1-2022

Managing stochastic bucket brigades on discrete work stations

Peng WANG
Singapore Management University, peng.wang.2016@pbs.smu.edu.sg

Kai PAN
Hong Kong Polytechnic University

Zhenzhen YAN
Nanyang Technological University

Yun Fong LIM
Singapore Management University, yflim@smu.edu.sg

Managing Stochastic Bucket Brigades on Discrete Work Stations

Peng WANG[1] • Kai PAN[2] • Zhenzhen YAN[3] • Yun Fong LIM[1*]

[1]Lee Kong Chian School of Business, Singapore Management University, Singapore 178899

[2]Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University, Kowloon, Hong Kong

[3]School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371

peng.wang.2016@pbs.smu.edu.sg • kai.pan@polyu.edu.hk • yanzz@ntu.edu.sg • yflim@smu.edu.sg

Jul 05, 2021

Abstract

Bucket brigades are notably used to coordinate workers in production systems. We study a J-station, I-worker bucket brigade system. The time duration for each worker to serve a job at a station is exponentially distributed with a rate that depends on the station's expected work content and the worker's work speed. Our goal is to maximize the system's productivity or to minimize its inter-completion time variability. We analytically derive the throughput and the coefficient of variation (CV) of the inter-completion time. We study the system under two cases. (i) If the work speeds depend only on the workers, the throughput gap between the stochastic and the deterministic systems can be up to 47% when the number of stations is small. Either maximizing the throughput or minimizing the CV of the inter-completion time, the slowest-to-fastest worker sequence always outperforms the reverse sequence for the stochastic bucket brigade. To maximize the throughput, more work content should be assigned to the stations near the faster workers. In contrast, to minimize the CV of the inter-completion time, more work content should be allocated to the stations near the slower workers. (ii) If the work speeds depend on the workers and the stations such that the workers may not dominate each other at every station, the asymptotic throughput can be expressed as a function of the average work speeds and the asymptotic expected blocked times of the workers, and can be interpreted as the sum of the effective production rates of all the workers.

Key words: Bucket brigade; Stochastic service time; Productivity; Variability

History: Received: April 2020; Accepted: July 2021 by Panos Kouvelis, after 2 revisions.

*Corresponding author

1 Introduction

To boost an assembly line's productivity, it is crucial to coordinate workers on the line such that their production capacity is effectively used. A well-known strategy is to coordinate the workers on the line as a *bucket brigade* (Bartholdi and Eisenstein, 1996a,b). When the workers are organized as a bucket brigade, each of them simultaneously assembles a single job (an instance of the product) along the line. Each worker carries and works on his job from work station to work station until either he hands off his job to a downstream co-worker or he completes his job at the end of the line. The worker then walks back to get another job, either from his co-worker upstream or from a buffer at the beginning of the line.

Bucket brigades are notably used in warehouses and distribution centers to coordinate workers for order-picking (Bartholdi et al., 2001). Companies that adopt the strategy include Ford, The Gap, and Walgreen's (Bartholdi and Eisenstein, 1996b). Bucket brigades are also applied to manufacturing environments where workers assemble jobs on discrete work stations. Examples include United Technologies Automotive (Villalobos et al., 1999a,b), Mitsubishi Consumer Electronics America, and Subway (Bartholdi and Eisenstein, 1996b). Bucket brigades are attractive in practice because they only require the workers to follow simple rules and the work-in-process is strictly under control by the number of workers.

More importantly, bucket brigades possess a *self-balancing* dynamic behavior that is first studied by Bartholdi and Eisenstein (1996a). In their normative model, they assume the work content is deterministic. Each worker proceeds forward with a deterministic, finite work velocity, and walks back instantaneously (with an infinite velocity). The authors proved that if the workers are sequenced from slowest to fastest in the production-flow direction based on their work velocities, then the hand-offs between any two adjacent workers will converge to a fixed location. As a result, every worker will repeatedly work on a fixed segment of the line eventually. This self-balancing behavior helps boost the productivity (in some cases, very significantly) of the implemented systems mentioned above (Bartholdi and Eisenstein, 1996b). Furthermore, it also allows a bucket brigade to spontaneously adapt to disruptions and seasonality.

Most papers in the literature study deterministic models of bucket brigades. However, many systems in practice exhibit a certain level of randomness in the service (or processing) times at work stations (Hopp and Spearman, 2008). To the best of our understanding, only two papers have analytically studied stochastic models of bucket brigades. Bartholdi et al. (2001) consider

an assembly line with stochastic work content at work stations. The authors assume that the work content at each station is exponentially distributed with a common mean and each worker has a constant work speed. They show that when the number of stations approaches infinity, the dynamics and throughput of the stochastic system will be similar to that of a deterministic system. In practice, many manufacturing systems have only a few work stations (Hopp and Spearman, 2008) and all order-picking lines in warehouses have a finite number of rack sections (Bartholdi et al., 2019). Thus, it is worthwhile and important to study a line with a small (finite) number of discrete work stations. Bukchin et al. (2018) consider a stochastic two-worker model with continuous work content along the line. They assume that upon each hand-off, the work speed of each worker is randomly re-generated from a distribution function, and the worker maintains this speed until the next hand-off. They find that a fastest-to-slowest sequence may be optimal as long as the standard deviation of the fastest worker's speed is sufficiently large.

