t)] = \lambda_j(1-p_{jj})$. Hence when applied to this case, Lemma 1 yields the equation

$$E_u[\{\mu^{D_j}(t) - \lambda^{A_j}(t)\}L_j(t)] = \lambda_j(1-p_{jj}). \tag{24}$$

Substituting from Equations (22) and (23) into (24) and utilising (18) we conclude that

$$-\alpha_j x_j^u + \sum_{l=1}^{n} \mu_l(\delta_{lj} - p_{lj}) x_{lj}^u = \lambda_j(1-p_{jj}), \tag{25}$$

where δ_{ij} is Kronecker's delta. We note that (25) is the jth diagonal equation in (21) scaled by $1/2$.

EQUATION (i, j). The equation corresponding to row i and column j in (21), with $i \neq j$, formulates the flow-balance identity (17) in Lemma 1 as it applies to the queueing system of $\{i, j\}$-customers, having $L_i(t) + L_j(t)$ customers in the system at time t. The stochastic intensity of *net* $\{i, j\}$-customer arrivals, i.e., excluding feedback epochs from classes in $\{i, j\}$ into $\{i, j\}$, is

$$\lambda^{A_{ij}}(t) = \alpha_i + \alpha_j + \sum_{l \in \mathcal{N}\setminus\{i,j\}} \mu_l(p_{li} + p_{lj}) \sum_{k=1}^{m} B_l^k(t),$$

whereas the stochastic intensity of net $\{i, j\}$-customer departures is

$$\mu^{D_{ij}}(t) = \mu_i(1-p_{ii}-p_{ij}) \sum_{k=1}^{m} B_i^k(t) + \mu_j(1-p_{jj}-p_{ji}) \sum_{k=1}^{m} B_j^k(t),$$

having as steady-state expectation at an arbitrary time

$$\begin{aligned} E_u[\lambda^{A_{ij}}(t)] &= E_u[\mu^{D_{ij}}(t)] \\ &= \lambda_i(1-p_{ii}-p_{ij}) + \lambda_j(1-p_{jj}-p_{ji}). \end{aligned}$$

We now substitute these expressions into (17) and simplify by using the ith and jth diagonal equations. This yields

$$-\alpha_i x_j^u - \alpha_j x_i^u + \sum_{l=1}^{n} \mu_l(\delta_{lj} - p_{lj}) x_{li}^u + \sum_{l=1}^{n} \mu_l(\delta_{li} - p_{li}) x_{lj}^u = -p_{ji}\lambda_j - p_{ij}\lambda_i,$$

which is precisely the equation in position (i, j) in (21). This completes the proof. □

STEP 2. WORKLOAD REFORMULATION. We now show that the equations in Lemma 2 can be reformulated to yield the work decomposition laws of Theorem 1. In developing the analysis we shall require the following notational conventions: If $S,\ T \subseteq \mathcal{N},\ \boldsymbol{z} = (z_i)_{i \in \mathcal{N}}$ is an n-vector, and $\boldsymbol{A} = (a_{ij})_{i,\, j \in \mathcal{N}}$ is an $n \times n$ matrix, we shall write

$$\boldsymbol{z}_S = (z_j)_{j \in S}, \quad \text{and} \quad \boldsymbol{A}_{ST} = (a_{ij})_{i \in S,\, j \in T}.$$

PROOF OF THEOREM 1. Let $S \subseteq \mathcal{N}$, and let $v(S)$ denote the n-vector

$$v(S) = \begin{pmatrix} \boldsymbol{V}_S^S \\ \boldsymbol{0} \end{pmatrix}.$$

We shall derive work decomposition identity (14) from (21) by pre- and post-multiplying both sides of (21) by $v(S)'$ (the transpose of $v(S)$) and $v(S)$, respectively. We shall then simplify the resulting equation using (20).

The calculation based on the right-hand side of (21) yields (we incorporate a $1/2$ scaling factor for convenience)

$$\begin{aligned} &\frac{1}{2}\begin{pmatrix} \boldsymbol{V}_S^{S'} & \boldsymbol{0} \end{pmatrix} \{(\boldsymbol{I}-\boldsymbol{P})'\boldsymbol{D}_{\boldsymbol{\lambda}} + \boldsymbol{D}_{\boldsymbol{\lambda}}(\boldsymbol{I}-\boldsymbol{P})\} \begin{pmatrix} \boldsymbol{V}_S^S \\ \boldsymbol{0} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{V}_S^{S'} & \boldsymbol{0} \end{pmatrix} (\boldsymbol{I}-\boldsymbol{P})'\boldsymbol{D}_{\boldsymbol{\lambda}} \begin{pmatrix} \boldsymbol{V}_S^S \\ \boldsymbol{0} \end{pmatrix} \\ &= \left\{ \begin{pmatrix} \boldsymbol{I}_S - \boldsymbol{P}_{SS} & -\boldsymbol{P}_{SS^c} \\ -\boldsymbol{P}_{S^cS} & \boldsymbol{I}_{S^c} - \boldsymbol{P}_{S^cS^c} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_S^S \\ \boldsymbol{0} \end{pmatrix} \right\}' \boldsymbol{D}_{\boldsymbol{\lambda}} \begin{pmatrix} \boldsymbol{V}_S^S \\ \boldsymbol{0} \end{pmatrix} \\ &= \left(\left(\tfrac{1}{\mu_i}\right)'_{i \in S} \left(\tfrac{1}{\mu_i}\right)'_{i \in S^c} - \boldsymbol{V}_{S^c}^{S\,\prime} \right) \boldsymbol{D}_{\boldsymbol{\lambda}} \begin{pmatrix} \boldsymbol{V}_S^S \\ \boldsymbol{0} \end{pmatrix} \\ &= \sum_{j \in S} \rho_j V_j^S. \end{aligned}$$

The calculation based on the left-hand side of (21) yields

$$\begin{aligned} &\frac{1}{2} v(S)' \{ -\boldsymbol{\alpha}\boldsymbol{x}^{u\prime} - \boldsymbol{x}^u\boldsymbol{\alpha}' \\ &\quad + (\boldsymbol{I}-\boldsymbol{P})'\boldsymbol{D}_{\boldsymbol{\mu}}\boldsymbol{X}^u + \boldsymbol{X}^{u\prime}\boldsymbol{D}_{\boldsymbol{\mu}}(\boldsymbol{I}-\boldsymbol{P}) \} v(S) \\ &= -(v(S)'\boldsymbol{\alpha})(v(S)'\boldsymbol{x}^u) \\ &\quad + \left\{ \begin{pmatrix} \boldsymbol{I}_S - \boldsymbol{P}_{SS} & -\boldsymbol{P}_{SS^c} \\ -\boldsymbol{P}_{S^cS} & \boldsymbol{I}_{S^c} - \boldsymbol{P}_{S^cS^c} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_S^S \\ \boldsymbol{0} \end{pmatrix} \right\}' \boldsymbol{D}_{\boldsymbol{\mu}} \boldsymbol{X}^u \begin{pmatrix} \boldsymbol{V}_S^S \\ \boldsymbol{0} \end{pmatrix} \\ &= -(v(S)'\boldsymbol{\alpha})(v(S)'\boldsymbol{x}^u) + \left(\left(\tfrac{1}{\mu_i}\right)'_{i \in S} \left(\tfrac{1}{\mu_i}\right)'_{i \in S^c} - \boldsymbol{V}_{S^c}^{S\prime} \right) \\ &\quad \times \boldsymbol{D}_{\boldsymbol{\mu}} \begin{pmatrix} \boldsymbol{X}_{SS}^u & \boldsymbol{X}_{SS^c}^u \\ \boldsymbol{X}_{S^cS}^u & \boldsymbol{X}_{S^cS^c}^u \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_S^S \\ \boldsymbol{0} \end{pmatrix} \\ &= -\rho^0(S) \sum_{j \in S} V_j^S x_j^u + \sum_{i \in S} \sum_{j \in S} V_j^S x_{ij}^u \\ &\quad - \sum_{i \in S^c} \sum_{j \in S} \mu_i \left(V_i^S - \frac{1}{\mu_i} \right) V_j^S x_{ij}^u. \end{aligned} \tag{26}$$

