Adv. Appl. Prob. **53,** 463–483 (2021)
*doi:*10.1017/apr.2020.63

KELLY AND JACKSON NETWORKS WITH INTERCHANGEABLE, COOPERATIVE SERVERS

CHIA-LI WANG,* *National Dong Hwa University*
RONALD W. WOLFF,** *University of California, Berkeley*

Abstract

In open Kelly and Jackson networks, servers are assigned to individual stations, serving customers only where they are assigned. We investigate the performance of modified networks where servers cooperate. A server who would be idle at the assigned station will serve customers at another station, speeding up service there. We assume *interchangeable* servers: the service rate of a server at a station depends only on the station, not the server. This gives *work conservation*, which is used in various ways. We investigate three levels of server cooperation, from *full cooperation*, where all servers are busy when there is work to do anywhere in the network, to *one-way cooperation*, where a server assigned to one station may assist a server at another, but not the converse. We obtain the same stability conditions for each level and, in a series of examples, obtain substantial performance improvement with server cooperation, even when stations before modification are moderately loaded.

Keywords: Stability; positive recurrent; work conservation; insensitivity; flexible servers

2010 Mathematics Subject Classification: Primary 60K25
Secondary 60K20

1. Introduction

An open queueing network with Poisson arrivals at one or more stations, independent exponential service, and *general routes* of customers through the network is called a Kelly network [6]. Kelly networks are a generalized model of Jackson networks [5], where customers have Markovian paths. We consider Kelly and Jackson networks with k single-server stations, where server i is assigned to station i, and only serves customers there. (Kelly and Jackson networks may be closed.)

We will treat *modified* networks where servers cooperate. When station i is empty (of customers), or, in some cases, even when it is busy, server i may assist a busy server at another station, increasing the service rate there. Intuitively, we expect server cooperation to improve system performance. We will model server cooperation and obtain results about the extent of any improvement over *original* networks.

Received 9 July 2018; revision received 27 August 2020.
* Postal address: Department of Applied Mathematics, National Dong Hwa University, Hualien 974, Taiwan, ROC. Email address: cwang@gms.ndhu.edu.tw
** Postal address: Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA.

We measure system performance by the reduction in the average number of customers in system. Of course, from Little's law, there is a comparable reduction in the average waiting time in system.

Since Jackson, there has been a substantial literature on queueing networks under weaker (or different) stochastic assumptions, with research monographs by Bramson [3] and Meyn [8]; the former is primarily concerned with stability, the latter with control. Neither considers models where servers cooperate.

For (stable) original Kelly networks, the joint stationary distribution of the number of customers in system at each station has a well-known and simple *product form*. For modified Kelly networks, the stationary distribution is unknown and will not in general have a simple form. Analyzing the performance of modified networks is difficult. Most existing results on these networks are limited to Jackson networks with $k = 2$ stations and are about what are called *asymptotic decay rates*. See [9] for results of this nature and discussion of earlier literature.

In that literature, the increase in service rate at a station where two servers cooperate is treated as an arbitrary parameter. Our approach is different. We assume that servers are *interchangeable*: the service rate of a server depends only on the station, not the server. For an original network, the total remaining work to be performed on every customer at each station is defined in Section 2. When servers are interchangeable and the network is modified, these quantities may change for Kelly networks, but do not change for Jackson networks.

Flexible servers is a related term usually referring to servers who can perform more than one task. See, for example, [2]. In the network context, they can serve customers at more than one station. They may or may not be interchangeable.

In Section 2, we review some basic properties of Kelly networks with single-server first-in-first-out (FIFO) stations. In Subsection 2.1 we introduce the modified networks with interchangeable servers that are determined by some stationary policy, and in Subsection 2.2 we derive special results for Jackson networks.

In Section 3, we treat modified networks under full cooperation (FC), which means that all servers are busy when there is work to do anywhere in the network. We define the total-work-in-system process and show under FC that when servers are interchangeable, this process is invariant to how the network is modified. This is *work conservation*; it enables us to obtain a simple if-and-only-if condition for the entire modified network to be stable and an expression for the time average of the (total) work in system (TAWS) for the modified networks. With a scale factor, the TAWS for the modified network is the same as that for a corresponding M/G/1 queue. When the original network is stable, we present examples where server cooperation reduces the average number of customers in system by a large factor. Of course, when the original network is unstable, the improvement may be much larger.

In Section 4, we make an assumption which we call AA: we assume that a customer may be served by only one server at a time, and that all servers are busy when the number of customers in system is at least as large as the number of servers. The method of analysis and results in Section 3 provide the basis for the analysis and results in this section. In particular, TAWS under FC is a lower bound for the corresponding quantity under AA. It is then shown that the stability condition (12) for FC also holds under AA. However, we do not have an expression for TAWS under AA as we do under FC. Observing that the rules for how many servers are busy under AA are the same as for a standard M/G/k queue, we propose an approximation for TAWS. Using this approximation, examples similar to those in Section 3 may be investigated under AA.

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

In Section 5, we assume that in the original network, some stations are stable and some are unstable, and we have one-way cooperation (OWC). Under OWC, servers assigned to stable stations may assist servers assigned to unstable stations, but not the converse. We show that the modified networks in Section 5, restricted to what we call preemption, are stable under the same condition we found for FC and AA.

A tandem queue of single-server stations was analyzed in [11], where there is only a single server, who allocates a fraction of the service capacity to each station. As the service rate at each station depends on the station, not the server, the single server is in effect interchangeable. Work conservation was a key property in the analysis. A large number of related papers referenced in [11] (in particular, see [10]) analyze this model under particular stochastic assumptions, without either formally recognizing the interchangeable property or making use of work conservation. For a different purpose, the authors of [1] considered modified Jackson networks where the service rate at a station either depends only on the station (interchangeable servers) or only on the server. They did not make use of work conservation.

In another relevant study, [7] compared the performance of a Jackson network to a system where the queues at each station are pooled into a single queue, and the servers are pooled into a single server with service rate equal to the sum of the service rates at the individual stations in the Jackson network (complete pooling). This results in a system where all customers are served at a single-server FIFO queue where the service time of a customer is the sum of the service times the same customer would have at the various stations visited in the Jackson network. It is easily shown that in some cases, for example that of tandem queues, the single-server queue performs much better than the Jackson network. The authors of [7] also show and explain cases where the single-server queue performs much worse.

In our approach, we don't collapse the network; servers move around. Because, under FC, all servers are busy when there is any work to do, our work-in-system process is the same as that for the single-server queue in [7]. However, because of our network structure, other quantities are quite different.

2. Kelly networks

We consider an open Kelly network with k single-server stations and R customer routes, where the arrival processes of customers on route r are independent and Poisson at rate λ_r, $r = 1, \ldots, R$, and we let the combined arrival rate of routes be $\lambda_T = \lambda_1 + \cdots + \lambda_R$. Note that R may be infinite.

Customers on route r visit the stations in the network in some finite deterministic sequence,

$$v(r, 1), v(r, 2), \ldots, v(r, e(r)),$$

before they leave the system. We say a customer on route r visits station i on *stage* s if $v(r, s) = i$, for $s = 1, 2, \ldots, e(r)$. We assume that every station is visited on at least one route. When necessary, we use station 0 to denote 'out of the system'. Because R may be infinite, a Jackson network is a special case.

Service times at station i are independent and identically distributed (i.i.d.) exponential at rate μ_i, $i = 1, \ldots, k$. They do not depend on route information. We assume first-in-first-out (FIFO); Kelly's formulation includes FIFO as a special case, and is not limited to single-server stations. Under FIFO, arrivals at each station join the end of the queue, the customer at the front of the queue is served, and on service completion, the customers that remain in the queue each move up one position.