If the service times at work stations are stochastic, will a bucket brigade converge to some "stationary state"? Will the system "balance" itself on a line with a *finite* number of stations as observed in the deterministic model? If the work speed of each worker varies over the stations, then how should we characterize the efficiency of each worker along the line? What is the effective production rate of each worker that contributes to the throughput of the line? These issues are not well studied in the literature.

In this paper, we consider a stochastic bucket brigade line with J stations and I workers. We assume that the time duration for each worker to serve a job at a station is exponentially distributed with a rate that depends on the station's expected work content and the worker's work speed. We analytically derive the system's average throughput and the coefficient of variation (CV) of the inter-completion time. If the work speeds are independent of the jobs, we prove that the stations where hand-offs occur follow a stationary probability distribution as the number of jobs approaches infinity. Furthermore, the average throughput and the CV of the inter-completion time converge to a constant.

We make other contributions by investigating the following two cases:

(i) The work speeds depend only on the workers. We assume the workers can be ranked from slowest to fastest. Interestingly, in this situation, the stationary probability distribution of the hand-off stations is analogous to the dynamic behavior of a deterministic model. Although the throughput difference between the stochastic and the deterministic systems

gets closer as the number of stations increases, the difference can be quite significant if the number of stations is small. Either maximizing the throughput or minimizing the CV of the inter-completion time, the slowest-to-fastest sequence always outperforms the reverse sequence for the stochastic bucket brigade. Furthermore, to maximize the throughput, more work content should be assigned to the stations near the faster workers. In contrast, to minimize the CV of the inter-completion time, more work content should be assigned to the stations near the slower workers.

(ii) The work speeds depend on the workers and the stations. Given that the workers may not dominate each other along the entire line, it becomes non-trivial to characterize the efficiency of each worker. We define the average work speed of each worker as a weighted average of his work speeds at all the stations. We also derive the expected blocked time of each worker along the line. The throughput of the stochastic bucket brigade can be expressed as a function of the average work speeds and the expected blocked times of the workers. Furthermore, the throughput can be interpreted as the sum of the effective production rates of all the workers.

Section 2 discusses the relevant literature. Section 3 specifies the assumptions and notation of our stochastic bucket brigade model. Section 4 derives the average throughput and the CV of the inter-completion time, and determines the system's asymptotic behavior as the number of jobs approaches infinity. Sections 5 and 6 study the above two cases in detail. Section 7 provides some concluding remarks. All proofs can be found in the online supplement.

2 Literature review

This paper is related to two streams of research: (i) bucket brigade assembly lines and (ii) dynamic server assignment on stochastic systems. We discuss each stream of work as follows.

2.1 Bucket brigade assembly lines

Bartholdi and Eisenstein (1996a) introduce a deterministic bucket brigade model. The authors find that an assembly line under the bucket brigade protocol can balance itself: If the workers are sequenced from slowest to fastest in the direction of the production flow, then the system always converges to a state where each worker repeatedly works on a fixed segment of the line.

Furthermore, if the work content is continuous along the line, then the system's throughput reaches a maximum level. Bartholdi et al. (1999) analyze the dynamics of a bucket brigade with two or three workers, and each worker has a constant work speed. For a two-worker line, they find that under the fastest-to-slowest worker sequence, the hand-offs between the two workers converge to a 2-cycle periodic state. Bartholdi and Eisenstein (2005) consider a model where each worker spends a constant walk-back time and a constant hand-off time to get work from his upstream colleague. They assume the worker's constant walk-back time is independent of his upstream colleague. They find that the system will still balance itself if the workers are sequenced from slowest to fastest. Bartholdi et al. (2009) generalize the bucket brigade protocol by allowing the workers to overtake and pass their colleagues. They show that a two-worker line may exhibit a chaotic behavior in which the hand-off locations never repeat on the line. More results of this model are summarized in Bartholdi et al. (2010).

Most of the above results are based on a line with continuous work content. Lim and Yang (2009) study a bucket brigade on discrete work stations. For a given work-content distribution on the stations, the authors identify the best cross-training and worker-sequencing strategy to maximize the system's throughput. They find that fully cross-training the workers and sequencing them from slowest to fastest may not be optimal. Other researchers generalize the model in different ways. Armbruster and Gel (2006) consider a two-worker bucket brigade in which workers' speeds do not dominate each other along the entire line. The authors present conditions under which bucket brigades are effective. Bartholdi et al. (2006) extend the ideas of bucket brigades to a network of sub-assembly lines so that all the sub-assembly lines are synchronized to produce at the same rate, and jobs are completed at regular, predictable time intervals. Webster et al. (2012) examine the performance of a bucket brigade order-picking system by varying the distribution of products along an aisle. Through simulations, the authors identify conditions in which the product distribution has large impact on the throughput.