If we now equate the above two equivalent expressions and utilise (20) within (26), we obtain

$$\{m-\rho^0(S)\}\sum_{j\in S} V_j^S x_j^u = \sum_{j\in S}\rho_j V_j^S + \sum_{i\in S^c}\mu_i V_i^S \sum_{j\in S} V_j^S x_{ij}^u + \sum_{j\in S} V_j^S x_{0j}^u,$$

which immediately yields Theorem 1. □

4. WORK DECOMPOSITION LAWS AND THE CLOSENESS TO OPTIMALITY OF KLIMOV'S RULE

We now broach the question of how to deploy the work decomposition laws in Theorem 1 to analyse the stochastic optimisation problem of interest to us. For the single-server case discussed at the end of §2, Theorem 1 yields immediately the *generalised conservation laws* (GCL) in Equations (9)–(10). As outlined in §2, satisfaction of these laws implies the optimality of Klimov's rule via an achievable region analysis. In the parallel-server model with $m \geqslant 2$, we can utilise the achievable region approach together with Theorem 1 to obtain information on the closeness to optimality of Klimov's rule. To emphasise the broad scope of the ideas, we shall develop the material in the context of a general service system. The results here expand and reformulate the account given by Glazebrook and Garbe (1999). We shall indicate after Theorem 2 how Glazebrook and Garbe's results emerge from our analysis. In §5, the results of this section are applied to the queueing network under study.

Consider a general dynamic and stochastic *service system* consisting of a set of servers that provide service to customers belonging to a finite set of classes $\mathcal{N} = \{1, \ldots, n\}$. The system evolution is controlled by a scheduling policy u, one of a set of *admissible* policies $\mathcal{U}$, which specifies dynamically how servers are allocated to available customers. System performance under policy $u \in \mathcal{U}$ is measured by a *performance vector* $\boldsymbol{x}^u = (x_j^u)_{j\in\mathcal{N}}$, where x_j^u is a non-negative performance measure (an expectation) for class j customers. A central notion in this framework is that of *priority*. Given a subset $S \subseteq \mathcal{N}$ of customer classes, we say that a scheduling policy gives priority to *S-customers* (whose class is in S) over S^c-customers (where $S^c = \mathcal{N}\backslash S$) if no S^c-customer is allowed to enter service at the expense of an available S-customer having to wait.

Let $\boldsymbol{V} = (V_i^S)_{i\in\mathcal{N}, S\subseteq\mathcal{N}}$ be a matrix with $V_i^S > 0$, for $i \in S$. For each subset $S \subseteq \mathcal{N}$, let us define $\beta(S) \geqslant 0$ to be the minimum value achievable by performance objective $\sum_{j\in S} V_j^S x_j^u$ under admissible policies, namely

$$\beta(S) = \inf\left\{\sum_{j\in S} V_j^S\, x_j^u :\ u \in \mathcal{U}\right\}. \tag{27}$$

For any policy $u \in \mathcal{U}$, let $\Phi^u(S) \geqslant 0$ denote its corresponding suboptimality gap with respect to the objective stated above,

$$\Phi^u(S) = \sum_{j\in S} V_j^S\, x_j^u - \beta(S). \tag{28}$$

Finally, let $\Phi(S)$ denote the corresponding worst-case suboptimality gap achievable under admissible policies that give priority to S-customers,

$$\Phi(S) = \sup\{\Phi^u(S) :\ u \text{ gives priority to } S\text{-customers}\}. \tag{29}$$

If the set function Φ defined in Equation (29) is identically zero, we say that the system satisfies generalised conservation laws and priority index policies are optimal for linear performance objectives (see Bertsimas and Niño-Mora 1996). Our goal in this section is to investigate the closeness to optimality of index policies when GCL are not satisfied. Before proceeding further with our general development, note that for our multiclass $M/M/m$ queue with feedback, we will choose the matrix $\boldsymbol{V} = (V_i^S)_{i\in\mathcal{N}, S\subseteq\mathcal{N}}$ to be that defined at Equation (7) and $\boldsymbol{x}^u$ to be the vector of mean queue lengths under admissible control u. From the work decomposition laws in Theorem 1, we note that for this model and for these choices we have

$$\beta(S) = b(S) + \Delta^*(S), \tag{30}$$

where

$$\Delta^*(S) = \inf\left\{\Delta_{\mathrm{pr}}^u + \Delta_{\mathrm{id}}^u(S) :\ u \in \mathcal{U}\right\}. \tag{31}$$

It then follows that

$$\Phi^u(S) = \Delta_{\mathrm{pr}}^u + \Delta_{\mathrm{id}}^u(S) - \Delta^*(S), \tag{32}$$

and

$$\Phi(S) = \sup\{\Delta_{\mathrm{pr}}^u(S) + \Delta_{\mathrm{id}}^u(S) :\ u \text{ gives priority to } S\text{-customers}\} - \Delta^*(S). \tag{33}$$

We postpone further consideration of this model to §5.