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

By letting the state of the system contain the information of route, stage, and *position* (in the queue) of every customer in the system, a Kelly network is an irreducible (continuous-time) Markov chain.

Let $\#(r, i)$ be the number of times a route-r customer visits station i, and define Λ_i as

$$\Lambda_i = \sum_{r \in R} \#(r, i)\lambda_r, \qquad i = 1, 2, \ldots, k. \tag{1}$$

We require that $\Lambda_i < \infty$ for all i, which implies that $\lambda_T < \infty$. For a stable (positive recurrent) network, where all arrivals are eventually served, Λ_i is the *composite* (external plus internal) arrival rate at station i.

Note that the Λ_i do not depend on the service rates (of course, having stability does depend on the service rates).

Because we will be comparing various quantities for original and modified networks, it will be important to distinguish the one from the other. We will do this with the superscripts ^{o} and ^{m}. For example, the average numbers of customers at station i in the original and some corresponding modified network we denote by L_i^o and L_i^m, respectively.

Similarly, let $L_T^o = L_1^o + \ldots + L_k^o$ and $L_T^m = L_1^m + \ldots + L_k^m$ be the average number of customers in system in the original and modified networks, respectively. Our goal is to reduce L_T^m.

For an original stable Kelly network, let N_i denote the stationary number of customers at station i. Below we state an important result (Corollary 3.4 on p. 63 of [6]).

Corollary 1. *Given $N_i = n$, the probability that a route-r, stage-s customer is in position $l \le n$ at station i is λ_r/Λ_i, provided that $v(r, s) = i$.*

Let W_i^o be the *total remaining work* (TRW) to be performed on any of the N_i customers at station i, that is, the remaining service time at station i plus the sum of the service times on this customer at all stations to be visited until departure. (From Corollary 1, they all have the same distribution.) Note that in a Kelly network, the TRW of any customer at station i depends on his route and the present stage.

From Corollary 1, the expected TRW (ETRW) to be performed on any customer at station i is

$$w_i^o = \sum_{r=1}^{R} \frac{\lambda_r}{\Lambda_i} \sum_{s=1}^{e(r)} I\{v(r, s) = i\} \sum_{j=s}^{e(r)} \mathbb{E}[S_{v(r,j)}], \tag{2}$$

where $S_{v(r,j)}$ is the service time of a customer on route r at stage j.

In [6], Kelly defines $\phi_i(l)$ to be the service rate provided at station i when l customers are there, and he shows (see p. 61 of [6]) that the network is stable if and only if

$$\sum_{n=0}^{\infty} \frac{\Lambda_i^n}{\prod_{l=1}^{n} \phi_i(l)}$$

is finite for every $i = 1, \ldots, k$. For the network of single-server stations considered here, $\phi_i(l) = \mu_i$ for all $l \ge 1$ and all i. Letting $\rho_i = \Lambda_i/\mu_i$, the above stability condition is reduced to

$$\rho_i < 1, \qquad i = 1, \ldots, k. \tag{3}$$

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

2.1. Modified networks with interchangeable servers

We will consider modified networks where servers are *interchangeable*. This means that a server at station i has service rate μ_i, regardless of where that server was originally assigned. When c servers are working at station i, their combined (exponential) service rate is $c\mu_i$. In Section 3, this holds even when the number of customers at station i is less than c, and all servers are busy whenever there is work to do at any station. In Section 4, we assume that a customer may be served by only one server at a time, and some servers are idle only when the total number of customers in system is less than k. These assumptions are not the only possibilities. For example, cooperation may be one way, where, as in Section 5, server i may serve customers at station j when station i is idle, but not the converse.

Under any of these possibilities, we assume that the network operates under some *stationary assignment policy* (SAP). We only consider the simplest case of an SAP, called *preemption*: the assignment of servers to stations is a deterministic function of the state of the system. For Kelly networks, an SAP also determines the service rate assigned to each customer at every station. The assignment of servers changes only when there is a change of state, which occurs when there is either an arrival or a service completion. When there is a change of state, under preemption, a server may be assigned to a different station, even when this would interrupt a service that is underway. For any SAP, the service rate on each customer at every station is fixed, as long as there is no change in state.

Let n_i be the number of customers at station i and $\mathbf{n} = (n_1, \ldots, n_k)$. For a Jackson network, the state space is the collection of vectors $\mathbf{n}$. For a Kelly network, the state space is large. It includes route, stage, and position information about every customer in system. The modified network is a Markov chain with the same state space as the original network, but the transition rates depend on the SAP.

An alternative to preemption we call *completion*: a server may be reassigned only when idle or on service completion by that server.

When (3) is satisfied in the original network, the joint stationary distribution of the number of customers at each station is the product of the marginals (has *product form*), where the marginal distributions are geometric with mean at station i given by

$$L_i^o = \rho_i/(1-\rho_i), \qquad i = 1, \ldots, k. \tag{4}$$

For an original or modified network, we define (*total*) *work in system* (TAWS) at time t, $V^o(t)$ or $V^m(t)$, as the sum of the TRW to be performed on every customer at each station, and over every station, at time t, with (when stable) finite TAWS $\mathbb{E}(V^o)$ and $\mathbb{E}(V^m)$. We write

$$\mathbb{E}(V^o) = \sum_{i=1}^{k} w_i^o L_i^o. \tag{5}$$

Thus (4) and (5) determine $\mathbb{E}(V^o)$, but it is of much greater interest to turn this around for particular modified networks, analyzing $\mathbb{E}(V^m)$ directly and using it to determine or approximate L_i^m.

2.2. Results for the case of Jackson networks

When we have a Jackson network, formulating it as a Kelly network is awkward, particularly when, as is usually the case, the number of routes is infinite.

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

When the routing of customers is Markovian, with the transition probability p_{ij} being the probability that on completion of service at station i, a customer goes to station j next, independent of the past, we have a Jackson network. We let p_{i0} be the probability of departure from the network. For Jackson networks *only*, let γ_i be the (external) arrival rate to station i; that is, γ_i is the arrival rate of all routes where station i is visited first. The Λ_i, as defined before, now are the unique solution to

$$\Lambda_i = \gamma_i + \sum_{j=1}^{k} \Lambda_j p_{ji}, \qquad i = 1, \ldots, k. \tag{6}$$

When the network is transient, the actual arrival rate of customers at station i may be smaller than that given by (1) and (6).

Note that because of Markovian paths and exponential service, TRW is independent of prior history, including prior states visited and the time spent in service at station i by this customer. Hence, W_i^o and W_i^m have the same distribution. We will use W_i for both here, with $w_i = \mathbb{E}(W_i)$, but note that the analysis below is for the original network.

Given that a customer at station i (with W_i, composed partly of remaining service S_i there) goes next to station j, W_i has the same distribution as $S_i + W_j$, where S_i has distribution $\exp(\mu_i)$, S_i and W_j are independent, and we define $W_0 = 0$. With this information, we write that with probability p_{ij},

$$W_i \text{ has the same distribution as } S_i + W_j, \ j = 0, \ldots, k.$$

It follows that expected values w_i satisfy

$$w_i = 1/\mu_i + \sum_{j=1}^{k} p_{ij} w_j, \qquad i = 1, \ldots, k. \tag{7}$$

Note that W_i has the form of a finite sum over all stations of a geometric sum of exponential service times at each station, and has finite moments of all orders. Later, we will need a way to find second moments $\mathbb{E}(W_i^2)$. Squaring $S_i + W_j$, and using independence, we find that they satisfy

$$\mathbb{E}\big(W_i^2\big) = 2/\mu_i^2 + \sum_{j=1}^{k} p_{ij}\big[2w_j/\mu_i + \mathbb{E}\big(W_j^2\big)\big], \qquad i = 1, \ldots, k. \tag{8}$$

For (6) and arguments that equations of form (6), (7), and (8) have unique solutions, see [12], p. 319 and Theorem 8 on p. 167.