Lim (2011) introduces a cellular bucket brigade, where each worker works on one side of an aisle when he proceeds in one direction and works on the other side of the aisle when he proceeds in the reverse direction. The idea is to eliminate the unproductive walk-back inherent in a traditional, serial bucket brigade. The author proposes extended rules to coordinate the workers under the new design, and identifies a sufficient condition for the cellular bucket brigade to self balance. Lim (2012) assesses the performance of cellular bucket brigades for warehouse

order-picking using data from a distribution center in North America. Lim (2017) analyzes a cellular bucket brigade in which any two adjacent workers may spend different time durations in a hand-off between them. Even with significant hand-off times, the cellular bucket brigade remains substantially more productive than a traditional bucket brigade especially if the team size is small and the workers' work speeds are close to their walk speed. Lim and Wu (2014) maximize the productivity of a cellular bucket brigade on a U-shape line with discrete work stations. They find conditions for the system to self balance.

Some papers study stochastic bucket brigade models. Bartholdi et al. (2001) assume that the time duration for a worker to serve a job at a station follows an exponential distribution with a rate that depends only on the worker. The authors prove that as the number of stations increases, the moment-to-moment behavior and the throughput of this stochastic model will increasingly resemble that of a deterministic model. Bukchin et al. (2018) consider a stochastic two-worker bucket brigade model with continuous work content. Immediately after each hand-off, they randomly re-generate each worker's work speed from a distribution function. The worker maintains this speed until the next hand-off. In contrast to these two papers, we assume that the time duration for each worker to serve a job at a station is exponentially distributed with a rate that depends on the station's expected work content and the worker's work speed. Furthermore, the work speed may depend on the worker, the station, and the job.

2.2 Dynamic server assignment on stochastic systems

Another stream of research studies dynamic server assignment on stochastic systems. Some papers in this stream *minimize holding costs* of systems with two stations. These include Harrison and López (1999), Williams (2000), Bell and Williams (2001), Ahn et al. (2004), and Mandelbaum and Stolyar (2004), which study flexible servers in parallel queues. Other examples are Rosberg et al. (1982), Farrar (1993), Iravani et al. (1997), Duenyas et al. (1998), Kaufman et al. (2005), and Armony et al. (2018), which study flexible servers in tandem queues. In contrast, our objective is to maximize the throughput or to minimize the CV of the inter-completion time.

Some papers find dynamic server assignment policies to *maximize throughput* of tandem lines with finite buffers. Andradóttir et al. (2001) consider a two-station, two-server Markovian system with service rates that are independent of the jobs. They identify an optimal policy that maximizes the long-run average throughput, and propose near-optimal heuristics for larger

systems. Andradóttir and Ayhan (2005) study the optimal policies for two-station Markovian systems with more servers than the stations. Kirkizlar et al. (2010) show that the policies in Andradóttir et al. (2001) and Andradóttir and Ayhan (2005) can also be effective for non-Markovian systems. Andradóttir et al. (2003) and Andradóttir et al. (2007a) study general queueing networks with infinite buffers, without or with server and station failures. Kirkizlar et al. (2014) consider the trade-off between throughput and holding costs. They maximize the long-run average profit of a serial system with finite buffers by finding an optimal server assignment policy. Isik et al. (2016) study dynamic server allocation for a tandem system with non-collaborative servers. They derive an optimal policy for Markovian systems with two servers and two stations, and propose heuristics for larger systems. In contrast to the above papers, we focus on the bucket brigade policy. Furthermore, we assume the service rates may depend on the jobs.

Some researchers investigate the benefits of partial flexibility in tandem systems. For example, Andradóttir et al. (2007b) study a Markovian system with two stations, and demonstrate that making only one server flexible when the buffer is sufficiently large can attain most of the benefits of full flexibility. Kirkizlar et al. (2012) study flexible servers in understaffed lines. They prove that the best possible production rate with full server flexibility and infinite buffers can be attained with partial flexibility and zero buffers. Hopp et al. (2004) consider a system with the same number of stations and servers under a constant work-in-process policy. They show that a skill-chaining strategy with two skills per server can outperform a "cherry picking" strategy that cross-trains some servers at bottleneck stations. In contrast, we assume each worker is fully cross-trained in our stochastic bucket brigade model. For a comprehensive review on cross-trained workforce, see Hopp and Van Oyen (2004).

3 Assumptions and notation

We consider a bucket brigade assembly line with work stations sequenced as $j = 1, \ldots, J$. Workers are sequenced as $i = 1, \ldots, I$ in the direction of the production flow. We call workers $i-1$ and $i+1$ the *predecessor* and the *successor*, respectively, of worker i. There are K jobs to be processed by the assembly line. Let $Z_{i,j}^{(k)}$ denote the time duration for worker i to serve (or process) job k at station j. We assume $Z_{i,j}^{(k)}$ follows an exponential distribution with rate $\mu_{i,j}^{(k)}$,

for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$. We define the rate $\mu_{i,j}^{(k)} = \left(s_j / v_{i,j}^{(k)}\right)^{-1}$, where s_j represents the *proportion of work content* at station j with $\sum_{j=1}^{J} s_j = 1$, and $v_{i,j}^{(k)}$ represents the work speed of worker i at station j on job k. We assume that all the workers are *fully cross-trained* such that $v_{i,j}^{(k)} > 0$, for all i, j, and k.

