Delay Optimal Server Assignment to Symmetric Parallel Queues with Random Connectivities

Hassan Halabian, Ioannis Lambadaris, Chung-Horng Lung
Department of Systems and Computer Engineering
Carleton University, 1125 Colonel By Drive, Ottawa, ON, K1S 5B6, Canada
Email: {hassanh, ioannis, chlung}@sce.carleton.ca

***Abstract*—In this paper, we investigate the problem of assignment of K identical servers to a set of N parallel queues in a time slotted queueing system. The connectivity of each queue to each server is randomly changing with time; each server can serve at most one queue and each queue can be served by at most one server per time slot. Such queueing systems were widely applied in modeling the scheduling (or resource allocation) problem in wireless networks. It has been previously proven that Maximum Weighted Matching (MWM) is a throughput optimal server assignment policy for such queueing systems [1], [2]. In this paper, we prove that for a symmetric system with i.i.d. Bernoulli packet arrivals and connectivities, MWM minimizes, in *stochastic ordering* sense, a broad range of cost functions of the queue lengths including total queue occupancy (or equivalently average queueing delay).**

I. Introduction

Optimal stochastic control of emerging wireless networks is one of the primary objectives in the design of such networks. In general, the main goal in the stochastic control of wireless networks is to distribute the shared resources in physical (e.g. power) and MAC layers (e.g. radio interfaces, relay stations and orthogonal channels) to multiple users such that a certain stochastic performance attribute is optimized. While various performance attributes including the stable throughput region, power consumption and utility functions of the admitted rates have been studied in many papers, average queueing delay has been considered far less in literature. This is due to the inherent difficulty of delay optimal scheduling problems in queueing systems with time varying channel conditions. In this paper, we consider a discrete time queueing system which is suitable in modeling of orthogonal resource assignment (e.g. radio interfaces/channel allocation) in multi-user wireless access networks. In our system, we model the available shared resources by a set of identical servers. The model also consists of a set of queues whose connectivities to each server is changing by time randomly. Therefore, the resource assignment problem is equivalent to finding a *matching* between the queues and the servers at each time slot such that some performance objectives are optimized. It has been already shown that Maximum Weighed Matching (MWM) is throughput optimal for such a system, i.e., it maximizes the stable throughput region of the system [1], [2]. MWM has also been extensively used in literature for treating the scheduling problem in crossbar packet switches [3]–[6]. In this paper, we prove that for a symmetric system with i.i.d. Bernoulli arrivals and connectivities (i.e. with the same arrival and connectivity parameters for all the queues), MWM is also optimal in minimizing, in *stochastic ordering* sense, a broad range of cost functions of queue lengths including total queue occupancy (or equivalently average queueing delay)[1]. In other words, we show that MWM policy minimizes stochastically a broad range of cost functions of queue length processes including the expected total queue occupancy across all possible server assignment policies.

The problem of optimal server allocation in queueing systems with random connectivities was mainly addressed in [1], [2], [7]–[13]. In [1], the authors introduced the notion of stability region of a general queueing network with time varying connectivities and they proposed back-pressure algorithm as a throughput optimal resource allocation policy for queueing networks. In [7], they considered a multi-queue single-server queueing system with random connectivities. They characterized the stability region by a set of linear inequalities and also proved that for a symmetric system with the same arrival and connectivity parameters for all the queues, LCQ (Longest Connected Queue) provides the optimal performance in terms of average queue occupancy.

In [11], Maximum Weight (MW) policy was proposed as a throughput optimal server allocation policy for multi-queue multi-server queueing systems with stationary channel processes. In [13], the authors characterized the network capacity region of multi-queue multi-server queueing systems with time varying connectivities. They also obtained an upper bound for the average queueing delay of AS/LCQ policy which is a throughput optimal server allocation policy for these systems. The results were fur-

[1]We order two discrete time random processes $A = \{A(t)\}_{t=1}^{\infty}$ and $B = \{B(t)\}_{t=1}^{\infty}$ stochastically as follows: We say A is stochastically less than B and we write $A \leq_{st} B$ if $\Pr(A(t) > r) \leq \Pr(B(t) > r)$ for all $t = 1, 2, \ldots$ and all $r \in \mathbb{R}$. The notion and relevant properties will be discussed in more detail in Section III-B.

ther extended in [14] for more general stationary channel distributions (and not just i.i.d. Bernoulli channels).