3. Full cooperation, a conservation law

We now consider modified networks with what we call *full cooperation* (FC): all servers are busy whenever there is work to do, anywhere in the network.

Let T_r be the *total work* (TW) to be performed on a route-r customer, that is, the sum of the service times at all stations to be visited in this route until departure; and let T_a be the total work to be performed on an arriving customer. Thus the distribution of T_a is a mixture of the T_r distributions with mixture probabilities $f_r = \lambda_r/\lambda_T$. In particular, we write

$$\mathbb{E}(T_a) = \sum_{r=1}^{R} f_r \mathbb{E}(T_r) \ \text{ and } \ \mathbb{E}\big(T_a^2\big) = \sum_{r=1}^{R} f_r \mathbb{E}\big(T_r^2\big). \tag{9}$$

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

It is easy to find an expression for $\mathbb{E}(T_a)$ and show that it is finite by equating two ways of representing the arrival rate of work to the system:

$$\lambda_T \mathbb{E}(T_a) = \sum_{i=1}^{k} \Lambda_i/\mu_i, \tag{10}$$

which shows that $\mathbb{E}(T_a) < \infty$.

Note that the amount of work brought by an arrival is defined as the amount of time it would take one server to do that work. Hence all servers work (do work) at rate 1.

Our use below of total work in system under FC also requires that $\mathbb{E}(T_a^2) < \infty$. However, this may not be true for Kelly networks. (Note that infinite $\mathbb{E}(T_a^2)$ cannot occur for Jackson networks.) In Subsection 3.1, we obtain as (19) an if-and-only-if condition for $\mathbb{E}(T_a^2) = \infty$, and investigate what happens when this is true.

The total-work-in-system process $\{V^m(t); t \geq 0\}$ has i.i.d. jumps (increases) distributed as T_a at the arrival times of external arrivals to the entire network (at rate λ_T). Between jumps, the process decreases at rate $-k$ when positive, because all servers are busy, and each server reduces total work at rate -1. Thus process $\{V^m(t)/k\}$, with i.i.d. jumps distributed as T_a/k, and decreases at rate -1, is equivalent to the work-in-system process for an M/G/1 queue with arrival rate λ_T and service times distributed as T_a/k, where this queue is stable if and only if

$$\lambda_T \mathbb{E}(T_a)/k < 1. \tag{11}$$

From (10), (11) is equivalent to

$$\rho_1 + \rho_2 + \cdots + \rho_k < k. \tag{12}$$

When (12) holds, this queue has time-average work

$$\frac{\lambda_T \mathbb{E}(T_a^2)/k^2}{2(1 - \lambda_T \mathbb{E}(T_a)/k)}.$$

Hence the modified network is stable if and only if (12) holds, and in this case, $\{V^m(t)\}$ has time average

$$\mathbb{E}\left(V_{FC}^m\right) = \frac{\lambda_T \mathbb{E}\left(T_a^2\right)}{2k(1 - \lambda_T \mathbb{E}(T_a)/k)}. \tag{13}$$

We have added the subscript FC to the left-hand side of (13) to distinguish it from a corresponding time average in Section 4, under an alternative to FC that we call AA.

For $t \geq 0$ and $i = 1, \ldots, k$, let $V_i^m(t)$ be the portion of $V^m(t)$ associated with customers at station i at time t, and let $N_i^m(t)$ be the number of customers at station i at time t. $V_i^m(t)$ is the sum of the *total remaining work* on each of the $N_i^m(t)$ customers at station i at time t.

When (12) holds, $\{N_i^m(t)\}$ has finite time average L_i^m. When we also have $\mathbb{E}(T_a^2) < \infty$, $\{V_i^m(t)\}$ has finite time average $\mathbb{E}(V_i^m)$, where

$$\sum_{i=1}^{k} \mathbb{E}\left(V_i^m\right) = \mathbb{E}\left(V_{FC}^m\right). \tag{14}$$

For every i, we now *define* w_i^m to be

$$w_i^m = \mathbb{E}\left(V_i^m\right)/L_i^m. \tag{15}$$

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

Thus w_i^m is no longer a property of a particular customer. It is an average. Actually, this is also true for w_i^o for Kelly networks. It is an average over the possible routes and stages a particular customer may be on.

From (13), (14), and (15), we have the following theorem.

Theorem 1. *Under FC, and when (12) holds and* $\mathbb{E}(T_a^2) < \infty$, *the* w_i^m *and* L_i^m *are finite and satisfy*

$$\frac{\lambda_T \mathbb{E}\left(T_a^2\right)}{2k(1-\lambda_T \mathbb{E}(T_a)/k)} = \sum_{i=1}^{k} w_i^m L_i^m. \tag{16}$$

Suppose a modified Jackson network is stationary at time t. At station i,

$$V_i^m(t) = \sum_{j=1}^{N_i^m(t)} W_{ij}^m(t),$$

where, summed in any order, $W_{ij}^m(t)$ is the TRW to be performed on the jth customer at station i at time t. Because of Markovian paths, the $W_{ij}^m(t)$ are i.i.d., independent to $N_i^m(t)$, with mean w_i^o. Taking expectations,

$$\mathbb{E}(V_i^m) = w_i^o L_i^m, \qquad i = 1, \ldots, k,$$

and we have the following result.

Corollary 2. *For modified Jackson networks, Theorem 1 holds, where for every i,* $w_i^m = w_i^o$, $i = 1, \ldots, k$.

For modified Kelly networks, we don't expect $w_i^m = w_i^o$. However, if the modified network doesn't change the 'mix' of customers at each station very much, w_i^o may be a good approximation of w_i^m. See Example 2, where this turns out to be true.

In other circumstances, w_i^m and w_i^o may be quite different. We return to this issue at the end of Subsection 3.2.

For a Jackson network, computation of the moments of T_a via (9) would be an arduous task. Instead, suppose that the arriving customer arrives at station i. Then T_a has the same distribution as W_i, and thus the distribution of T_a is a mixture of the W_i distributions with mixture probabilities $g_i = \gamma_i/\Lambda_i$. In parallel to (9), we write

$$\mathbb{E}(T_a) = \sum_{i=1}^{k} g_i \mathbb{E}(W_i) \text{ and } \mathbb{E}(T_a^2) = \sum_{i=1}^{k} g_i \mathbb{E}(W_i^2),$$

and these are determined by the solutions to (7) and (8).

We have already seen that server cooperation can transform an unstable original network to a stable modified network, and we expect that large improvements in performance measures can be achieved for stable original networks when some but not all stations are heavily loaded.

We now illustrate that large improvements in performance measures are obtained even when the original-network stations are equally loaded. To make comparisons easier, these examples have *symmetric stations*, meaning that the stations are stochastically equivalent. The term *symmetric queue* has a different meaning; see Section 3.1.

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

For the Jackson-network examples, we have not only symmetric stations, but also $w_i^m = w_i^o = w_i$ in (16). This greatly simplifies our analysis. For Kelly networks, w_i^m and w_i^o are not, in general, equal. In Example 2, we bound w_i^m and approximate it by w_i^o.