Defining the service (or processing) rates in this manner allows us to investigate the effects of the stations' expected work contents and the work speeds of the workers separately. Note that the proportion of work content s_j at station j is independent of job k. On the other hand, the work speed of worker i at station j may vary across the jobs. For example, some jobs are easier to process than other jobs because of the difference in the complexity levels or the material of the items. It is worth noting that the work speed $v_{i,j}^{(k)}$ is the inverse of the expected time to finish a *unit work content* at station j, whereas the service rate $\mu_{i,j}^{(k)}$ is the inverse of the expected time to finish a job at station j.

When worker i is working at any station, we assume the station's work content is *preemptible* such that worker $i+1$ can interrupt and take over the former's job. Worker i is *blocked* in front of a station if his job is ready to enter the station but some worker is still working at the station. We assume the workers spend negligible time when they move from one station to another and when they walk back to get more work. When the last worker I completes job k at the last station J, the system has its kth *reset*: Worker I walks back to take over the job of worker $I-1$, who in turn walks back to take over the job of worker $I-2$, and so on, until worker 1 initiates a new job at the beginning of the line. Since the workers spend negligible time to walk back, each reset is instantaneous.

Let $T^{(k)}$ denote the time point when worker I completes job k at the last station J. Let $H_i^{(k)}$ denote the station where worker i is working at immediately before $T^{(k)}$, for $i = 1, \ldots, I-1$. If worker i is blocked in front of station j immediately before $T^{(k)}$, then we set $H_i^{(k)} = j$. Thus, we have $1 \leq H_1^{(k)} \leq \ldots \leq H_{I-1}^{(k)} \leq J$. Since worker i hands off his job to worker $i+1$ at station $H_i^{(k)}$ in the kth reset, we call $H_i^{(k)}$ the kth *hand-off station* between workers i and $i+1$. Define $\boldsymbol{H}^{(k)} = \left(H_1^{(k)}, \ldots, H_{I-1}^{(k)}\right)$ as the kth *hand-off station vector*, for $k = 1, \ldots, K-1$. We set $T^{(0)} = 0$ and $\boldsymbol{H}^{(0)} = (1, \ldots, 1)$, which means that at time 0, worker I starts working on job 1 at station 1, while the first $I-1$ workers are blocked in front of station 1 (at the start of the line). Since the service times $Z_{i,j}^{(k)}$ are exponentially distributed and independent of each other, for all i, j, and k, the probability distribution of $\boldsymbol{H}^{(k)}$ can be determined by $\boldsymbol{H}^{(k-1)}$. Thus,

$\left\{\boldsymbol{H}^{(k)}, k = 1, \ldots, K-1\right\}$ is a Markov process.

Define $\mathcal{H} = \{\boldsymbol{h} = (h_1, \ldots, h_{I-1}) | 1 \leq h_1 \leq \ldots \leq h_{I-1} \leq J\}$ as a set of all possible hand-off station vectors. It is straightforward to show that the cardinality of $\mathcal{H}$ is $|\mathcal{H}| = \binom{I+J-2}{I-1}$. For any $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{H}$, we say $\boldsymbol{a} < \boldsymbol{b}$ if and only if there exists some m such that $a_m < b_m$ and $a_n = b_n$, for $n = m+1 \ldots, I-1$. We order the vectors in $\mathcal{H}$ such that $\boldsymbol{h}^1 < \boldsymbol{h}^2 < \ldots < \boldsymbol{h}^{|\mathcal{H}|}$, where $\boldsymbol{h}^1 = (1, \ldots, 1)$. For each $\boldsymbol{h}^n \in \mathcal{H}$, define $\pi_n^{(k)} = Pr\left\{\boldsymbol{H}^{(k)} = \boldsymbol{h}^n\right\}$ as the probability of $\boldsymbol{h}^n$ being the kth hand-off station vector. Define $\boldsymbol{\pi}^{(k)}$ as an $|\mathcal{H}|$-dimensional vector with its nth entry equals $\pi_n^{(k)}$. Note that $\boldsymbol{\pi}^{(k)}$ represents the probability distribution of the kth hand-off station vector $\boldsymbol{H}^{(k)}$. Since $\boldsymbol{H}^{(0)} = (1, \ldots, 1)$, we have $\boldsymbol{\pi}^{(0)} = (1, 0, ..., 0)$. For $\boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}$, define $p_{\boldsymbol{h},\boldsymbol{h}'}^{(k)} = Pr\left\{\boldsymbol{H}^{(k)} = \boldsymbol{h}' | \boldsymbol{H}^{(k-1)} = \boldsymbol{h}\right\}$ as the probability of $\boldsymbol{h}'$ being the kth hand-off station vector, conditioned on $\boldsymbol{h}$ being the $(k-1)$st hand-off station vector. Let $\boldsymbol{P}^{(k)} = \left(p_{\boldsymbol{h},\boldsymbol{h}'}^{(k)}\right)_{|\mathcal{H}| \times |\mathcal{H}|}$ denote the corresponding transition probability matrix.