The authors in [8] considered a queueing model with a set of symmetrical parallel queues competing for K identical servers. The connectivity of each queue to all the servers is assumed to be the same at each time slot and during each time slot, each queue can attract at most one server. The authors proposed LCQ policy in which the servers are allocated to the K longest connected queues at each time slot. They proved the optimality of LCQ policy by using dynamic coupling and stochastic ordering method.

The work in [9], [10], [12], [15] focuses on the optimal server allocation problem in multi-queue multi-server queueing systems in terms of average queueing delay. In [9], [10], [15], the authors introduced MTLB (Maximum-Throughput Load-Balancing) policy and showed that this policy minimizes a class of cost functions including total average delay for the case of two symmetric queues. The work in [12] considers this problem for general number of symmetric queues and servers. In [12], a class of *Most Balancing* (MB) policies was characterized among all work conserving policies which are minimizing, in stochastic ordering sense, a class of cost functions including total average delay. Note that in the model used in [9], [10], [12], [13], [15], there is no restriction on the number of servers that are serving a queue at each time slot. In [2], it was shown that for a multi-queue multi-server system in which queues are restricted to attract at most one server at each time slot, Maximum Weighted Matching (MWM) policy is throughput optimal. The authors also considered the effect of infrequent channel state measurements on the stability region.

The rest of the paper is organized as follows. Section II describes the model and the notation required throughout the paper. In section III, we introduce Maximum Weighted Matching (MWM) policy as the optimal policy for the described model. We will also review the concepts of stochastic ordering and dynamic coupling method which are the main mathematical tools used in proving the optimality of MWM policy. In section IV, we present the main result of this paper, that is proving the optimality of MWM server assignment policy. Section V summarizes the conclusions of the paper.

II. Model Description

We consider a time slotted parallel queueing system with a set of parallel symmetrical queues $\mathcal{N} = \{1, 2, ..., N\}$ and infinite buffer space for each queue. Packets in this system are assumed to have constant length and require one time slot to complete service. The service to this set of queues is provided through a set of identical servers namely $\mathcal{K} = \{1, 2, ..., K\}$. The connectivity of each queue $n \in \mathcal{N}$ to each server $k \in \mathcal{K}$ at each time slot t is random and follows a Bernoulli distribution. We denote the connectivity of queue n to server k at time slot t by $C_{n,k}(t)$. Note that

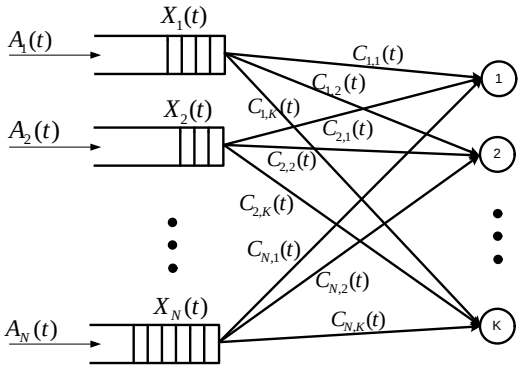


Fig. 1: Discrete time queueing system with N parallel queues and K servers

$C_{n,k}(t) \in \{0, 1\}$ and $E[C_{n,k}(t)] = p$ for all $n \in \mathcal{N}$ and $k \in \mathcal{K}$ and $t = 1, 2, ...$

At each time slot, each server can serve at most one packet from a connected non-empty queue. Note that in the system we do not have server sharing i.e., a server can serve at most one queue at each time slot. We also assume that a queue which is being serviced by a server at a given time slot, cannot get service from other servers during the same time slot.

Let $A_n(t)$ be the packet arrival process (number of packet arrivals) to queue n at time slot t. We assume that new arrivals at each time slot are added to the queues at the end of the time slot. Assume that the arrival processes $A_n(t)$ at each time slot t are independent Bernoulli random variables with the same parameter for all n and t. We denote the length of queue n at the end of time slot t (i.e., after adding the new arrivals) by $X_n(t)$. In other words, $X_n(t)$ represents the number of packets in the nth queue at the end of time slot t (or beginning of time slot $t + 1$).