Example 1. Consider a Jackson network with two symmetric stations, $\gamma_1 = \gamma_2 = \gamma$, $\mu_i = \mu$, and $p_{12} = p_{21} = p_{10} = p_{20} = 1/2$. We have $\Lambda_1 = \Lambda_2 = 2\gamma$ and $\rho_i = \rho = 2\gamma/\mu < 1$ for all i. It is easy to see that each W_i is exponentially distributed with mean $2/\mu$, as is T_a. We assume cooperation is also symmetric. For example, when an arrival occurs at an empty station while the other is busy, a server returns immediately to that station (the assumption in [4]). This implies $L_1^m = L_2^m$. Evaluating the left-hand side of (16), we get

$$L_i^m = \rho/2(1-\rho) = L_i^o/2, \qquad i = 1, 2,$$

half the corresponding original quantities.

We now generalize the result under FC in stages. For a k-station network, suppose $\mu_i = \mu$ and $p_{i0} = \alpha \in (0, 1)$ for every station i. T_a and the W_i are random sums of i.i.d. exponential service times, where the probability that a customer leaves the system after n service completions is $\alpha(1-\alpha)^{n-1}$, $n \geq 1$.

To preserve symmetry in the modified network when $k > 2$, we allocate servers to busy stations symmetrically, where if b stations are busy, $(k-b)/b$ servers from empty stations are allocated to each busy server; this number may not be an integer. Splitting a server's service capacity between stations is done in [10] and [11].

The following is easily shown.

Lemma 1. *For a k-station Jackson network with $\mu_i = \mu$ and $p_{i0} = \alpha$ for every station i, T_a and the W_i have the same exponential distribution, with mean*

$$\mathbb{E}(T_a) = \mathbb{E}(W_i) = 1/\alpha\mu.$$

For this result, the transition probabilities between stations within the network are irrelevant. If the stations are symmetric, the L_i^m are equal; under FC, we evaluate (16) to obtain

$$L_i^m = L_i^o/k \tag{17}$$

for a k-station network. On reflection, this is obvious. The original network has what amounts to k M/M/1 queues, each with arrival rate $\Lambda_i = \gamma/\alpha$, service rate μ, and $\rho_i = \gamma/\alpha\mu$. The modified network operates like one M/M/1 queue with arrival rate $k\gamma$, service rate (of T_a/k) $k\alpha\mu$, and $\rho = k\gamma/k\alpha\mu = \gamma/\alpha\mu$. The same ρ!

We now obtain a similar result for a Kelly network.

Example 2. Consider a Kelly network with k FIFO symmetric stations and k routes, where service rate at every station is μ, Route 1 = (1, 2), Route 2 = (2, 3), . . ., Route k = (k, 1), and every route has arrival rate λ. Then T_a is 2-Erlang(μ) and $\rho_i = \rho = 2\lambda/\mu$.

Under FC, the right-hand side of (16) can be computed to be $3\rho/2\mu(1-\rho)$, but the w_i^m may not satisfy (2) and are unknown. From these simple routes, however, we have the bounds $1/\mu < w_i^m < 2/\mu$.

From these and the obvious fact that $L_1^m = L_2^m = \cdots = L_k^m$, we get

$$L_i^o/1.33k \leq L_i^m \leq L_i^o/0.67k, \qquad i = 1, \ldots, k.$$

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

TABLE 1: Results from 2.5×10^5 busy cycles for 2-station Kelly network.

	Original Kelly	Modified Kelly
ρ	Exact L_T^o	Estimated L_T^m $\pm$ 95% c.i.
0.1	0.222	0.111 ± 0.001
0.5	2.000	1.017 ± 0.003
0.9	18.00	9.23 ± 0.03
0.95	38.00	19.84 ± 0.07

TABLE 2: Results from 2.5×10^5 busy cycles for 4-station Kelly network.

	Original Kelly	Modified Kelly
ρ	Exact L_T^o	Estimated L_T^m $\pm$ 95% c.i.
0.1	0.444	0.112 ± 0.001
0.5	4.000	1.019 ± 0.003
0.9	36.00	9.09 ± 0.03
0.95	76.00	19.13 ± 0.07

For an approximation, we use (2) with $w_i^m \approx w_i^o = 3/2\mu$ to obtain

$$L_i^m \approx \rho/k(1-\rho) = L_i^o/k,$$

which is exact in (17) for a corresponding Jackson network.

To explore the accuracy of the approximation $w_i^m \approx w_i^o$ for Kelly networks, we present the results of two simulation experiments. Table 1 is for the model with $k = 2$ stations. Because of the narrow range of possible values of w_i^m in that example, we simulated, with the results in Table 2, a symmetric network with $k = 4$ stations, with routes (1, 2, 3, 4), (2, 3, 4, 1), (3, 4, 1, 2), and (4, 1, 2, 3). Results are presented for $L_T^\bullet \equiv L_1^\bullet + \cdots + L_k^\bullet$, where $^\bullet$ denotes either o or m.

All point estimates are remarkably close to L_T^o/k, which is exact for corresponding Jackson networks, and the confidence intervals are narrow. Except for $\rho = 0.1$, however, the confidence intervals do not cover this quantity, and all the estimates are on the high side. This is strong evidence that w_i^m is (slightly) smaller than w_i^o.

3.1. How infinite $\mathbb{E}(T_a^2)$ occurs; implications for Kelly networks

We will find it useful to call the collection of service times on a route a *batch*, and the number of these service times a *batch size*. The effect of batch size variance here is similar to what happens at a single-server FIFO queue when we have batch arrivals.

We now sort routes by their length (batch size). Let a_b be the combined arrival rate of all routes of batch size b, $b = 1, 2, \ldots$, where $\sum_{b=1}^{\infty} a_b = \lambda_T$, and $c_b = a_b/\lambda_T$. Let B be a random batch size with distribution $\mathbb{P}(B = b) = c_b$, $b = 1, 2, \ldots$.

Let S_{Tb} be the sum of the b service times in the batch. First consider the special case where all service times in the batch have the same rate μ (when the k stations all have the same service

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

rate). In this case, S_{Tb} is the sum of b i.i.d. exponential random variables. When we write out the b^2 terms in S_{Tb}^2, there are b terms that are squared exponentials and $b(b-1)$ cross-product terms. It follows that

$$\mathbb{E}\big(S_{Tb}^2\big) = \frac{2b}{\mu^2} + \frac{b(b-1)}{\mu^2} = \frac{b(b+1)}{\mu^2}.$$

In general, the μ_i will be different. As there is a fixed number of stations, there will be a maximum and minimum service rate μ_M and μ_m that are upper and lower bounds on all service rates in a batch, independent of b. In terms of these quantities, we have the upper and lower bounds

$$\frac{b(b+1)}{\mu_M^2} \le \mathbb{E}(S_{Tb}^2) \le \frac{b(b+1)}{\mu_m^2}. \tag{18}$$

Now, $\mathbb{E}\big(T_a^2\big) = \mathbb{E}\big[\mathbb{E}\big(T_a^2|B\big)\big]$, where $\mathbb{E}\big(T_a^2|B\big) = \mathbb{E}\big(S_{TB}^2\big)$, and hence from (18),

$$\frac{\mathbb{E}\big(B^2\big) + \mathbb{E}(B)}{\mu_M^2} \le \mathbb{E}(T_a^2) \le \frac{\mathbb{E}\big(B^2\big) + \mathbb{E}(B)}{\mu_m^2}.$$

The conclusion from above is that

$$\mathbb{E}(T_a^2) = \infty \quad \text{if and only if } B \text{ has infinite variance.} \tag{19}$$

When $\mathbb{E}(T_a^2) = \infty$, the modified network has infinite time-average work. As work in the modified network is a lower bound, work in the original Kelly network also is infinite. The only way this can happen is if $w_i^o = \infty$ for at least one station in the original network. We now present an example where this happens, using only the Kelly formulation and results.