Returning to our general dynamic and stochastic service system, we introduce the system's *achievable performance region* $\mathcal{X}$, defined by

$$\mathcal{X} = \{\boldsymbol{x}^u :\ u \in \mathcal{U}\}.$$

It may be hard to fully characterise the performance region $\mathcal{X}$ by means of constraints. However, note from (27) that $\mathcal{X}$ is contained in the polyhedron

$$\mathcal{P} = \left\{\boldsymbol{x} \in \Re_+^n :\ \sum_{j\in S} V_j^S x_j \geqslant \beta(S),\ S \subseteq \mathcal{N}\right\}.$$

Suppose that the *optimal scheduling problem* of interest is to choose an admissible control to minimise a linear performance objective. If the optimal value is $Z^{\min}$, we write

$$Z^{\min} = \inf\left\{\sum_{j\in\mathcal{N}} c_j x_j^u :\ u \in \mathcal{U}\right\}, \tag{34}$$

where $\boldsymbol{c} = (c_j)_{j\in\mathcal{N}} \geqslant 0$ is a given cost vector. In what follows, we shall write Z^u for the cost under control u,

namely $\sum_{j\in\mathcal{N}} c_j x_j^u$. We now consider an LP relaxation of this scheduling problem that consists of minimising the objective $\sum_{j\in\mathcal{N}} c_j x_j$ over the polyhedron $\mathcal{P}$. If the optimal value of this LP is Z^{LP}, we write

$$Z^{\mathrm{LP}} = \min\left\{\sum_{j\in\mathcal{N}} c_j x_j : \boldsymbol{x}\in\mathcal{P}\right\}. \tag{35}$$

Because $\mathcal{X}\subseteq\mathcal{P}$, it must follow that

$$Z^{\mathrm{LP}} \leqslant Z^{\min}. \tag{36}$$

We develop next a *primal-dual* approach to the optimal scheduling problem in (34), based on constructing simultaneously a heuristic solution for it and a feasible solution to the dual of LP relaxation (35). A suboptimality bound for the heuristic will follow by comparing the values of both solutions. The procedure is as follows:

Step 1. Run Klimov's adaptive greedy algorithm (see Figure 1) on input $(\boldsymbol{c}, \boldsymbol{V})$ to obtain output $(\boldsymbol{\pi}, \bar{\boldsymbol{y}}, \boldsymbol{\gamma})$.

Step 2. Take as a heuristic solution to Problem (34) the priority-index rule that gives higher priority to classes with higher index γ_i. We denote by Z^{KR} (KR ≡ Klimov's rule) its corresponding objective value.

Step 3. Take as a feasible solution to the dual of LP relaxation (34) vector $\bar{\mathbf{y}}$, with corresponding dual value Z^{D}.

We assume in what follows that classes are renumbered so that permutation π returned by Klimov's algorithm is $\pi = (1,\ldots,n)$, (i.e., class n has highest priority), and hence

$$\gamma_1 \leqslant \cdots \leqslant \gamma_n.$$

From weak LP duality and (36), we have the inequalities

$$Z^{\mathrm{D}} \leqslant Z^{\mathrm{LP}} \leqslant Z^{\min} \leqslant Z^{\mathrm{KR}}. \tag{37}$$

The next result gives a representation of objective Z^u under a general policy $u\in\mathcal{U}$. This is fundamental to our subsequent analysis. Here and elsewhere we adopt the convention that $\gamma_0 = 0$.

LEMMA 3. *For any admissible scheduling policy $u\in\mathcal{U}$,*

$$Z^u = Z^{\mathrm{D}} + \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\,\Phi^u(\{j,\ldots,n\}), \tag{38}$$

where

$$Z^{\mathrm{D}} = \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\,\beta(\{j,\ldots,n\}). \tag{39}$$

PROOF. The dual of LP problem (35) is

$$\max \sum_{S\subseteq\mathcal{N}} \beta(S)\, y(S),$$

subject to

$$\sum_{S\ni i} V_i^S y(S) \leqslant c_i, \quad \text{for } i\in\mathcal{N},$$

$$y(S) \geqslant 0, \quad \text{for } S\subseteq\mathcal{N}.$$

Now it is easily verified that the $\bar{\boldsymbol{y}}$ computed by Klimov's algorithm is a dual feasible solution that satisfies the constraints with equality (see Bertsimas and Niño-Mora 1996 for details):

$$\sum_{S\ni i} V_i^S \bar{y}(S) = c_i, \quad \text{for } i\in\mathcal{N}. \tag{40}$$

Note also from Klimov's algorithm that $\bar{\boldsymbol{y}}(S) = 0$ when $S \neq \{i,\ldots,n\}$ for some i, and that

$$\gamma_i - \gamma_{i-1} = \bar{\boldsymbol{y}}(\{i,\ldots,n\}), \quad \text{for } i\in\mathcal{N}. \tag{41}$$

It is straightforward from these observations that the dual value of $\bar{\mathbf{y}}$, given by

$$Z^{\mathrm{D}} = \sum_{S\subseteq\mathcal{N}} \beta(S)\bar{\mathbf{y}}(S),$$

reduces to the expression in (39).

We now use (40) and (41) to develop an expression for Z^u as follows:

$$\begin{aligned} Z^u &= \sum_{i=1}^{n} c_i x_i^u \\ &= \sum_{i=1}^{n}\sum_{S\ni i} V_i^S \bar{y}(S) x_i^u \\ &= \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1}) \sum_{i=j}^{n} V_i^{\{j,\ldots,n\}} x_i^u \\ &= \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})[\beta(\{j,\ldots,n\}) + \Phi^u(\{j,\ldots,n\})], \end{aligned} \tag{42}$$

by (28). Identity (38) now follows from (39) and (42). □

Our next result gives alternative representations for the gaps $Z^{\min} - Z^{\mathrm{D}}$ and $Z^{\mathrm{KR}} - Z^{\min}$ in the string of Inequalities (37), in terms of the functions $\Phi^u(\cdot)$ introduced above. Our prime interest is in the difference $Z^{\mathrm{KR}} - Z^{\min}$, which measures the suboptimality gap of Klimov's rule. The notation $\Phi^{\mathrm{KR}}(S)$ to be used in what follows refers to $\Phi^u(S)$ when u is the priority-index (Klimov's) rule previously defined.

LEMMA 4. *The gaps $Z^{\min} - Z^{\mathrm{D}}$ and $Z^{\mathrm{KR}} - Z^{\min}$ can be represented as follows*:

(a)

$$Z^{\min} - Z^{\mathrm{D}} = \inf\left\{\sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\Phi^u(\{j, j+1,\ldots,n\}) : u\in\mathcal{U}\right\};$$

(b)

$$Z^{KR} - Z^{\min} = \sup\left\{\sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\big[\Phi^{\mathrm{KR}}(\{j, j+1,\ldots,n\}) - \Phi^u(\{j, j+1,\ldots,n\})\big] : u\in\mathcal{U}\right\}.$$

PROOF. Part (a) is an immediate consequence of Identity (38) in Lemma 3. As for part (b), we first use (38) to obtain, for any policy $u \in \mathcal{U}$, the identity

$$Z^{\mathrm{KR}} - Z^u = \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1}) \big[(\Phi^{\mathrm{KR}}\{j, j+1, \ldots, n\}) - \Phi^u(\{j, j+1, \ldots, n\}) \big],$$

and then maximise both sides over $u \in \mathcal{U}$. □

We next apply Lemmas 3 and 4 to obtain upper bounds for the suboptimality gap of Klimov's rule, namely $Z^{\mathrm{KR}} - Z^{\min}$.