A server assignment policy at each time slot determines an assignment of servers of set $\mathcal{K}$ to the queues of set $\mathcal{N}$. In other words, at each time slot the scheduler has to decide about a *bipartite matching* (matching in bipartite graphs) between sets $\mathcal{N}$ and $\mathcal{K}$. This should be accomplished based on the available information about the connectivities $C_{n,k}(t)$ and also the queue length process at the beginning of time slot t (which is $X(t-1) = (X_1(t-1), X_2(t-1), ..., X_N(t-1))$). For a given policy π, suppose that indicator variable $I^{(\pi)}_{n,k}(t)$ is defined to be "1" if server k is assigned to queue n at time slot t and "0" otherwise. We define $M^{(\pi)}(t) = \{I^{(\pi)}_{n,k}(t), \forall n \in \mathcal{N}, k \in \mathcal{K}\}$ as the employed *matching* by policy π at time slot t. Therefore, a server scheduling policy π is defined as $\pi = \{M^{(\pi)}(t)\}_{t=1}^{\infty}$.

According to the above discussion, we can see that the queue length random variable $X_n(t), \forall n \in \mathcal{N}$ evolves with time according to the following rule:

$$X_n(t) = \left(X_n(t-1) - \sum_{k=1}^{K} C_{n,k}(t) I^{(\pi)}_{n,k}(t) \right)^{+} + A_n(t)$$

where $(\cdot)^+$ returns the term inside the brackets if it is non-negative and zero otherwise. Note that a server can be assigned to an empty queue however it cannot serve it since there is no packet to be served. That is why we have used operator $(\cdot)^+$ in (1).

As we discussed earlier, the queueing model introduced in this section is useful in modeling the resource assignment problem in various systems with shared resources. In wireless communication systems, communication resources such as communication sub-channels, relay stations, etc. are shared among users and therefore can be studied using our model (e.g. [2], [16]). Bipartite Matching also has been extensively used in literature (e.g. [3]–[6]) to model the scheduling problem in crossbar packet switching systems. In this paper, random variables are represented by CAPITAL letters and lower case letters are used to represent sample values of the random variables.

III. Background

A. Maximum Weighted Matching

In [1], [2], [17]–[19], it was shown that Back-pressure algorithm maximizes the stable throughput region of a general data network. For the model introduced in section II, Back-pressure algorithm is equivalent to solving the following optimization problem at each time slot t [2].

$$\begin{aligned} \texttt{Maximize} \quad & \sum_{n=1}^{N} x_n(t-1) \sum_{k=1}^{K} I_{n,k}(t) c_{n,k}(t) \\ \texttt{s.t.} \quad & \sum_{k=1}^{K} I_{n,k}(t) \leq 1 \quad (n = 1, 2.., N) \\ & \sum_{n=1}^{N} I_{n,k}(t) \leq 1 \quad (k = 1, 2.., K) \end{aligned} \tag{1}$$

where $x_n(t-1)$ and $c_{n,k}(t)$ are the values of random variables $X_n(t-1)$ and $C_{n,k}(t)$ at time slots $t-1$ and t, respectively. Note that finding the solutions of problem (1) is equivalent to finding a maximum weighted matching in the bipartite graph $G_t = (\mathcal{N}, \mathcal{K}, \mathcal{E})$ (see Figure 2). In G_t, $\mathcal{N}$ and $\mathcal{K}$ are the two sets of vertices in each part of the graph and $\mathcal{E} = \{e_{n,k}, \forall n \in \mathcal{N}, \forall k \in \mathcal{K}\}$ is the set of edges between these two parts. Note that the associated weight to each edge $e_{n,k}$ is $x_n(t-1)c_{n,k}(t)$. A matching in graph G_t is basically a sub-graph of G_t in which no two edges share a common vertex. Note that any matching $M^{(\pi)}(t)$ at any time slot t is corresponding to a sub-graph of G_t namely $G_t^{(\pi)} = (\mathcal{N}, \mathcal{K}, \mathcal{E}^{(\pi)})$ in which $e_{n,k} \in \mathcal{E}^{(\pi)}$ if and only if $I_{n,k}^{(\pi)}(t) = 1$. Suppose that $M^{(\text{MWM})}(t) = \{I_{n,k}^{(\text{MWM})}(t), \forall n \in \mathcal{N}, k \in \mathcal{K}\}$ be the matching whose indicator variables are the solution of the optimization problem (1). Thus, we define Maximum Weighted Matching (MWM) server assignment policy as MWM $= \{M^{(\text{MWM})}(t)\}_{t=1}^{\infty}$. There are several algorithms to find the maximum weighted matching in bipartite graphs. The most well known algorithm is Hungarian algorithm whose complexity is of $O((\min\{N,K\})(\max\{N,K\})^2)$ [20].