Example 3. Consider a 2-station Kelly network with service rate $\mu_i = \mu$, $i = 1, 2$, and infinitely many customer routes with arrival rates $\{\lambda_r, r \ge 1\}$. In particular, for every odd r, a customer on route r visits stations 1 and 2 alternately r times each, and on every even route $r+1$ a customer visits stations 2 and 1 alternately r times each.

B, the route length (batch size) of a randomly arriving customer, has distribution

$$\mathbb{P}(B = 2r) = (\lambda_r + \lambda_{r+1})/\lambda_T, \qquad r = 1, 2, \ldots,$$

and consequently

$$\mathbb{E}(B) = \sum_{r=1}^{\infty} 2(2r-1)(\lambda_{2r-1} + \lambda_{2r})/\lambda_T,$$

where finite Λ_i implies B has finite first moment, but it may have an infinite variance.

Note that W_1^o is a random sum of i.i.d. exponential service times, $W_1^o = \sum_{i=1}^{Z} S_i$, where Z is the number of remaining stages on the route of a randomly selected customer at station 1. We have

$$w_1^o = \mathbb{E}(Z)/\mu.$$

To access $\mathbb{E}(Z)$, we first note that

$$\Lambda_1 = \sum_{r=1}^{\infty} (2r-1)(\lambda_{2r-1} + \lambda_{2r}) = \lambda_T \mathbb{E}(B)/2.$$

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

Then, from Corollary 1, we have

$$\mathbb{P}(Z=n)=\sum_{r=\lceil n/2\rceil}^{\infty}\frac{\lambda_r}{\Lambda_1}=\frac{1}{\Lambda_1}\frac{\lambda_T\mathbb{P}(B\geq n)}{2}=\frac{\mathbb{P}(B\geq n)}{\mathbb{E}(B)},\qquad n=1,2,\ldots. \tag{20}$$

It is easily shown from (20) and a standard result that

$$\mathbb{E}(Z)=\mathbb{E}[B(B+1)]/2\mathbb{E}(B).$$

Thus $\mathbb{E}(Z)$ and hence w_1^o are infinite if and only if B has an infinite variance.

In Kelly networks, we found that time-average work may be infinite when L_T^o is finite. This possibility occurs throughout the earlier related literature. For example, an M/G/1 queue with server utilization $\rho<1$ that operates as a *symmetric queue* (see [6, p. 72]) has $L=\rho/(1-\rho)$. The remaining service of a customer in service has *equilibrium distribution* with infinite mean when the corresponding service distribution G has infinite variance. Note that preemptive last-in-first-out and processor sharing (called *server-sharing* in [6]) are examples of a symmetric queue, but FIFO is not.

From the exclusion of FIFO in the preceding paragraph, one may wonder how FIFO is allowed in Kelly's formulation in [6, Chapter 2], where service times are exponential. 'For simplicity' (see [6, p. 57]), Kelly does not allow customers to visit the same station twice in a row. So on completion of service at one station, a customer either departs or (under FIFO) joins the end of the queue at another station. Suppose we allow this, and a twice-in-a-row customer at station i completes service there for the first time. What happens? If this customer is served again immediately, Kelly's results do not hold. If instead this customer joins the end of the queue at station i, they do.

For example, consider a network with one station and one route of length 2. If on the first service completion a customer is served again immediately, we have an M/E_2/1 queue. Kelly's results do not hold.

3.2. Optimal SAP and a counterexample

When we don't have symmetry, (16) is still valid, and a severe restriction on the L_i^m. Time-average work in system is substantially reduced. Now we have an interesting problem: how should servers be allocated to stations so as to reduce or even minimize L_T^m? Because the w_i^m are unknown for Kelly networks and depend on the SAP, we now consider only Jackson networks; we briefly return to Kelly networks at the end of this section.

Intuitively, reducing the L_i^m with small w_i will increase the L_j^m with large w_j, but not by as much. This suggests the following policy.

Definition 1. The SETRW rule assigns, at all times, all the servers to the non-empty station where customers have the shortest expected TRW.

In [11], it was shown that for a tandem queue, allocating all the service capacity to the last station that has work to do minimizes the waiting time in system of every customer. (At every station, FIFO was assumed.) This policy is the SETRW rule. This result holds on sample paths, without making any stochastic assumptions, and implies that the number of customers in system is minimized at every time t, and of course L_T^m is minimized. The service capacity of a single server was allocated over the network, but the same argument holds for multiple servers. Hence SETRW is optimal in this strong sense for tandem queues.

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

We now briefly investigate when SETRW is optimal in the sense that it minimizes L_T^m. In Example 4, we show that for a k-station network and an arbitrary station i, allocating all the service capacity to station i whenever it is busy minimizes L_i^m. Showing this is trivial when station i only has external arrivals. More care is required when station i also has internal arrivals. We then use this result for an easy proof that SETRW is optimal for 2-station networks. In Example 5, we present a counterexample where SETRW is not optimal for a 3-station network.

Example 4. For a k-station Jackson network that is stable under FC, let P be an FC policy that allocates all the service capacity to an arbitrary station i whenever it is busy. The allocation of service capacity to other stations when station i is empty is unspecified.

We assume FIFO order of service at station i. This is convenient in the analysis but not a restriction, because under Markovian paths, the stochastic evolution of the number-of-customers-in-system vector $\mathbf{n}$ is not affected by order of service.

We need some additional notation. Let d_i^m be the average delay in queue at station i, and let d_i^{em} and d_i^{im} be the same quantity for external and internal arrivals at station i. Let Q_i^m be the average number of customers in queue and L_{si}^m the average number of customers in service at station i.

Under policy P, d_i^m is the weighted average

$$d_i^m = (\gamma_i/\Lambda_i)d_i^{em} + (1 - \gamma_i/\Lambda_i)d_i^{im}, \tag{21}$$

where, because internal arrivals occur at station i only when it is empty,

$$d_i^{im} = 0. \tag{22}$$

From PASTA and under P, the service rate on a customer in service is $k\mu_i$, so

$$d_i^{em} = L_i^m/k\mu_i. \tag{23}$$

From Little's law (twice) and (21), (22), and (23),

$$\begin{aligned} Q_i^m &= \Lambda_i d_i^m = (\gamma_i/k\mu_i)L_i^m, \quad \text{and} \\ L_i^m &= Q_i^m + L_{si}^m = (\gamma_i/k\mu_i)L_i^m + \Lambda_i/k\mu_i. \end{aligned} \tag{24}$$

From (24), we now solve for L_i^m. Under P,

$$L_i^m = \Lambda_i/(k\mu_i - \gamma_i) \equiv L_i^{mP}, \tag{25}$$

the average number of customers at station i under policy P.

Now repeat the steps for any FC policy. Equation (21) is the same, but $d_i^{im} \geq 0$ in (22). Dropping the second term on the right in (21) gives

$$d_i^m \geq (\gamma_i/\Lambda_i)d_i^{em}. \tag{26}$$

From PASTA and that the service rate is $\leq k\mu_i$, we have

$$d_i^{em} \geq L_i^m/k\mu_i. \tag{27}$$

From Little's law, (26) and (27), and $L_{si}^m \geq \Lambda_i/k\mu_i$,

$$\begin{aligned} Q_i^m &= \Lambda_i d_i^m \geq (\gamma_i/k\mu_i)L_i^m, \quad \text{and} \\ L_i^m &= Q_i^m + L_{si}^m \geq (\gamma_i/k\mu_i)L_i^m + \Lambda_i/k\mu_i. \end{aligned} \tag{28}$$

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

From (25) and (28), we have that L_i^m under any FC policy has as a lower bound the quantity found in (25),

$$L_i^m \geq L_i^{mP}.$$

Now consider a 2-station Jackson network with stations 1 and 2, where $w_1 < w_2$. Policy P applied to station 1 is an SETRW policy. Under any FC policy, we have from (16) and $w_i^m = w_i$ for Jackson networks that $w_1 L_1^m + w_2 L_2^m$ is a constant; decreasing L_1^m by Δ increases L_2^m, but not by as much, so $L_T = L_1^m + L_2^m$ also decreases. Minimizing L_1 minimizes L_T. Hence, SETRW is optimal for 2-station networks.