4 Performance measures and asymptotic behavior

To derive the transition probability $p_{\boldsymbol{h},\boldsymbol{h}'}^{(k)}$, we need to analyze the movements of the I workers between the $(k-1)$st and the kth resets. Let $X(i)$ denote the station where worker i is located, for $i = 1, \ldots, I$. We set $X(I) = J+1$ when worker I finishes his job at station J. Define a *state* of the system as $\boldsymbol{X} = (X(1), X(2), \ldots, X(I))$. Recall that $H_i^{(k)}$ represents the kth hand-off station between workers i and $i+1$. Immediately after $T^{(k-1)}$, the state of the system is $\boldsymbol{X} = \left(1, H_1^{(k-1)}, \ldots, H_{I-1}^{(k-1)}\right)$. The workers keep working forward until $T^{(k)}$ when worker I finishes his job at station J, and the system's state becomes $\boldsymbol{X} = \left(H_1^{(k)}, \ldots, H_{I-1}^{(k)}, J+1\right)$. We call a state with $X(I) \leq J$ a *transient state*, and a state with $X(I) = J+1$ an *absorbing state*. Thus, immediately after the $(k-1)$st reset, the system progresses from one transient state to another transient state until it reaches an absorbing state when the kth reset occurs. Define $\mathcal{X} = \{\boldsymbol{x} = (x(1), \ldots, x(I)) | 1 \leq x(1) \leq x(2) \leq \ldots \leq x(I-1) \leq J,\ x(I-1) \leq x(I) \leq J+1\}$ as a set of all possible states between any two resets. Similar to $\mathcal{H}$, the cardinality of $\mathcal{X}$ is $|\mathcal{X}| = \binom{I+J-1}{I} + \binom{I+J-2}{I-1}$.

Between the $(k-1)$st and the kth resets, instead of keeping track of the system's state continuously, we only need to consider the *epochs* when a worker finishes his job at a station. Given that the service times are exponentially distributed, the probability that more than one worker finish their jobs at their respective stations at the same time is 0. For any transient state

$\boldsymbol{x}$, suppose there are $N(\boldsymbol{x})$ workers that are not blocked. Let $l_1, \ldots, l_{N(\boldsymbol{x})}$ denote these $N(\boldsymbol{x})$ workers from upstream to downstream. Note that worker I is never blocked, which implies $l_{N(\boldsymbol{x})} = I$. Thus, we have

$$1 \leq x(1) = \ldots = x(l_1) < x(l_1+1) = \ldots = x(l_2) < \ldots = x(l_{N(\boldsymbol{x})-1}) < x(l_{N(\boldsymbol{x})-1}+1) = \ldots = x(I) \leq J.$$

Among the $N(\boldsymbol{x})$ workers that are not blocked, one of them, say worker l_n, will finish his job in the next epoch. Let $\boldsymbol{x}_n$ denote the new state immediately after worker l_n finishes his job. We have $x_n(l_n) = x(l_n) + 1$, $x_n(i) = x(i)$, for $i \neq l_n, 1 \leq i \leq I$. Between the $(k-1)$st and the kth resets, the system progresses from state $\boldsymbol{x}$ to another state $\boldsymbol{x'}$. Let $q^{(k)}_{\boldsymbol{x},\boldsymbol{x'}}$ denote a one-step transition probability from $\boldsymbol{x}$ to $\boldsymbol{x'}$. For any transient state $\boldsymbol{x}$, we have

$$q^{(k)}_{\boldsymbol{x},\boldsymbol{x'}} = \begin{cases} \mu^{(k+I-l_n)}_{l_n,x(l_n)} / \sum_{m=1}^{N(\boldsymbol{x})} \mu^{(k+I-l_m)}_{l_m,x(l_m)}, & if\ \boldsymbol{x'} = \boldsymbol{x}_n, n = 1, \ldots, N(\boldsymbol{x}), \\ 0, & otherwise. \end{cases}$$

For any absorbing state $\boldsymbol{x}$, we set

$$q^{(k)}_{\boldsymbol{x},\boldsymbol{x'}} = \begin{cases} 1, & if\ \boldsymbol{x'} = \boldsymbol{x}, \\ 0, & otherwise, \end{cases}$$

such that the system stays in the absorbing state $\boldsymbol{x}$. Let $\boldsymbol{Q}^{(k)} = \left(q^{(k)}_{\boldsymbol{x},\boldsymbol{x'}}\right)_{|\mathcal{X}|\times|\mathcal{X}|}$ denote the corresponding one-step transition probability matrix. Starting from any state $\boldsymbol{x}$, the system will take at most $(J-1)I+1$ epochs to reach an absorbing state. Let $\boldsymbol{R}^{(k)} = \left(\boldsymbol{Q}^{(k)}\right)^{JI-I+1}$, and let $r^{(k)}_{\boldsymbol{x},\boldsymbol{x'}}$ denote the $\boldsymbol{x}$-$\boldsymbol{x'}$ entry of $\boldsymbol{R}^{(k)}$. As the system transitions from the hand-off station vector $\boldsymbol{h}$ immediately after the $(k-1)$st reset to the hand-off station vector $\boldsymbol{h'}$ at the kth reset, the system's state progresses from $(1, \boldsymbol{h})$ to $(\boldsymbol{h'}, J+1)$. Thus, we have the following lemma.