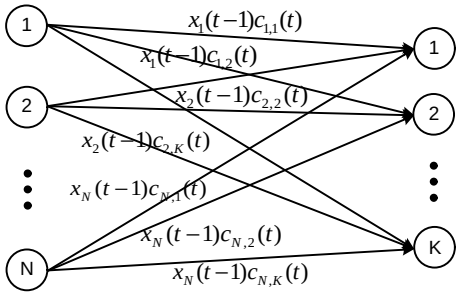


Fig. 2: Bipartite graph corresponding to problem (1)

As explained before, MWM is known to be throughput optimal for the queueing system described in section II [2]. Our contribution in this paper is to prove that MWM is also optimal in minimizing, in stochastic ordering sense, a class of cost functions of the queue length processes including the total system occupancy (or equivalently total average queueing delay) for the symmetric queueing system of Figure 1 (which can be used to model a homogeneous wireless access network). We will introduce a detailed description of those class of cost functions in the following section.

B. Stochastic Ordering and Dynamic Coupling

In this section, we briefly review the concepts of stochastic ordering (stochastic dominance) and dynamic coupling techniques. Consider two discrete time stochastic processes $A = \{A(t)\}_{t=1}^{\infty}$ and $B = \{B(t)\}_{t=1}^{\infty}$ in $\mathbb{R}$. We say A is stochastically less than B and we write $A \leq_{st} B$ if $\Pr(A(t) > r) \leq \Pr(B(t) > r)$ for all $t = 1, 2, ...$ and all $r \in \mathbb{R}$ [21], [22]. Some properties of stochastic ordering are the following. If $A \leq_{st} B$ then $f(A) \leq_{st} f(B)$ for all non-decreasing functions f. If $A \leq_{st} B$ then $E[A(t)] \leq E[B(t)]$. A is stochastically smaller than B ($A \leq_{st} B$), if there exists process $\tilde{A} = \{\tilde{A}(t)\}_{t=1}^{\infty}$ defined on the same probability space as B with the same probability distribution as A and satisfy $\tilde{A}(t) \leq B(t)$ almost surely for every $t = 1, 2, ...$ [8]. The last statement is known as coupling of A and $\tilde{A}$. In fact, when applying coupling technique, we are given the process A and we try to construct a coupled process $\tilde{A}$ with the same distribution as A and $\tilde{A}(t) \leq B(t)$ a.s. for all t. This gives us a tool for comparing processes A and B stochastically. This is specially useful when it is infeasible to derive the distributions of A and B (e.g. in our queueing model when comparing the total occupancy process for different server assignment policies).

IV. Optimality of MWM

In this section, we present the main result of this paper, that is proving the optimality of MWM with respect to minimization of a class of cost functions of queue lengths

including the average queueing delay. Suppose that $\mathbb{Z}_+$ be the set of non-negative integers and $\mathbb{Z}_+^N$ be the N dimensional Cartesian space of non-negative integers. We define relation " $\preceq$ " over $\mathbb{Z}_+^N$ as follows.

Definition 1: For two vectors x , $\tilde{x} \in \mathbb{Z}_+^N$, we write $\tilde{x} \preceq x$ if one of the following relations holds:

D1: $\tilde{x}_n \leq x_n$ for all $n = 1, 2, ..., N$

D2: $\tilde{x}$ is obtained by permutation of two distinct elements of x, i.e., $\tilde{x}$ and x are different in only two elements n and m such that $\tilde{x}_n = x_m$ and $\tilde{x}_m = x_n$.