THEOREM 2. *The following relations hold*:

$$Z^{\mathrm{KR}} - Z^{\min} \leqslant Z^{\mathrm{KR}} - Z^{\mathrm{D}} \tag{43}$$

$$= \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\, \Phi^{\mathrm{KR}}(\{j, \ldots, n\}) \tag{44}$$

$$\leqslant \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\, \Phi(\{j, \ldots, n\}). \tag{45}$$

PROOF. Inequality (43) follows from (37), whereas Identity (44) follows from (38) in Lemma 3. Inequality (45) is a consequence of the definition of Φ in (29) and the fact that Klimov's rule gives priority to classes in $\{j, \ldots, n\}$ over those in $\{1, \ldots, j-1\}$ for all j. □

As a convenient shorthand we shall refer to the suboptimality bounds in Identity (44) and Inequality (45) as ϵ^{KR} and ϵ, respectively. Note that it is the bound ϵ that is recovered from Glazebrook and Garbe's (1999) analysis based on approximate conservation laws. For systems satisfying GCL, $\Phi \equiv 0$ so that $\epsilon = 0$. Hence, (45) implies the optimality of Klimov's rule in such cases. This is the main result of Bertsimas and Niño-Mora (1996). Note further that the bound ϵ^{KR} is tight in the special case when all nonzero Klimov indices are equal. To see this, suppose that, for some $S \subseteq \mathcal{N}$,

$$\gamma_j = \begin{cases} \gamma & \text{for } j \in S \\ 0 & \text{for } j \in S^c = \mathcal{N} \setminus S, \end{cases} \tag{46}$$

with $\gamma > 0$.

COROLLARY 1. *When all nonzero Klimov indices are equal, it follows that*

$$Z^{\mathrm{KR}} - Z^{\min} = \epsilon^{\mathrm{KR}}.$$

PROOF. Because ϵ^{KR} is defined by the right-hand side in Identity (44), it follows from (46) that

$$\epsilon^{\mathrm{KR}} = \gamma\, \Phi^{\mathrm{KR}}(S),$$

while from the form of Klimov's adaptive greedy algorithm we can conclude that

$$\gamma = c_j / V_j^S, \quad \text{for } j \in S,$$

and

$$c_j = 0, \quad \text{for } j \in S^c.$$

Hence, we see that Condition (46) implies that the objective is of the form

$$Z^u = \gamma \sum_{j \in S} V_j^S\, x_j^u.$$

This fact, together with the definition of Φ^u in (28), immediately yields the result. □

5. ANALYSIS OF KLIMOV'S RULE FOR THE PARALLEL-SERVER SYSTEM

In this section, we deploy the theoretical framework and results developed in §4 in support of our principal objective, namely, the assessment of Klimov's rule as a scheduling policy for the multiclass queueing network on parallel servers described in §2. In principle, the application of the material in §4 to the model of interest is a straightforward matter. As indicated in Equations (30)–(32), the work decomposition laws in Theorem 1 provide us with appropriate choices for matrix $\boldsymbol{V}$, performance vector $\boldsymbol{x}^u$, and set functions β and Φ^u. The space of admissible scheduling policies $\mathcal{U}$ is as outlined in §2. We utilise these choices within Lemma 4(b) and Theorem 2 to obtain, respectively, an exact expression for and bounds on $Z^{\mathrm{KR}} - Z^{\min}$, the suboptimality gap of Klimov's rule. It will assist the reader to refer back to the definition and interpretation of $b(S)$, $\Delta_{\mathrm{pr}}^u(S)$ and $\Delta_{\mathrm{id}}^u(S)$ given just before the start of §3.1. It is only in §5.1 that we make extensive use of the precise forms of $\Delta_{\mathrm{pr}}^u(S)$ and $\Delta_{\mathrm{id}}^u(S)$. Recall from (32) that, for each admissible policy u, the set function Φ^u takes the form

$$\Phi^u(S) = \Delta_{\mathrm{pr}}^u + \Delta_{\mathrm{id}}^u(S) - \Delta^*(S).$$

We can now express the suboptimality bound for Klimov's rule ϵ^{KR} given by Identity (44) in the previous section as the right-hand side in (47):

$$\begin{aligned} Z^{\mathrm{KR}} - Z^{\min} &\leqslant \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\, \Phi^{\mathrm{KR}}(\{j, \ldots, n\}) \\ &= \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1}) \big[\Delta_{\mathrm{pr}}^{\mathrm{KR}}(\{j, \ldots, n\}) \\ &\qquad + \Delta_{\mathrm{id}}^{\mathrm{KR}}(\{j, \ldots, n\}) - \Delta^*(\{j, \ldots, n\}) \big]. \end{aligned} \tag{47}$$

The reformulation of bound ϵ^{KR} given in (47) reveals its intuitive interpretation: the jth term of the sum, which gives ϵ^{KR} involves a weight $\gamma_j - \gamma_{j-1} \geqslant 0$, which depends on cost vector $\boldsymbol{c}$, and a difference $\Delta_{\mathrm{pr}}^{\mathrm{KR}}(\{j, \ldots, n\}) + \Delta_{\mathrm{id}}^{\mathrm{KR}}(\{j, \ldots, n\}) - \Delta^*(\{j, \ldots, n\}) \geqslant 0$. Note that the latter does not involve $\boldsymbol{c}$ and can be thought of as a measure of excess expected workload incurred by Klimov's rule because of inefficiencies in the handling of priorities and processing capacity over $\{j, \ldots, n\}$-customers.

In what follows, the performance objective $Z^{\min}_{\text{pooled}}$ achieved by an optimally scheduled single-server pooled resource, as discussed in §2.1, will play a key role in our analysis. We shall employ the representation of $Z^{\min}_{\text{pooled}}$ given in (13), namely

$$Z^{\min}_{\text{pooled}} = \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\, b(\{j, \ldots, n\}).$$

We now state and prove three identities relating $Z^{\min}_{\text{pooled}}$, with the objective $Z^u = \sum_{i=1}^n c_i x_i^u$ achieved by an arbitrary policy $u \in \mathcal{U}$, the lower bound Z^{D} defined in the previous section and the optimal objective value $Z^{\min}$. We shall make use of these results in establishing our approximate optimality results in Theorem 3.

LEMMA 5. *The following identities hold*:

(*a*) *for any admissible scheduling policy* $u \in \mathcal{U}$,

$$Z^u = Z^{\min}_{\text{pooled}} + \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1}) \left[\Delta^u_{\text{pr}}(S) + \Delta^u_{\text{id}}(S)\right]; \tag{48}$$

(*b*)

$$Z^{\min} - Z^{\min}_{\text{pooled}} = \inf \left\{ \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1}) \left[\Delta^u_{\text{pr}}(\{j, \ldots, n\}) + \Delta^u_{\text{id}}(\{j, \ldots, n\})\right] : \; u \in \mathcal{U} \right\}; \tag{49}$$

(*c*)

$$Z^{\text{D}} - Z^{\min}_{\text{pooled}} = \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\, \Delta^*(\{j, \ldots, n\}). \tag{50}$$

PROOF. (a) The result follows directly from Lemma 3 together with Identities (13), (30), and (32).

(b) This follows directly from part (a).

(c) This is a direct consequence of (13), (30), and (39). □

Note that because $\Delta^*(S) \geqslant 0$ for all S, it follows from the above that

$$Z^{\min}_{\text{pooled}} \leqslant Z^{\text{D}},$$

i.e., $Z^{\min}_{\text{pooled}}$ is a weaker lower bound than Z^{D} for the optimal problem value $Z^{\min}$. Furthermore, note that the positive gap $Z^{\min} - Z^{\min}_{\text{pooled}}$ in Lemma 5(b) may be thought of as the parallel-server system's *performance degradation from the pooled ideal*, or the *cost of parallelism*.