Lemma 1. *The probability of $\boldsymbol{h'}$ being the kth hand-off station vector, conditioned on $\boldsymbol{h}$ being the $(k-1)$st hand-off station vector is*

$$p^{(k)}_{\boldsymbol{h},\boldsymbol{h'}} = r^{(k)}_{(1,\boldsymbol{h}),(\boldsymbol{h'},J+1)}.$$

Since $\boldsymbol{\pi}^{(0)}$ is given and $\boldsymbol{\pi}^{(k)} = \boldsymbol{\pi}^{(k-1)}\boldsymbol{P}^{(k)}$, we can derive the probability distribution of $\boldsymbol{H}^{(k)}$ as $\boldsymbol{\pi}^{(k)} = \boldsymbol{\pi}^{(0)}\boldsymbol{P}^{(1)} \ldots \boldsymbol{P}^{(k)}$, for $k = 1, \ldots, K-1$.

Define $E\left[T^{(K)}\right]$ as the *expected makespan* of a bucket brigade assembly line with K jobs. To derive $E\left[T^{(k)}\right]$, we define $Y(k) = T^{(k)} - T^{(k-1)}$ as the inter-completion time between job $k-1$ and job k, which equals the total service time of worker I on job k. Suppose $\boldsymbol{H}^{(k-1)} = \boldsymbol{h}$,

we have $Y(k) = \sum_{j=h_{I-1}}^{J} Z_{I,j}^{(k)}$. Thus, we have

$$E\left[Y(k)|\boldsymbol{H}^{(k-1)} = \boldsymbol{h}\right] = \sum_{j=h_{I-1}}^{J} \frac{1}{\mu_{I,j}^{(k)}}, \quad \boldsymbol{h} \in \mathcal{H}. \tag{1}$$

Define $\boldsymbol{z}^{(k)}$ as an $|\mathcal{H}|$-dimensional column vector with its nth component equals $E\left[Y(k)|\boldsymbol{H}^{(k-1)} = \boldsymbol{h}^n\right]$, for $n = 1, \ldots, |\mathcal{H}|$. From Equation (1), we have

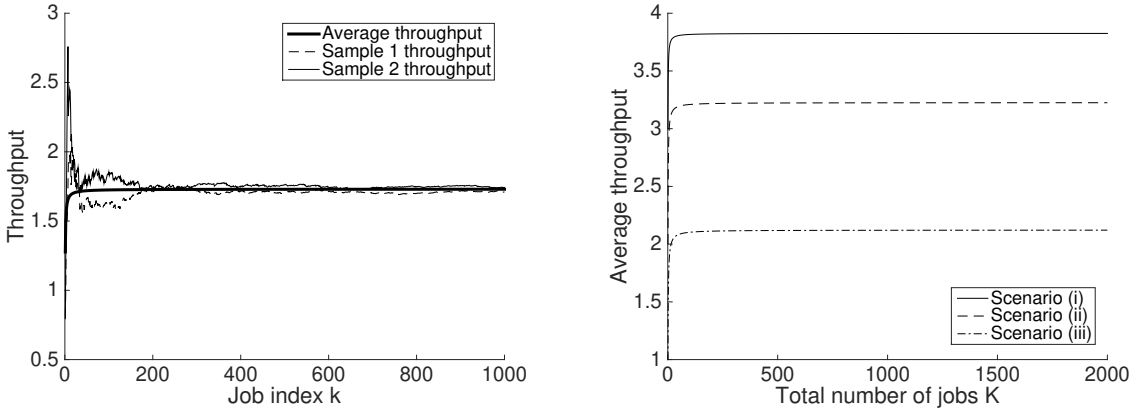
$$E\left[Y(k)\right] = \sum_{n=1}^{|\mathcal{H}|} E\left[Y(k)|\boldsymbol{H}^{(k-1)} = \boldsymbol{h}^n\right] \pi_n^{(k-1)} = \boldsymbol{\pi}^{(k-1)}\boldsymbol{z}^{(k)}. \tag{2}$$

Define the *average throughput* of a bucket brigade assembly line with K jobs as $\rho(K) = K/E\left[T^{(K)}\right]$. The following theorem determines the average throughput.

Theorem 1. *The average throughput of a bucket brigade assembly line with K jobs is*

$$\rho(K) = K \Big/ E\left[T^{(K)}\right] = K \Big/ \sum_{k=1}^{K} \boldsymbol{\pi}^{(0)}\boldsymbol{P}^{(1)} \cdots \boldsymbol{P}^{(k-1)}\boldsymbol{z}^{(k)} \,. \tag{3}$$

Given $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{z}^{(k)}$, we can determine the average throughput of the line using the transition probability matrices $\boldsymbol{P}^{(k)}$, for $k = 1, \ldots, K-1$. The bold solid line in Figure 1(a) represents the average throughput $\rho(k)$ of a bucket brigade assembly line up to job k determined by Equation (3). We set $I = 2, J = 3, s_1 = s_2 = s_3 = 1/3, K = 1,000$, and $v_{i,j}^{(k)} = i + j + k/K$, for $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, \ldots, K$. The dashed line and the thin solid line represent two sample paths of the *actual throughput* $k/T^{(k)}$ based on simulations.