D3: $\tilde{x}$ and x are different in only two elements n and m such that $x_n < \tilde{x}_n \leq \tilde{x}_m < x_m$ and the following constraints are satisfied: $\tilde{x}_n = x_n + 1$ and $\tilde{x}_m = x_m - 1$.

In **D3**, we say that $\tilde{x}$ is more balanced than x and can be obtained by decreasing a larger element of x (between m and n) by "1" and increasing a smaller element (between m and n) by "1". We call such an interchange a *balancing interchange* on vector x. Thus, the result of a balancing interchange on a vector x would be a vector $\tilde{x}$ such that $\tilde{x} \preceq x$. Suppose that vector $x \in \mathbb{Z}_+^N$ represents the queue length vector at a given time slot. Then, a balancing interchange is equivalent to taking a packet from a larger queue and adding it to a smaller queue.

We define the partial order " $\preceq_p$ " on $\mathbb{Z}_+^N$ as the transitive closure of relation " $\preceq$ " [23]. In other words, $\tilde{x} \preceq_p x$ if and only if $\tilde{x}$ is obtained from x by performing a sequence of reductions, permutations of two elements and/or balancing interchanges. When x and $\tilde{x}$ are two queue length vectors, we write $\tilde{x} \preceq_p x$ if and only if queue length vector $\tilde{x}$ is obtained from x by applying a series of packet removal, two queues permutations and balancing interchanges.

We define $\mathcal{F}$ as the class of real-valued functions on $\mathbb{Z}_+^N$ that are monotone and non-decreasing with respect to the partial order " $\preceq_p$ ", i.e.,

$$f \in \mathcal{F} \iff \tilde{x} \preceq_p x \Rightarrow f(\tilde{x}) \leq f(x). \tag{2}$$

We can easily check that function $f(x) = \sum_{n=1}^N x_n$ belongs to $\mathcal{F}$. This function captures the total queue occupancy of the system.

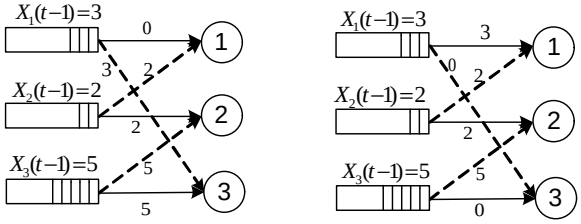
Let $X'(t) = (X'_1(t), X'_2(t), ..., X'_N(t))$ denote the queue length vector at time slot t exactly after serving the queues according to a server assignment policy π and before adding the new arrivals of time slot t, i.e.,

$$X'_n(t) = \left(X_n(t-1) - \sum_{k=1}^K C_{n,k}(t) I^{(\pi)}_{n,k}(t) \right)^+. \tag{3}$$

Given $x'(t)$ as a sample value of random variable $X'(t)$, we define a *balancing server reallocation* at time slot t as follows:

Definition 2: A balancing server reallocation on vector $x'(t)$ is a matching that results in vector $\tilde{x}'(t)$ such that one of the following conditions is satisfied.

(C1): $\tilde{x}'_n(t) \leq x'_n(t)$ for all $n = 1, 2, ..., N$ and there exists $m \in \{1, 2, ..., N\}$ such that $\tilde{x}'_m(t) < x'_m(t)$.



(a) Satisfying condition **C1** (b) Satisfying condition **C2**

Fig. 3: Examples of balancing server reallocations

(C2): $\tilde{x}'(t)$ and $x'(t)$ are different in only two elements n and m such that $x'_n(t) < \tilde{x}'_n(t) \leq \tilde{x}'_m(t) < x'_m(t)$ and the following constraints are satisfied: $\tilde{x}'_n(t) = x'_n(t) + 1$ and $\tilde{x}'_m(t) = x'_m(t) - 1$.

Figures 3a and 3b show two examples of balancing server reallocations in two sample graphs. In these figures, the original allocations are specified by solid lines while the balancing reallocations are specified by dashed lines.