We shall not take this general discussion any further here. Many questions of interest are prompted by the above discussion. For example, the results obtained for the batch case (where there are no arrivals but a finite number of jobs that have to be scheduled) by Weiss (1990, 1992, 1995) suggest that we might expect $Z^{\text{KR}} - Z^{\min}$ to be much smaller in general than $Z^{\min} - Z^{\min}_{\text{pooled}}$. From the previous discussion, this conjecture could be explored via appropriate study of the quantities $\Delta^{\text{KR}}_{\text{pr}}$, $\Delta^{\text{KR}}_{\text{id}}$, and Δ^u_{pr}, Δ^u_{id}, for $u \in \mathcal{U}$. This will be the subject of future work.

What we shall establish in §5.1 is the result that the gap $Z^{\text{KR}} - Z^{\min}_{\text{pooled}}$, which measures the performance degradation of Klimov's rule due to parallelism, is uniformly bounded above by a constant with respect to (i) external arrival rates, as long as they vary within system capacity and (ii) the number of servers. From (i) it will follow that the corresponding relative gap vanishes in heavy traffic (as external arrival rates approach system capacity). Because the suboptimality gap of Klimov's rule, $Z^{\text{KR}} - Z^{\min}$, is bounded above by $Z^{\text{KR}} - Z^{\min}_{\text{pooled}}$, the corresponding results extend to the former.

5.1. Approximate and Heavy-Traffic Optimality of Klimov's Rule

Our concern in this section will be to develop simple and interpretable bounds for the suboptimality gap of Klimov's rule, $Z^{\text{KR}} - Z^{\min}$, and for its performance degradation due to parallelism, $Z^{\text{KR}} - Z^{\min}_{\text{pooled}}$, expressed in terms of model parameters. We will use these bounds to establish the asymptotic optimality of Klimov's rule in an appropriate heavy-traffic limit. The bounds we develop in the course of the analysis will certainly not be the tightest available, but they will be sufficient for our purposes. In particular, as mentioned above, our bounds imply that the gap $Z^{\text{KR}} - Z^{\min}_{\text{pooled}}$, and hence, $Z^{\text{KR}} - Z^{\min}$ remains uniformly bounded above by a constant with respect to external arrival rates, as long as they vary within system capacity, and with respect to the number of servers.

We begin in Lemmas 6 and 7 by developing simple bounds on the crucial last two terms in the Work Decomposition Law (14). We require the notation

$$V^{\mathcal{N}}_{\max} = \max_{j \in \mathcal{N}} V^{\mathcal{N}}_j,$$

and

$$(\mu V)^{\mathcal{N}}_{\max} = \max_{j \in \mathcal{N}} \mu_j V^{\mathcal{N}}_j.$$

The next result gives an upper bound on the priority term $\Delta^u_{\text{pr}}(S)$.

LEMMA 6. *For any admissible scheduling policy* u *that gives preemptive priority to* S*-customers over* S^c*-customers,*

$$\Delta^u_{\text{pr}}(S) \leqslant \rho\, V^{\mathcal{N}}_{\max} (\mu V)^{\mathcal{N}}_{\max} 1\{m > 1\}, \quad \textit{for } S \subset \mathcal{N}.$$

PROOF. Inspection of the $\Delta^u_{\text{pr}}(S)$ term in (14) yields immediately the conclusion that in the case $m = 1$, it becomes 0 under any policy u that gives preemptive priority to S-customers.

Suppose now $m \geqslant 2$. Let B_S (respectively, B_{S^c}) denote the number of S-customers (respectively, S^c-customers) in service. Note that because u gives preemptive priority to S-customers, it follows that with probability 1,

$$B_{S^c} \sum_{j \in S} L_j = B_{S^c} B_S \leqslant (m - B_S) B_S. \tag{51}$$

By standard results, we have

$$\rho(S) = E_u[B_S] \leqslant \sqrt{E_u\big[(B_S)^2\big]},$$

and hence taking expectations through (51), we conclude that

$$E_u\left[B_{S^c}\sum_{j\in S} L_j\right] \leqslant \rho(S)\,[m-\rho(S)].$$

The result now follows from the form of $\Delta^u_{\rm pr}(S)$ in (14) and from the facts that $\rho^0(S) \leqslant \rho(S) \leqslant \rho$, $\mu_i V_i^S \leqslant (\mu V)^{\mathcal{N}}_{\max}$ and $V_j^S \leqslant V^{\mathcal{N}}_{\max}$ for each i, j. □

A very similar argument to that in Lemma 6 yields the upper bound on idleness term $\Delta^u_{\rm id}(S)$ stated in the following result. Recall that admissible policies are required to be nonidling. This is necessary for the bound given next to hold.

LEMMA 7. *For any admissible scheduling policy u,*

$$\Delta^u_{\rm id}(S) \leqslant \rho\, V^{\mathcal{N}}_{\max} 1\{m>1\}, \quad \textit{for } S \subseteq \mathcal{N}.$$

Major simplifications result when we consider the version of our model without feedback (where $p_{ij} = 0$ for all i, j). We write

$$\left(\frac{1}{\mu}\right)_{\max} = \max_{j\in\mathcal{N}}\left(\frac{1}{\mu_j}\right).$$

LEMMA 8 (NO-FEEDBACK CASE). *For any admissible scheduling policy u that gives preemptive priority to S-customers over S^c-customers,*

$$\Delta^u_{\rm pr}(S) + \Delta^u_{\rm id}(S) \leqslant \rho\left(\frac{1}{\mu}\right)_{\max} 1\{m>1\}, \qquad \textit{for } S \subseteq \mathcal{N}.$$

PROOF. Fix $S \subseteq \mathcal{N}$. In the no-feedback case the definition of the matrix $\boldsymbol{V}$ given in §2 yields

$$V_i^S = \frac{1}{\mu_i}, \qquad \text{for } i \in \mathcal{N}. \tag{52}$$

We now observe that in this case the performance over S-customers of a policy that gives preemptive priority to S-customers is identical to that obtained in a *reduced system* in which there are only S-customers. Applying Work Decomposition Law (14) to this reduced system and using (52) yields

$$\Delta^u_{\rm pr}(S) + \Delta^u_{\rm id}(S) = \frac{\sum_{j\in S}\frac{1}{\mu_j}E_u[(\#\text{ servers not busy with } S\text{-customers})\,L_j]}{m-\rho(S)}. \tag{53}$$

From this point, the calculation follows closely that of Lemma 6. We omit the details. □

We shall next use the bounds in Lemmas 6–8 together with the results of §4 and the beginning of this section to derive some simple suboptimality bounds for Klimov's rule for our multiclass queueing system on parallel servers. We deal with the general model (with feedback) first, and retain the customer numbering in which $\gamma_1 \leqslant \gamma_2 \leqslant \cdots \leqslant \gamma_n$. We shall further use the notation $Z^{c\mu}$ to denote the performance objective Z^u in the no-feedback case under the $c\mu$ rule (where Klimov's indices reduce to $\gamma_j = c_j\mu_j$, for all j), and write

$$(c\mu)_{\max} = \max_{1\leqslant j\leqslant n} c_j\mu_j.$$

THEOREM 3 (APPROXIMATE OPTIMALITY OF KLIMOV'S RULE). *The following inequalities hold*:

(*a*)

$$\begin{aligned} Z^{\rm KR} - Z^{\min} &\leqslant Z^{\rm KR} - Z^{\min}_{\rm pooled} \\ &\leqslant \rho\left[\gamma_n V^{\mathcal{N}}_{\max} + (\gamma_n-\gamma_1)(\mu V)^{\mathcal{N}}_{\max} V^{\mathcal{N}}_{\max}\right] \\ &\quad\cdot 1\{m>1\}; \end{aligned} \tag{54}$$

(*b*) *In the no-feedback case, where Klimov's rule reduces to the $c\mu$ rule, we have*

$$\begin{aligned} Z^{c\mu} - Z^{\min} &\leqslant Z^{c\mu} - Z^{\min}_{\rm pooled} \\ &\leqslant \rho(c\mu)_{\max}\left(\frac{1}{\mu}\right)_{\max} 1\{m>1\}. \end{aligned} \tag{55}$$

PROOF. (a) The inequality $Z^{\rm KR} - Z^{\min} \leqslant Z^{\rm KR} - Z^{\min}_{\rm pooled}$ follows from Lemma 5(b). As for Inequality (54), we have

$$\begin{aligned} \Phi(S) &\leqslant \sup\big\{\Delta^u_{\rm pr}(S) + \Delta^u_{\rm id}(S) : \\ &\qquad u \text{ gives preemptive priority to } S\text{-customers}\big\} \\ &\leqslant \begin{cases} \rho\big[(\mu V)^{\mathcal{N}}_{\max} + 1\big] V^{\mathcal{N}}_{\max} 1\{m>1\} & \text{if } S \subset \mathcal{N} \\ \rho V^{\mathcal{N}}_{\max} 1\{m>1\} & \text{if } S = \mathcal{N}, \end{cases} \end{aligned}$$

where the first inequality uses (33) and the subsequent inequality combines the results in Lemmas 6 and 7. The result then follows immediately from Inequality (45) in Theorem 2.

(b) Inequality (55) follows by applying the same argument as in part (a), but using Lemma 8 to bound the $\Phi(S)$ terms. □

Corollary 2 is an immediate consequence of the above approximate optimality result.

COROLLARY 2. *The gaps $Z^{KR} - Z^{min}$ and $Z^{KR} - Z^{min}_{pooled}$ are uniformly bounded above with respect to* (*i*) *external arrival rates vector α, as long as it stays within system capacity ($\rho < m$); and* (*ii*) *the number m of servers.*

PROOF. From Bound (54) in Theorem 3 it follows that as the vector α of external arrival rates varies within system capacity we have $\rho < m$, and hence

$$\begin{aligned} Z^{\rm KR} - Z^{\min} &\leqslant Z^{\rm KR} - Z^{\min}_{\rm pooled} \\ &\leqslant m\big[\gamma_n V^{\mathcal{N}}_{\max} + (\gamma_n-\gamma_1)(\mu V)^{\mathcal{N}}_{\max} V^{\mathcal{N}}_{\max}\big] 1\{m>1\}. \end{aligned}$$

The result follows by noting that the last bound in the previous equation does not depend on α (see the corresponding parameter definitions in §2).

It follows by noting that for any number $m \geqslant 2$ of servers, the last bound in the string of inequalities

$$Z^{\mathrm{KR}} - Z^{\min} \leqslant Z^{\mathrm{KR}} - Z^{\min}_{\mathrm{pooled}} \leqslant \rho\big[\gamma_n V^{\mathcal{N}}_{\max} + (\gamma_n - \gamma_1)\,(\mu V)^{\mathcal{N}}_{\max}\, V^{\mathcal{N}}_{\max}\big]$$

does not depend on m. □

We next investigate the asymptotic performance of Klimov's rule in the heavy-traffic limit, as $\rho \nearrow m$. To this end we consider a sequence of models in which only the external arrival rates vary, while all other parameters remain fixed. In particular, we suppose that the vector of external arrival rates $\boldsymbol{\alpha} = (\alpha_j)_{j\in\mathcal{N}}$ varies according to a convergent sequence $\{\boldsymbol{\alpha}^k\}_{k=1}^{\infty}$ with limit $\boldsymbol{\alpha}^*$ such that, in an obvious notation,

$$\lim_{k\to\infty} \rho(\boldsymbol{\alpha}^k) = \rho(\boldsymbol{\alpha}^*) = m,$$

with

$$\rho(\boldsymbol{\alpha}^k) < m, \qquad \text{for each } k \geqslant 1.$$

We shall write, in the same way, $Z^{\mathrm{KR}}(\boldsymbol{\alpha}^k)$, $Z^{\min}(\boldsymbol{\alpha}^k)$, and $Z^{\min}_{\mathrm{pooled}}(\boldsymbol{\alpha}^k)$, for $k \geqslant 1$.

It is a simple consequence of the uniform boundedness result in Corollary 2(i), which we deduced from Theorem 3 that both the relative suboptimality gap of Klimov's rule, and its relative performance degradation because of parallelism, vanish as external arrival rates approach network capacity. We assume now for convenience that $\boldsymbol{c} \neq \mathbf{0}$. The new notation $V^{\mathcal{N}}_{\min}$, $(c\mu)_{\min}$, $(1/\mu)_{\min}$ used below has the obvious meaning, consistent with that introduced above for, e.g., $V^{\mathcal{N}}_{\max}$.

COROLLARY 3 (HEAVY-TRAFFIC OPTIMALITY OF KLIMOV'S RULE). *We have* (*a*)

$$\begin{aligned}\frac{Z^{\mathrm{KR}}(\boldsymbol{\alpha}^k) - Z^{\min}(\boldsymbol{\alpha}^k)}{Z^{\min}(\boldsymbol{\alpha}^k)} &\leqslant \frac{Z^{\mathrm{KR}}(\boldsymbol{\alpha}^k) - Z^{\min}_{\mathrm{pooled}}(\boldsymbol{\alpha}^k)}{Z^{\min}_{\mathrm{pooled}}(\boldsymbol{\alpha}^k)} \\ &\leqslant \big(m - \rho(\boldsymbol{\alpha}^k)\big) \\ &\quad \cdot \frac{\gamma_n V^{\mathcal{N}}_{\max} + (\gamma_n - \gamma_1)\,(\mu V)^{\mathcal{N}}_{\max}\, V^{\mathcal{N}}_{\max}}{\gamma_1 V^{\mathcal{N}}_{\min}} \\ &\quad \cdot 1\{m > 1\} \\ &= O\big(m - \rho(\boldsymbol{\alpha}^k)\big) \to 0 \quad \textit{as } k \to \infty;\end{aligned}$$