(a) (b)

Figure 1: The throughput of a stochastic bucket brigade assembly line

For a special case in which the workers' service rates are independent of the jobs, we can drop the superscripts such that $\mu_{i,j}^{(k)} = \mu_{i,j}$, $\boldsymbol{P}^{(k)} = \boldsymbol{P}$, and $\boldsymbol{z}^{(k)} = \boldsymbol{z}$. Then, $\left\{\boldsymbol{H}^{(k)}, k = 1, \ldots, K-1\right\}$ becomes a homogeneous Markov process. Theorem 2 shows that, for this special case, the distribution of the kth hand-off station vector converges to a stationary distribution and the

average throughput converges to a constant as the number of jobs approaches infinity. Let $\boldsymbol{e}$ denote an $|\mathcal{H}|$-dimensional unit column vector.

Theorem 2. *If the workers' service rates are independent of the jobs, then* $\lim_{k\to\infty} \boldsymbol{\pi}^{(k)} = \boldsymbol{\pi}$, *where* $\boldsymbol{\pi}$ *is a unique stationary distribution that satisfies the equations* $\boldsymbol{\pi P} = \boldsymbol{\pi}$ *and* $\boldsymbol{\pi e} = 1$. *Furthermore, the asymptotic throughput* $\rho_\infty = \lim_{K\to\infty} \rho(K) = 1/(\boldsymbol{\pi z})$, *where* $\boldsymbol{\pi z}$ *represents the asymptotic expected inter-completion time.*

It is interesting to see that the convergence to the stationary distribution is independent of the sequence of the workers. While we need to sequence the workers from slowest to fastest in the direction of the production flow for the deterministic bucket brigade to self balance (Bartholdi and Eisenstein, 1996a), we do not need such a worker sequence for the stochastic bucket brigade to converge to the stationary distribution.

Figure 1(b) shows an example with $I = 3, J = 5, s_1 = s_2 = 0.1$, $s_3 = s_4 = 0.3$, and $s_5 = 0.2$. We consider three different scenarios: (i) $v_{1,j} = 1$, $v_{2,j} = 2$, $v_{3,j} = 3$, (ii) $v_{1,j} = v_{2,j} = v_{3,j} = 2$, and (iii) $v_{1,j} = 3$, $v_{2,j} = 2$, $v_{3,j} = 1$. The average throughput $\rho(K)$ for each scenario converges to a constant as K increases, which is consistent with Theorem 2. Furthermore, the average throughput decreases as the worker sequence changes from slowest-to-fastest (scenario (i)) to fastest-to-slowest (scenario (iii)). Section 5.2 demonstrates that the slowest-to-fastest sequence is always more productive than the reverse sequence for the stochastic bucket brigade.

Using the probability distribution $\boldsymbol{\pi}^{(k)}$ of the kth hand-off station vector, we can also derive the variance of the inter-completion time $Y(k)$. Let h_i^n denote the ith component of $\boldsymbol{h}^n$.

Lemma 2. *The variance of the inter-completion time* $Y(k)$ *can be determined as*

$$Var\left(Y(k)\right) = \sum_{n=1}^{|\mathcal{H}|} \pi_n^{(k-1)} \sum_{j=h_{I-1}^n}^{J} \left(\frac{1}{\mu_{I,j}^{(k)}}\right)^2 + \sum_{n=1}^{|\mathcal{H}|} \pi_n^{(k-1)} \left(\sum_{j=h_{I-1}^n}^{J} \frac{1}{\mu_{I,j}^{(k)}}\right)^2 - \left(\sum_{n=1}^{|\mathcal{H}|} \pi_n^{(k-1)} \sum_{j=h_{I-1}^n}^{J} \frac{1}{\mu_{I,j}^{(k)}}\right)^2. \quad (4)$$

We define the *coefficient of variation* (CV) of the inter-completion time $Y(k)$ as $CV = \sqrt{Var\left(Y(k)\right)}$ $/E\left[Y(k)\right]$, where $E\left[Y(k)\right]$ and $Var\left(Y(k)\right)$ are determined by Equations (2) and (4) respectively. Note that a small CV of the inter-completion time ensures a more predictable output process of the line, which facilitates planning of the downstream processes of the supply chain. Lemma 2 implies the following theorem.

Theorem 3. *If the service rates are independent of the jobs, and $\boldsymbol{\pi}$ is the stationary distribution of the hand-off station vectors. The asymptotic variance of the inter-completion time is*

$$\lim_{k\to\infty} Var\left(Y(k)\right) = \sum_{n=1}^{|\mathcal{H}|} \pi_n \sum_{j=h_{I-1}^n}^{J} \left(\frac{1}{\mu_{I,j}}\right)^2 + \sum_{n=1}^{|\mathcal{H}|} \pi_n \left(\sum_{j=h_{I-1}^n}^{J} \frac{1}{\mu_{I,j}}\right)^2 - \left(\sum_{n=1}^{|\mathcal{H}|} \pi_n \sum_{j=h_{I-1}^n}^{J} \frac{1}{\mu_{I,j}}\right)^2. \quad (5)$$

The asymptotic CV of the inter-completion time equals

$$\lim_{k\to\infty} \frac{\sqrt{Var\left(Y(k)\right)}}{E\left[Y(k)\right]} = \frac{1}{\boldsymbol{\pi z}} \left[\sum_{n=1}^{|\mathcal{H}|} \pi_n \sum_{j=h_{I-1}^n}^{J} \left(\frac{1}{\mu_{I,j}}\right)^2 + \sum_{n=1}^{|\mathcal{H}|} \pi_n \left(\sum_{j=h_{I-1}^n}^{J} \frac{1}{\mu_{I,j}}\right)^2 - \left(\sum_{n=1}^{|\mathcal{H}|} \pi_n \sum_{j=h_{I-1}^n}^{J} \frac{1}{\mu_{I,j}}\right)^2\right].$$

5 Case I: The work speeds depend only on the workers

In this section, we consider a special case in which $v_{i,j}^{(k)} = v_i$, for all i, j, and k.