Example 5. Consider a 3-station network with $1/\mu_1 = 1$, $1/\mu_2 = 3$, $1/\mu_3 = 80$, $p_{10} = 0.1$, $p_{12} = 0.9$, $p_{20} = 0.9$, $p_{23} = 0.1$, and $p_{30} = 1$. External arrivals enter the system at station 1 at rate γ_1. The value of γ_1 is arbitrary except that it must be small enough for the 3-state network to be stable under FC. From (7), $w_1 = 10.9$, $w_2 = 11$, and $w_3 = 80$, so that SETRW assigns service priority order to be 123. In what follows, we are comparing policies 123 and 213.

Under any FC policy, replacing the w_i by their numerical values, we have

$$10.9L_1^m + 11L_2^m + 80L_3^m = K_1, \tag{29}$$

a constant. Observe that under either policy, customers at station 3 are served only when there are no customers at stations 1 and 2.

Hence the time-average performance of the 2-station network consisting of stations 1 and 2 may be analyzed separately from station 3 by this simple change in the network: change transition probabilities out of station 2 to $p'_{20} = 1$ and $p'_{23} = 0$. We 'prime' the other quantities that change, namely the expected total remaining work, w'_1 and w'_2, which are easily found to be $w'_1 = 3.7$ and $w'_2 = 3$.

Under any FC policy for this new network,

$$3.7L_1^m + 3L_2^m = K_2, \tag{30}$$

a constant.

We know that policy 213 minimizes L_2^m. We now show that it also minimizes L_T^m. Let L_i^{ma} and L_i^{mb}, $i = 1, 2, 3$, be the average number of customers at station i under policies $a = 123$ and $b = 213$, respectively, where $L_2^{ma} - L_2^{mb} \equiv \Delta > 0$. Then, from (30), we get $L_1^{mb} = L_1^{ma} + (3/3.7)\Delta$. We plug these results into (29) to find $L_3^{mb} = L_3^{ma} + 0.03\Delta$. Combining these results, we have

$$L_T^{mb} - L_T^{ma} = -0.16\Delta < 0,$$

which shows that 213 is better than the SETRW policy 123.

The model in Example 5 helps verify that Equation (25) is correct. Under 123, we are applying P to station 1, which has no internal arrivals. We have $\Lambda_1 = \gamma_1$, and (25) becomes $\gamma_1/(3\mu_1 - \gamma_i)$. This is correct because station 1 operates as an M/M/1 queue with these arrival and service rates. Under 213, we are applying P to station 2, where $\gamma_2 = 0$, and (25) becomes $\Lambda_1/3\mu_2$. This is correct because at all times there is at most one customer at station 2, and L_2^m is the fraction of time station 2 is busy.

This example also shows that without special structure such as tandem queues, using the w_i alone to allocate service capacity may produce poor results. It fails to distinguish between two situations: w_i may be large simply because service at station i is very slow, or maybe station i

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

conducts inspections that are performed quickly, and when necessary, a lengthy repair will be done elsewhere. It may make sense to allocate more servers to inspections so that those that pass leave the system as quickly as possible.

We now return to Kelly networks. If we know the route and stage of every customer, we can allocate service capacity to customers with the shortest expected TRW. (We expect this would greatly increase some of the w_i^m.) Without special structure, it is unlikely that such a rule will be optimal. However, because we are using more information than is available in a Jackson network, the potential for improving performance may be greater in Kelly networks.

4. Another model of server cooperation

Under FC, all k servers are fully busy when there are fewer than k customers in system. Sometimes this is unrealistic. Here we assume that a customer may be served by only one server at a time, and that a server originally assigned to an idle station will serve a customer somewhere else, if needed. Under our preemption assumption in Section 2, this server will immediately return to the idle station when a customer arrives there.

Whatever the details, we make Assumption A (AA): the network operates under some SAP such that at all times, the number of busy servers is the number of customers in system when that number is less than k, and k when that number is at least k.

Note that in terms of whether servers are busy or not, AA operates the same as a standard M/G/k queue. This observation is useful for approximating time-average work in system, but as under FC, we still have a network. The flow of customers is different.

Under what conditions are these modified networks stable? Compared with FC, the total-work-in-system process for these models has the same jumps at arrival times, but from time to time will decrease at a rate less rapid than $-k$ when positive. Hence total work in system for these models will be larger (no smaller) at every time t than the corresponding quantity under FC. This holds on sample paths, and hence for time averages. Actually, this is true under any model of server cooperation with interchangeable servers. When the FC model is stable, time-average work under FC is a lower bound on time-average work under AA.

It is immediate that (12) is necessary for modified networks under AA to be stable. We now show that it is sufficient, but under a restriction that, as we will discuss, is of little practical importance.

Theorem 2. *Modified networks under AA are stable if and only if (12) holds for Jackson networks and for Kelly networks when, in addition, the number of routes is finite.*

Proof. We will show that when (12) holds under the stated conditions, the modified network is a Markov chain that spends a strictly positive fraction of time in a finite set of states. This implies the chain is stable (positive recurrent), and hence (12) is sufficient. We first prove the result for Jackson networks and then extend it to Kelly networks.

Let $\mathcal{S}$ be the set of states $\{\mathbf{n} : n_1 + \cdots + n_k \leq k-1\}$, and let $\mathcal{S}^c$ be the set of states not in $\mathcal{S}$. For Jackson networks under preemption, the $\mathbf{n}$ are states and $\mathcal{S}$ is a finite set.

Each chain leaves $\mathcal{S}$ whenever an external arrival finds that chain in a state such that $n_1 + \cdots + n_k = k-1$. Suppose such an event occurs at some arbitrary time t, and from that point forward, let X be the first passage time of the chain from $\mathcal{S}^c$ to $\mathcal{S}$. The duration of X depends on the total work in system to be performed when X begins.

We represent this total work when X begins by $T_a + T_f$, where T_a is the total work to be performed on the arriving customer and T_f is the sum of the TRW to be performed on each of the $k-1$ customers in system found by this arrival (called *found customers*). For a Jackson

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

network, a found customer at station i has TRW distributed as W_i, and T_f is the sum of $k-1$ such quantities. Clearly, $\mathbb{E}(T_a + T_f) < \infty$. Note that the distribution of T_f depends on how many found customers are at each station.

Let $\{V^m(t)\}$ be the total-work-in-system process for a modified network. Because all servers are busy during X, $\{V^m(t)/k\}$ is equivalent during X to the work-in-system process for the M/G/1 queue introduced in Section 3. Let B_a be an *exceptional-first-service busy period* (see [12, p. 392]) for this queue that begins at t with exceptional first service $S_a = (T_a + T_f)/k$. As the equivalent (11) holds, B_a is a proper random variable with expectation

$$\mathbb{E}(B_a) = \frac{\mathbb{E}(S_a)}{1 - \lambda_T \mathbb{E}(T_a)/k} < \infty.$$

It is also true and easily shown that if we replace S_a by S'_a, to generate B'_a,

$$S_a \le S'_a \Longrightarrow B_a \le B'_a$$

Because B_a ends when work in system hits zero, and the number of customers in system will fall below k earlier, we have

$$X \le B_a.$$

Now X is a generic member of a sequence $\{X_j\}$ of first passage times from $\mathcal{S}^c$ to $\mathcal{S}$, with corresponding quantities $S_{aj} = (T_{aj} + T_{fj})/k$ and B_{aj}. The X_j are not i.i.d., because the number of found customers at each station may change. We want not only that $\mathbb{E}(X) < \infty$, which now is obvious, but also that

$$X_j \le B'_{aj}, \tag{31}$$

where $\{B'_{aj}\}$ is an i.i.d. sequence to be identified, with $\mathbb{E}(B'_{aj}) < \infty$.