(*b*) *in the no-feedback case, where Klimov's rule reduces to the* $c\mu$ *rule, we have*

$$\begin{aligned}\frac{Z^{c\mu}(\boldsymbol{\alpha}^k) - Z^{\min}(\boldsymbol{\alpha}^k)}{Z^{\min}(\boldsymbol{\alpha}^k)} &\leqslant \frac{Z^{c\mu}(\boldsymbol{\alpha}^k) - Z^{\min}_{\mathrm{pooled}}(\boldsymbol{\alpha}^k)}{Z^{\min}_{\mathrm{pooled}}(\boldsymbol{\alpha}^k)} \\ &= \big(m - \rho(\boldsymbol{\alpha}^k)\big)\, \frac{(c\mu)_{\max}\left(\frac{1}{\mu}\right)_{\max}}{(c\mu)_{\min}\left(\frac{1}{\mu}\right)_{\min}} \\ &\quad \cdot 1\{m > 1\} \\ &= O\big(m - \rho(\boldsymbol{\alpha}^k)\big) \to 0 \quad \textit{as } k \to \infty.\end{aligned}$$

PROOF. (a) The first inequality follows from the fact that $Z^{\min}_{\mathrm{pooled}}(\boldsymbol{\alpha}^k) \leqslant Z^{\min}(\boldsymbol{\alpha}^k) \leqslant Z^{\mathrm{KR}}(\boldsymbol{\alpha}^k)$. In the second inequality, we note that (11) and (13), together with the fact that γ_1 is the smallest index, yield

$$Z^{\min}_{\mathrm{pooled}}(\boldsymbol{\alpha}^k) \geqslant \gamma_1\, \frac{\sum_{j=1}^{n} \rho_j(\boldsymbol{\alpha}^k)\, V^{\mathcal{N}}_j}{m - \rho(\boldsymbol{\alpha}^k)} \geqslant \gamma_1\, V^{\mathcal{N}}_{\min}\, \frac{\rho(\boldsymbol{\alpha}^k)}{m - \rho(\boldsymbol{\alpha}^k)}.$$

The second inequality in the statement of Corollary 3(a) is now immediate from Theorem 3(a), and the remainder of the result follows easily. Part(b) follows by applying the same line of argument as in part(a), but using the bounds in Theorem 3(b). □

Note the presence of the factor $1\{m > 1\}$ in Theorem 3 and Corollary 3. This ensures that the well-known optimality of the corresponding index rule in the single-server case is recovered from our analyses.

5.2. Numerical Investigation of ϵ^{KR}

As explained at the beginning of §5.1, the goal was to develop bounds on the suboptimality gaps of interest that were simple and adequate to be deployed in proving the theoretical results above. If we return to Theorem 2 in §4, we see that all the bounds developed in §5.1 in fact utilised Inequality (45), rather than the tighter Inequality (44). However, the bound ϵ^{KR} given by the r.h.s. of (44) and (47) has intuitive appeal (see the comments following (47)) and is known to be tight under simply stated conditions (see Corollary 1). To investigate the tightness of ϵ^{KR} more widely, computational experiments were conducted. The system studied was a two-class M/M/2 queue with no customer feedback. The vectors of costs and service rates are given, respectively, by (c_1, c_2) and (μ_1, μ_2). Arrival rates α_1 and α_2 are always set equal to 1, as is c_1. The Klimov indices in this case are $c_i\mu_i$, $i = 1, 2$ with Klimov's rule now the celebrated $c\mu$ rule. Hence, we introduce the quantity

$$K = (c_1\mu_1 - c_2\mu_2)/(c_1\mu_1 + c_2\mu_2)$$

as a natural measure of the extent to which the condition of equal indices (which guarantees the tightness of $\epsilon^{c\mu}$) fails; see Corollary 1. For each (ρ, K)-pair with traffic intensity ρ drawn from the set $\{0.5, 0.8, 1.0, 1.2, 1.4\}$ and K from $\{0, 0.025, 0.05, 0.075, 0.100, 0.150, 0.200\}$, we investigated 50 examples where

$$\frac{1}{\mu_1} \sim U(0.2, \rho - 0.2) \text{ and } \frac{1}{\mu_2} = \rho - \frac{1}{\mu_1},$$

with c_2 set equal to

$$\mu_1(1-K)/[\mu_2(1+K)].$$

For each problem studied, approximate optimal costs ($Z^{\min}$) were computed by truncating the state space of queue lengths and using the Value Iteration Algorithm of dynamic programming. Initially, the truncation was set at

Table 1. Computational results.

		ρ									
		0.5		0.8		1.0		1.2		1.4	
K	0.000	0.202	0.026	0.668	0.095	0.820	0.122	0.910	0.138	0.843	0.132
		0.202	0.026	0.668	0.095	0.820	0.122	0.910	0.138	0.843	0.132
	0.025	0.066	0.014	0.373	0.070	0.493	0.089	0.539	0.095	0.517	0.090
		0.154	0.026	0.577	0.094	0.765	0.123	0.855	0.135	0.830	0.131
	0.050	0.020	0.006	0.222	0.048	0.306	0.063	0.343	0.069	0.342	0.066
		0.142	0.026	0.535	0.092	0.711	0.119	0.795	0.132	0.773	0.127
	0.075	0.010	0.001	0.149	0.032	0.229	0.045	0.279	0.052	0.296	0.052
		0.162	0.024	0.635	0.089	0.856	0.118	0.968	0.132	0.949	0.128
	0.100	0.000	0.000	0.089	0.022	0.148	0.033	0.184	0.040	0.202	0.042
		0.153	0.023	0.601	0.086	0.811	0.113	0.919	0.126	0.903	0.123
	0.150	0.000	0.000	0.017	0.006	0.047	0.014	0.075	0.021	0.125	0.029
		0.118	0.021	0.476	0.083	0.650	0.112	0.741	0.126	0.805	0.127
	0.200	0.000	0.000	0.003	0.002	0.015	0.005	0.025	0.009	0.068	0.017
		0.088	0.019	0.361	0.075	0.495	0.106	0.570	0.114	0.920	0.125

25 customers for each class. This limit was increased until the difference between successive calculations was negligible. A similar approach was used for the expected cost under the $c\mu$-rule ($Z^{c\mu}$) and the suboptimality bound $\epsilon^{c\mu}$.

The results of the study are given in Table 1, where the table entries corresponding to each pair (K, ρ) are in the form:

$$\frac{A_1 \mid A_2}{A_3 \mid A_4},$$

where

$$A_1 = \widehat{E}\left[\frac{Z^{c\mu} - Z^{\min}}{Z^{\min}}\right] \times 100;$$

$$A_2 = \hat{\sigma}\left[\frac{Z^{c\mu} - Z^{\min}}{Z^{\min}}\right] \times 100;$$

$$A_3 = \widehat{E}\left[\frac{\epsilon^{c\mu}}{Z^{\min}}\right] \times 100;$$

$$A_4 = \hat{\sigma}\left[\frac{\epsilon^{c\mu}}{Z^{\min}}\right] \times 100.$$

In the above expressions, $\widehat{E}$ and $\hat{\sigma}$ denote the sample mean and sample standard deviation, respectively.