Consider an arbitrary server assignment policy π with the allocation variables $\{I^{(\pi)}_{n,k}(t)\}_{t=1}^{\infty}$ for all $k \in \mathcal{K}$ and $n \in \mathcal{N}$. We introduce Matching Weight (MW) index associated to a server allocation policy π at time slot t by

$$\mathsf{MW}_\pi(t) = \sum_{n=1}^N x_n(t-1) \sum_{k=1}^K c_{n,k}(t) I^{(\pi)}_{n,k}(t) \tag{4}$$

Note that MW index is exactly the objective of the optimization problem (1). According to Definition 2 and definition of MW index, we can prove the following Lemma.

Lemma 1: For a given policy π employing matching $M^{(\pi)}(t)$ at time slot t, by applying a balancing server reallocation at time slot t (if there exists any) we will have a new policy $\tilde{\pi}$ differing from π only at time slot t such that $\mathsf{MW}_\pi(t) < \mathsf{MW}_{\tilde{\pi}}(t)$.

The proof is omitted here due to space limitations. The detailed proof of the lemma is given in [24]. Based on Lemma 1, we can state the following corollary.

Corollary 1: For a given policy π at time slot t, if $\mathsf{MW}_\pi(t)$ is maximized, i.e., policy π employs a maximum weighted matching at time slot t, then there exists no balancing server reallocation at that time slot.

Note that Lemma 1 just states that any balancing reallocation increases the matching weight index. However, it does not imply the existence of a balancing server reallocation when $\mathsf{MW}_\pi(t)$ is not maximized. In the following, we will prove the reverse of Lemma 1.

Lemma 2: For a given policy π at time slot t, if $\mathsf{MW}_\pi(t)$ is not maximized, i.e., $\mathsf{MW}_\pi(t) < \mathsf{MW}_{\mathrm{MWM}}(t)$, then there exists a balancing server reallocation at that time slot.

The proof is lengthy and is omitted here due to space limitations. For the detailed proof, please refer to [24].

By Π^{MWM}, we denote the set of all policies who employ maximum weighted matching at all time slots. We also

define Π_t as the set of all policies that employ maximum weighted matching exactly until time slot t (including t). We can easily observe that $\Pi_t \subseteq \Pi_{t-1}$ and $\Pi^{\text{MWM}} = \bigcap_{t=1}^{\infty} \Pi_t$. From Lemmas 1 and 2 we conclude that given a policy $\pi \in \Pi_{t-1}$ which is using an arbitrary matching at time slot t, we can reach to a policy $\pi^\star \in \Pi_t$ by applying a sequence of balancing server reallocations. Suppose that h_t^π represents the number of balancing server reallocations required to convert the employed matching in policy π at time slot t to a maximum weighted matching. In this case, we say that the distance of π from Π_t is h_t^π balancing server reallocations. Note that if the distance of π from Π_t is h_t^π, after applying the first balancing server reallocation, we get to a policy $\tilde{\pi}$ whose distance from Π_t is $h_t^\pi - 1$ balancing server reallocations. By repeating this procedure we finally get to a policy whose distance to Π_t is zero, i.e., it belongs to Π_t. By Π_t^h $(0 \leq h \leq h_t^\pi)$ we denote the set of all server assignment policies in Π_{t-1} whose distance from Π_t is at most h balancing sever reallocations. Note that $\Pi_t^0 = \Pi_t$.

Consider any two policies π and $\tilde{\pi}$ such that $f(\tilde{X}) \leq_{st} f(X)$, $f \in \mathcal{F}$ where $X = \{X(t)\}_{t=1}^{\infty}$ and $\tilde{X} = \{\tilde{X}(t)\}_{t=1}^{\infty}$ are the queue length processes when policies π and $\tilde{\pi}$ are applied respectively. For such a system, we say policy $\tilde{\pi}$ *dominates* π. Therefore, if $\tilde{\pi}$ dominates π we have $E[f(\tilde{X})] \leq E[f(X)]$. Given $f(x) = \sum_{n=1}^{N} x_n$, we conclude that the average queue occupancy (or equivalently average queueing delay) of policy $\tilde{\pi}$ is smaller than that of policy π. According to the above discussion, we can prove the following Lemma.