This is easily accomplished. Replace exceptional first service S_{aj} with $S'_{aj} = (T_{aj} + Y_j)/k$, where for each j, Y_j is the sum over i, $i = 1, \ldots, k$, of the sum of $k-1$ i.i.d. replicates of W_i for each station i. We have

$$T_{fj} \le Y_j,$$

where the Y_j are i.i.d. and $\mathbb{E}(Y_j) < \infty$. The S'_{aj} generate i.i.d. sequence $\{B'_{aj}\}$, and we have (31). Whenever the chain visits the set $\mathcal{S}$, it remains in $\mathcal{S}$ at least until the next arrival. These results imply that the chain spends a strictly positive fraction of time in the finite set $\mathcal{S}$, which completes the proof for Jackson networks.

Now consider a Kelly network with a finite number of routes. For every customer found in $\mathcal{S}$, we must enlarge the state space to include the route, stage, and position of every customer at each station. For the number of states in $\mathcal{S}$ to be finite, the number of routes must be finite, which is required for our method of proof. With notation X_j, T_{aj}, and T_{fj}, as defined earlier in the proof, we proceed directly to obtain the upper bound Y_j on T_{fj}.

The TRW on a found customer at some station is the remaining service time at that station plus the service times at the stations on the remainder of the route it is on. That sum is bounded above by the total of the service times of all stations visited on the same entire route.

Recall that $e(r)$ is the length (number of service times) of route r, and let l_R be the maximum over r of $e(r)$. Then $(k-1)l_R$ is an upper bound on the total number of service times that remain to be performed on all $k-1$ found customers. We don't have separate bounds on how many of these service times are to be performed at each station i, but as we are only seeking an upper bound, we use the same bound for it. Now define Y as the sum of $k(k-1)l_R$ independent

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

random variables, where $(k-1)l_R$ of them have the service distribution at station i, exp (μ_i), $i = 1, \ldots, k$.

For X_j, let Y_j be the jth replicate of Y. We have $T_{fj} \leq Y_j$ where the Y_j are i.i.d. and $\mathbb{E}(Y_j) < \infty$. The remainder of the argument is the same as for Jackson networks, completing the proof. □

The restriction to a finite number of routes is not really a restriction. Jackson networks usually have an infinite number of routes. Suppose we have R collections of customers, where R is finite. Each collection, when served at a k-station network, is a Jackson network, where all collections have the same service rate at each station, but each station, but each collection has a different transition probability matrix. Now serve all R collections at the same k-station network.

The final network is a Kelly network with a countable number of finite deterministic routes, but with special structure. It is a Markov chain with states that specify station, position, and transition probability matrix of every customer in system. Because R is finite, the number of states in set $\mathcal{S}$ is finite, and the same result holds.

To distinguish it from time-average work under FC, given in (13), we denote time-average work under AA by $\mathbb{E}(V^m_{AA})$. While the first quantity is known, the second is not. Furthermore, the second depends on the SAP, because the rate at which work in system declines under AA at any time t depends on the number of customers in system at that time, which in turn depends on the SAP.

To evaluate performance measures under AA for simple models with symmetric stations, as was done under FC in Examples 1 and 2, we would need to approximate $\mathbb{E}(V^m_{AA})$.

We obtained (13) by showing that, up to a scale factor, the work-in-system process for a modified Kelly or Jackson network under FC is the same as that for an M/G/1 queue. Under AA, the rate at which the work-in-system process decreases is the same function of the number of customers in system as that for an M/G/k queue. Of course, the way the number of customers in system decreases is much more complicated in the network case. Another complication is that we don't have exact expressions for M/G/k performance measures. We now propose an approximation for time-average work under AA that combines time-average work under FC with what amounts to an approximation of an M/G/k performance measure:

$$\mathbb{E}(V^m_{AA}) \approx \mathbb{E}(V^m_{FC}) \frac{L_{M/M/k}}{L_{M/M/1}}, \tag{32}$$

where service rate for M/M/1 is k times service rate for M/M/k. Examples may exist where (32) is very poor, but it works well in Example 1, where we have symmetric stations.

For Jackson networks with symmetric stations, W_i under AA and W_i under FC have the same distribution, so that $\mathbb{E}(V^m_{AA})/\mathbb{E}(V^m_{FC}) = L^m_{AA}/L^m_{FC}$. Furthermore, L^m_{AA} is the same as for a corresponding standard M/M/k queue, while L^m_{FC} is that of a corresponding M/M/1 queue. Thus (32) reduces to (33):

$$L^m_{AA} \approx L^m_{FC} \frac{L_{M/M/k}}{L_{M/M/1}}. \tag{33}$$

On the other hand, for Kelly networks with symmetric stations, we use the approximation that worked well in Example 2, namely $w^m_i \approx w^o_i$, to again reduce (32) to (33). It turns out that the above approximation is pretty good for all values of ρ. A numerical example is below.

Example 6. Under AA, consider a 3-station Kelly network with symmetric stations, and with routes (1, 2, 3), (2, 3, 1), and (3, 1, 2). We first estimate L^m_{AA} by simulation, then compare it to the approximation (33). The results are in Table 3.

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

TABLE 3: Results from 2.5×10^5 busy cycles for 3-station Kelly network.

ρ	L^m_{AA}	Approximation in (33)
0.1	0.289	0.302
0.5	1.652	1.789
0.9	9.619	9.985
0.95	19.41	20.33

5. One-way cooperation

For an original k-station network, we call each station i *stable* if $\rho_i < 1$, and *unstable* otherwise. We now assume one-way cooperation (OWC): a server assigned to a stable station will, when that station is empty, assist a server at some unstable station. Servers are interchangeable. Multiple servers may serve the same customer, as under FC and in [4]. We assume preemption.

We now investigate stability conditions for an entire OWC network (*not* individual stations, as defined above). For a stable Kelly network, ρ_i is the fraction of time station i is busy. Whether stable or not, ρ_i is the *offered load* at station i, which is the expected number of (interchangeable) servers required to serve all the workload there. Summing over i, (12) means that the *total* offered load to the network is less than the total number of servers; this is necessary for stability.

A 2-station example of server cooperation was analyzed in [4], where $\rho_1 > 1$ and $\rho_2 < 1$. Thus the original network is unstable (transient in this case). The queue length $\to \infty$ at station 1, but remains finite at station 2. The authors of [4] modify the original network as follows: when station 2 is empty, server 2 assists server 1 at station 1, increasing the (combined) exponential service rate there to some value $\mu_1^* > \mu_1$. They show that the modified network is stable when

$$\mu_1^*(1-\rho_2) + \mu_1\rho_2 > \Lambda_1, \tag{34}$$

that is, when the time-average available service rate at station 1 exceeds the composite arrival rate there.