From Table 1 we observe that the $c\mu$ rule exhibits an excellent level of performance across the range of problems investigated. Its relative suboptimality gap never exceeds 1%. As indicated by Corollary 1, the bound $\epsilon^{c\mu}$ is tight when $K = 0$. As expected, the quality of the bound deteriorates with increasing K, although never exceeds 1% of the optimal cost on average. We note that it follows from the results in §5.1 that the ratio $\epsilon^{c\mu}/Z^{\min}$ tends to zero in the heavy traffic limit $\rho \nearrow 2$.

6. CONCLUDING REMARKS

We have analyzed a simple heuristic index policy that extends Klimov's classical solution for the single-server case to the general parallel-server model, presenting closed-form suboptimality bounds that imply its asymptotic optimality in a heavy-traffic limit. Ideas that emerge from our analysis include the following: (1) Understanding of a simple single-server system has yielded useful insights into the performance of its more complex parallel-server counterpart; (2) understanding the fundamental laws of a complex parallel-server model (flow balance and work decomposition) has yielded a key to its analysis; (3) investigating strong linear programming relaxations of a complex stochastic optimization problem has yielded an approximate and asymptotic analysis of a heuristic, which had resisted traditional approaches. We believe these ideas, which guided our approach, should prove fruitful for addressing other complex stochastic optimization problems. We refer the reader back to the discussion in the paragraph following Lemma 5 for indications of further work to be done on the current model.

In a companion paper (see Glazebrook and Niño-Mora 1999) we carry out a corresponding analysis of priority index rules for scheduling Markovian multiclass queueing networks with multiple service stations. Although such rules are known to perform poorly in general for the latter type of networks, in that paper we present suboptimality bounds under appropriate light-traffic conditions.

It should be remarked that while Klimov's optimal solution for the single-server model applied to multiclass $M/G/1$ networks, i.e., it was valid under general service time distributions, our analysis requires the latter to be exponential. Extending our approach to a model with general service time distributions and nonpreemptive policies would require the development and application of work decomposition laws for such a model. Carrying out this extension remains a challenging problem.

ACKNOWLEDGMENT

The authors thank Phil Ansell for carrying out the computational experiments reported in §5.2. They are grateful to the associate editors and referees for their stimulus to further work and their assistance in improving the presentation of the paper. The research of Glazebrook was supported by the Engineering and Physical Sciences Research Council by means of grant nos. GR/K03043 and GR/M09308. Part

of this work was carried out during Niño-Mora's stay at the Center for Operations Research and Econometrics (CORE) of the Université Catholique de Louvain, Belgium. His research was supported by a CORE Research Fellowship, by European Commission Individual Marie Curie Postdoctoral Fellowship no. ERBFMBICT961480, and by Spain's National R & D Program Grant CICYT TAP98-0229. A preliminary version of this paper was published in the Proceedings of the 5th European Symposium on Algorithms, ESA 97, Graz, Austria.

REFERENCES

Bertsimas, D., J. Niño-Mora. 1996. Conservation laws, extended polymatroids and multi-armed bandit problems: a polyhedral approach to indexable systems. *Math. Oper. Res.* **21** 257–306.

——, ——. 1999a. Optimization of multiclass queueing networks with changeover times via the achievable region approach: Part I, the single-station case. *Math. Oper. Res.* **24** 306–330.

——, ——. 1999b. Optimization of multiclass queueing networks with changeover times via the achievable region approach: Part II, the multi-station case. *Math. Oper. Res.* **24** 331–361.

——, I. C. Paschalidis, J. N. Tsitsiklis. 1994. Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance. *Ann. Appl. Probab.* **4** 43–75.

Boxma, O. J. 1989. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Syst.* **5** 185–214.

Burke, P. J. 1956. The output of a queueing system. *Oper. Res.* **4** 699–704.

Coffman, E., I. Mitrani. 1980. A characterization of waiting time performance realizable by single server queues. *Oper. Res.* **28** 810–821.

Federgruen, A., H. Groenevelt. 1988. Characterization and optimization of achievable performance in general queueing systems. *Oper. Res.* **36** 733–741.

Finch, P. D. 1959. On the distribution of queue size in queueing problems. *Acta Math. Hungar.* **10** 327–336.

Gelenbe, E., I. Mitrani. 1980. *Analysis and Synthesis of Computer Systems*. Academic Press, London.

Gittins, J. C., D. M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. Gani, J., Sarkadi, K., and Vince, I., eds., *Progress in Statistics: European Meeting of Statisticians, Budapest, 1972*, North-Holland, Amsterdam, 241–266.

Glazebrook, K. D., R. Garbe. 1999. Almost optimal policies for stochastic systems which almost satisfy conservation laws. *Ann. Oper. Res.* **92**, 19–43.

——, J. Niño-Mora. 1997. Scheduling multi-class queueing networks on parallel servers: approximate and heavy-traffic optimality of Klimov's rule. R. Burkard and G. Woeginger, (eds.), *Algorithms—ESA 97*, volume 1284 of *Spinger Lecture Notes in Computer Science.* 232–245.

——, ——. 1999. A linear programming approach to stability, optimization and performance analysis for Markovian queueing networks. *Ann. Oper. Res.* **92**, 1–18.

Harrison, J. M. 1998. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Ann. Appl. Probab.* **8** 822–848.

Kelly, F. P., C. N. Laws. 1993. Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Syst. Theory Appl.* **13** 47–86.

Klimov, G. P. 1974. Time sharing service systems I. *Theory Probab. Appl.* **19** 532–551.

——. 1978. Time sharing service systems II. *Theory Probab. Appl.* **23** 314–321.

Kumar, S., P. R. Kumar. 1994. Performance bounds for queueing networks and scheduling policies. *IEEE Trans. Autom. Control* **39** 1600–1611.

Papangelou, F. 1972. Integrability of expected increments of point processes and a related random change of time scale. *Trans. Amer. Math. Soc.* **165** 483–506.

Ross, K. W., D. D. Yao. 1989. Optimal dynamic scheduling in Jackson networks. *IEEE Trans. Aut. Control* **34** 47–53.

Shanthikumar, J. G., D. D. Yao. 1992. Multi-class queueing systems: polyhedral structure and optimal scheduling control. *Oper. Res.* **40** 293–299.

Smith, W. E. 1956. Various optimizers for single stage production. *Naval Res. Logist. Quart.* **3** 59–66.

Tsoucas, P. 1991. The region of achievable performance in a model of Klimov. Technical Report RC16543, IBM T.J. Watson Research Center, Yorktown Heights, NY.

Weiss, G. 1990. Approximation results in parallel machines stochastic scheduling. Special Volume on Production Planning and Scheduling M. Queyranne (ed.), *Ann. Oper. Res.* **26** 195–242.

——. 1992. Turnpike optimality of Smith's rule in parallel machines stochastic scheduling. *Math. Oper. Res.* **17** 255–270.

——. 1995. On almost optimal priority rules for preemptive scheduling of stochastic jobs on parallel machines. *Adv. Appl. Probab.* **27** 821–839.