Lemma 3: For any policy $\pi \in \Pi_t^h$ and $0 < h \leq h_t^\pi$ we can construct a policy $\tilde{\pi} \in \Pi_t^{h-1}$ such that $\tilde{\pi}$ dominates π.

Here, we just give the outline of the proof. For the detailed proof please refer to [24]. The proof follows by applying dynamic coupling method over random variables $C(t) = (C_{n,k}(t)), \forall n \in \mathcal{N}, \forall k \in \mathcal{K}$ and $A(t) = (A_1(t), A_2(t), ..., A_N(t))$. In other words, we will show that given an arbitrary sample path $\omega = (x(0), c(1), a(1), x(1), c(2), a(2), x(2), c(3), a(3), x(3)...)$ we can construct policy $\tilde{\pi}$ and a new sample path $\tilde{\omega} = (\tilde{x}(0), \tilde{c}(1), \tilde{a}(1), \tilde{x}(1), \tilde{c}(2), \tilde{a}(2), \tilde{x}(2), \tilde{c}(3), \tilde{a}(3), \tilde{x}(3), ...)$ resulting in a new sequence of random variables $(\tilde{X}(0), \tilde{C}(1), \tilde{A}(1), \tilde{X}(1), \tilde{C}(2), \tilde{A}(2), \tilde{X}(2), \tilde{C}(3),...)$ with $X(0) = \tilde{X}(0)$ such that $\tilde{x}(t) \preceq_p x(t)$ *for all* t. In fact, we construct $\tilde{\omega}$ and $\tilde{\pi} \in \Pi_t^{h-1}$ in such a fashion that *for all* the sample paths and all time slots we have $\tilde{x}(t) \preceq_p x(t)$. The construction of $\tilde{\pi}$ is consisting of two main steps: construction for time slots before and including t and construction for time slots after t. The construction before and including t follows by using the matchings of policy π for time slots before t. For time slot t, we apply the balancing server reallocation. The construction after t follows by using mathematical induction. The detailed proof is lengthy and is omitted at this point. We refer the interested readers to [24] for more detail.

Based on Lemma 3, we can prove the main result of this paper in the following Theorem.

Theorem 1: Maximum Weighted Matching policy dominates any server assignment policy.

Proof: Let π_0 be any arbitrary policy. Then $\pi_0 \in \Pi_0 = \Pi_1^{H_1}$ where $H_1 = h_1^{\pi_0}$. By applying Lemma 3 repeatedly, we can construct a sequence of policies such that each policy dominates the previous one. Thus, we obtain policies that belong to $\Pi_0 = \Pi_1^{H_1}, \Pi_1^{H_1-1}, \Pi_1^{H_1-2}, ..., \Pi_1^0 = \Pi_1$. The last policy is called π_1. Note that $\pi_1 \in \Pi_2^{H_2}$ where $H_2 = h_2^{\pi_1}$. By recursively continuing such argument we obtain a sequence of policies $\pi_t \in \Pi_t$, $t = 1, 2, ...$ such that π_j dominates π_i for $j > i$. Note that this sequence of policies defines a limiting policy π^* that agrees with MWM at all time slots. Thus, π^* is an MWM policy who dominates all the previous policies, including the starting policy π_0. ■

V. Conclusions

In this paper, we considered the problem of assignment of K identical servers to a set of N parallel queues in a symmetrical time slotted queueing system with random connectivities from the queues to the servers. For such a queueing system, it has been previously shown that MWM is throughput optimal, i.e. has the maximum stability region. Our contribution in this work is the development of a method to prove the optimality of MWM in minimizing, in stochastic ordering sense, a class of cost functions of queue lengths (including total queue occupancy or equivalently average queueing delay). Our method to achieve this goal used stochastic ordering and dynamic coupling techniques.

References

[1] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Auto. Control*, vol. 37, no. 12, pp. 1936–1949, Dec. 1992.

[2] K. Kar, X. Luo, and S. Sarkar, "Throughput-optimal scheduling in multichannel access point networks under infrequent channel measurements," *IEEE Trans. Wireless Comm.*, vol. 7, no. 7, pp. 2619–2629, July 2008.