In this example, we have OWC with preemption. The authors of [4] show that server cooperation can improve system performance enormously, but the analysis is difficult and confined to the 2-station case. Under our interchangeable assumption, we would set $\mu_1^* = 2\mu_1$ so that servers are interchangeable. The stability condition (34) of [4] is now (35), and it is trivial to show that this is the same as (12) when $k = 2$:

$$2\mu_1(1-\rho_2) + \mu_1\rho_2 > \Lambda_1. \tag{35}$$

We present an elementary argument that (35) is sufficient for stability. If this model is unstable (either transient or null recurrent), the fraction of time station 1 is busy is 1, and the fraction of time station 2 is idle is *at least* $1-\rho_2$ (it could be larger if some service times intended for station 2 never arrive there). Hence the departure rate of customers from station 1 is at least the left-hand side of (35). But this is impossible because it cannot exceed the arrival rate of customers at station 1, which is at most Λ_1, the arrival rate there when all customers are served. Applying the same argument to (34), we have that it is sufficient for the stability of the model in [4], but without work conservation, our analysis does not show that it is necessary.

Now suppose that we have k stations, where at least one is stable and at least one is unstable, and we have OWC with preemption. Let S be the set of stable stations and U the set of unstable

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

ones. As defined in Subsection 2.1, we consider only stationary allocation policies (SAP). Let $G(\mathbf{n})$ be an SAP that determines which servers from stable stations are assigned to each unstable station when the state is $\mathbf{n}$.

Let g_{ij} be the fraction of time that server i spends at station j under G, where $g_{ii} = \rho_i$ for $i \in S$. It is immediate that $\sum_{j\in U} g_{ij} = 1 - \rho_i$ for every $i \in S$. By the argument used for the 2-station network, the sufficient sufficient condition of stability for every $j \in U$ is

$$\mu_j\left(1 + \sum_{i\in S} g_{ij}\right) > \Lambda_j,$$

or, equivalently,

$$1 + \sum_{i\in S} g_{ij} > \rho_j.$$

Now, sum the associated inequalities over all unstable stations:

$$\sum_{j\in U}\left(1 + \sum_{i\in S} g_{ij}\right) > \sum_{j\in U} \rho_j.$$

Interchanging the order of summation and rearranging, we have

$$\sum_{j\in U}(1 - \rho_j) + \sum_{i\in S}(1 - \rho_i) > 0,$$

which is (12). We conclude the following.

Theorem 3. *A k-station OWC network with preemption, where some stations are unstable and the rest are stable, is stable if and only if (12) holds.*

6. Concluding remarks

The use of work in system and the concept of work conservation have a long history in the analysis of queues. Because our entire paper depends on an analysis of time-average work, it was essential that we determine an if-and-only-if condition for it to be finite. The fact that this is a new result for original Kelly networks should come as no surprise, because the analysis of these networks makes no formal use of work. The only result about work is Corollary 1, which determines the stationary distribution of the TRW of a customer at each station. The fact that the mean of this distribution may be infinite is of little importance in original networks.

Now consider FC. Except in symmetric examples, and of course when some stations in the original network are unstable, it is difficult to determine the extent to which a modified network improves performance. Comparisons are easier for Jackson networks because $w_i^o = w_i^m$, and when we also have symmetric stations, reducing time-average work reduces the average number of customers in system by the same factor. While comparisons are more difficult for Kelly networks, we expect improvements to be comparable or even better. As the number of stations increases, the reduction by a factor of k in the symmetric case suggests large reductions in other cases. In fact, when a few stations are initially heavily loaded, allocating more idle servers to them may produce even better results.

FC gives greater potential improvement than AA. Under AA, in fact, it is easy to see that in light traffic, there will be little improvement, as the time spent in queue will be small compared with the time spent in service, while the time spent in service will not change. In heavy traffic,

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

on the other hand, we expect the potential performance under AA to approach that under FC. What is clear is the potential for large improvement, even for moderately loaded systems.

In terms of the applicability of the stochastic assumptions we have made, Poisson arrivals will often be judged as reasonable, whereas Jackson's Markovian paths have neither the modeling flexibility nor the applicability that Kelly's formulation has. Exponential service? Our approach applies to general service distributions. Under FC, the work-in-system process will still be that of an M/G/1 queue, and the stability condition will be the same. Of course, investigating the performance improvement of our approach will be more difficult.

In [6, Section 3.3], beginning on p. 72, Kelly lets each station in the network operate as a symmetric queue. As we discussed in Subsection 3.1, when individual stations in isolation with Poisson arrivals operate this way, the stationary distribution of the number of customers in system is insensitive to the service distribution (depends only on the mean). While Kelly does not give a formal proof, his results carry over to general service distributions at those stations that operate this way. However, as we remarked earlier, FIFO does not satisfy the assumptions of a symmetric queue, and in fact, these assumptions are quite restrictive.

To illustrate that Kelly's results don't hold for general service distributions under FIFO, consider a FIFO tandem queue of single-server stations with Poisson arrivals and constant service times at each station. If the stations are arranged from longest to shortest service times, queueing occurs only at the first station. The number of customers at each of the other stations is either 0 or 1 at all times. For any other arrangement, this is clearly not so. Kelly's results are very different, and don't depend on the arrangement of the stations.

Not only that, operating as a symmetric queue can lead to very poor performance when service times are regular. For example, suppose we have a single-server queue in isolation with Poisson arrivals and constant service, and we operate under processor sharing, where for service rate μ, the remaining service of every customer in system decreases at $1/n\mu$ when there are n customers in system. Because service times are constant, customers will depart FIFO, but when they do, some work will have been performed on the customers that remain in the system. From work conservation, this implies that every customer departs later under processor sharing than under (the usual) FIFO—strictly later, except for those who end busy periods.

Finally, with interchangeable servers, our approach applies when multiple servers are assigned to the stations or even when only one server is assigned to the entire network.

Acknowledgements

This research was done during the first author's visit at the Department of Industrial Engineering and Operations Research at the University of California, Berkeley, and was supported by the Ministry of Science and Technology of Taiwan under the grant MOST 107-2410-H-259-019.

References

[1] ANDRADÓTTIR, S., AYHAN, H. AND DOWN, D. G. (2001). Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Manag. Sci.* **47**, 1421–1439.

[2] ANDRADÓTTIR, S., AYHAN, H. AND DOWN, D. G. (2003). Dynamic server allocation for queueing networks with flexible servers. *Operat. Res.* **51**, 952–968.

[3] BRAMSON, M. (2008). *Stability of Queueing Networks*. Springer, Berlin.

[4] FOLEY, R. D. AND MACDONALD, D. R. (2005). Large deviations of a modified Jackson network: stability and rough asymptotics. *Ann. Appl. Prob.* **15**, 519–541.

[5] JACKSON, J. R. (1963). Jobshop-like queueing systems. *Manag. Sci.* **10**, 131–142.

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press

[6] KELLY, F. P. (2011). *Reversibility and Stochastic Networks*. John Wiley, New York.
[7] MANDELBAUM, A. AND REIMAN, M. I. (1998). On pooling in queueing networks. *Manag. Sci.* **44**, 971–981.
[8] MEYN, S. (2008). *Control Techniques for Complex Networks*. Cambridge University Press.
[9] MIYAZAWA, M. (2009). Tail decay rates in double QBN processes and related reflected random walks. *Math. Operat. Res.* **34**, 547–575.
[10] RESING, J. AND ÖRMECI, L. (2003). A tandem queueing model with coupled processors. *Operat. Res. Lett.* **31**, 383–389.
[11] WANG, C.-L. AND WOLFF, R. W. (2005). Work-conserving tandem queues. *Queueing Systems* **49**, 283–296.
[12] WOLFF, R. W. (1989). *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, New Jersey.

https://doi.org/10.1017/apr.2020.63 Published online by Cambridge University Press