[3] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "On achieving 100% throughput in an input-queued switch," *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1260–1272, Aug. 1999.

[4] L. Tassiulas, "Linear complexity algorithms for maximum throughput in radio networks and input queued switches," in *Proc. of IEEE INFOCOM*, San Francisco, CA, USA, Apr. 1998.

[5] M. J. Neely and E. Modiano, "Logarithmic delay for nxn packet switches," in *Proc. of IEEE Workshop on High Performance Switching and Routing*, Phoenix, AZ, USA, Apr. 2004.

[6] E. Leonardi, M. Mellia, F. Neri, and M. A. Marsan, "Bounds on average delays and queue size averages and variances in input-queued cell-based switches," in *Proc. of IEEE INFOCOM*, Anchorage, AK, USA, Apr. 2001.

[7] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. Inform. Theory*, vol. 39, no. 2, pp. 466–478, Mar. 1993.

[8] A. Ganti, E. Modiano, and J. N. Tsitsiklis, "Optimal transmission scheduling in symmetric communication models with intermittent connectivity," *IEEE Trans. Inform. Theory*, vol. 53, no. 3, pp. 998–1008, Mar. 2007.

[9] S. Kittipiyakul and T. Javidi, "Delay-optimal server allocation in multi-queue multi-server systems with time-varying connectivities," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2319–2333, May 2009.

[10] ——, "Resource allocation in ofdma with time-varying channel and bursty arrivals," *IEEE Commun. Lett.*, vol. 11, no. 9, pp. 708–710, Sep. 2007.

[11] T. Javidi, "Rate stable resource allocation in ofdm systems: from waterfilling to queue-balancing," in *Proc. Allerton Conference on Communication, Control, and Computing*, Oct. 2004.

[12] H. Al-Zubaidy, I. Lambadaris, and I. Viniotis, "Optimal resource scheduling in wireless multi-service systems with random channel connectivity," in *Proc. of IEEE Global Communications Conference (GLOBECOM 2009)*, Honolulu, HI, USA, Nov. 2009.

[13] H. Halabian, I. Lambadaris, and C.-H. Lung, "Network capacity region of multi-queue multi-server queueing system with time varying connectivities," in *Proc. of IEEE Int. Symp. on Inform. Theory (ISIT'10)*, Austin, TX, USA, June 2010.

[14] ——, "On the stability region of multi-queue multi-server queueing systems with stationary channel distribution," in *Proc. of IEEE Int. Symp. on Inform. Theory (ISIT'11)*, Saint Petersburg, Russia, Aug. 2011.

[15] S. Kittipiyakul and T. Javidi, "A fresh look at optimal subcarrier allocation in ofdma systems," in *Proc. IEEE Conference on Decision and Control*, Dec. 2004.

[16] H. Halabian, I. Lambadaris, C.-H. Lung, and A. Srinivasan, "Throughput-optimal relay selection in multiuser cooperative relaying networks," in *IEEE MILCOM 2010*, San Jose, CA, USA, Nov. 2010.

[17] M. J. Neely, "Dynamic power allocation and routing for satellite and wireless networks with time varying channels," Ph.D. dissertation, Massachusetts Institute of Technology, LIDS, 2003.

[18] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," *IEEE Journal on Selected Areas in Communications, Special Issue on Wireless Ad-hoc Networks*, vol. 23, no. 1, pp. 89–103, Jan. 2005.

[19] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource Allocation and Cross Layer Control in Wireless Networks*. Now Publisher, 2006.

[20] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, pp. 2:83–97, 1955.

[21] D. Stoyan, *Comparison Methods for Queues and other Stochastic Models*. Chichester: J. Wiley and Sons, 1983.

[22] S. M. Ross, *Stochastic Processes, 2nd ed.* New York: J. Wiley and Sons, 1996.

[23] R. Lidl and G. Pilz, *Applied abstract algebra, 2nd edition*. New York: Springer, 1998.

[24] H. Halabian, "Optimal server assignment in multi-server queueing systems with random connectivities," SCE-Carleton University, Tech. Rep., Sept. 2011.