5.1 The distribution of the hand-off station vector

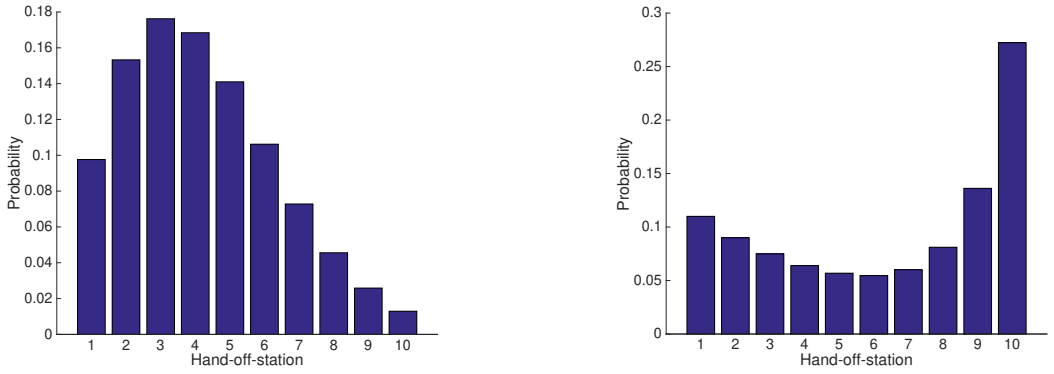
Since each worker's service time at each station is stochastic, the hand-offs between any two neighboring workers will not converge to a fixed location (station) as observed in the deterministic model studied by Bartholdi and Eisenstein (1996a). Instead, Theorem 2 shows that the probability distribution of the kth hand-off station vector $\boldsymbol{H}^{(k)}$ always converges to a stationary distribution $\boldsymbol{\pi}$ as k goes to infinity, independent of the worker sequence.

Figure 2(a) shows the stationary distribution of a ten-station two-worker line with $s_j = 0.1$, $j = 1, \ldots, 10$. The workers are sequenced from slowest to fastest with $v_1 = 1$ and $v_2 = 2$. The stationary distribution of the hand-off station is unimodal with a single peak at station 3. This is analogous to the behavior of the deterministic model in which the hand-offs between the two workers converge to a fixed location (Bartholdi and Eisenstein, 1996a).

Figure 2(b) shows the stationary distribution of the hand-off station when the workers are sequenced from fastest to slowest with $v_1 = 2$ and $v_2 = 1$. The stationary distribution has two peaks at stations 1 and 10. This is analogous to the dynamic behavior of the deterministic model studied by Bartholdi et al. (1999) in which the hand-offs converge to a 2-cycle.

5.2 Comparing with the deterministic system

To examine the effect of the random service times, we compare the stochastic system with the deterministic system with discrete work stations. We consider two workers with work



(a) $v_1 = 1, v_2 = 2$ (b) $v_1 = 2, v_2 = 1$

Figure 2: The stationary distributions of the hand-off station for a line with ten stations and two workers

speeds equal to 1 and 2. The expected work content is identical for all the stations such that $s_j = 1/J, j = 1, \ldots, J$. We compare four different settings: A deterministic system with a slowest-to-fastest worker sequence (denoted as d-SF), a stochastic system with a slowest-to-fastest worker sequence (denoted as s-SF), a deterministic system with a fastest-to-slowest worker sequence (denoted as d-FS), and a stochastic system with a fastest-to-slowest worker sequence (denoted as s-FS). We examine the asymptotic throughput of each setting as the number of jobs K approaches infinity. The asymptotic throughput of each stochastic system is determined by Theorem 2, and the asymptotic throughput of each deterministic system can be derived analytically. Figure 3 shows the asymptotic throughput of each setting as the number of stations J varies from 4 to 50. As J increases, the throughput difference between the deterministic and the stochastic systems becomes smaller for both worker sequences. This is consistent with the result of Bartholdi et al. (2001). The d-SF setting achieves the maximum throughput v_1+v_2, but the s-SF setting cannot always achieve the maximum throughput because of blocking. However, the s-SF setting's performance approaches that of the d-SF setting as J increases.

Under the fastest-to-slowest worker sequence, the asymptotic throughput of the d-FS setting decreases with J, whereas the asymptotic throughput of the s-FS setting first increases and then slightly decreases with J. Note that under the fastest-to-slowest sequence, the stochastic system (s-FS) may outperform the deterministic system (d-FS). For $J \geq 4$, the d-FS setting converges to a period-2 cycle as $K \to \infty$ with hand-offs alternating between $2/J$ and $(J-1)/J$. Thus, the asymptotic throughput equals $2 + 2/(J-1)$, which approaches 2 as J increases. Note

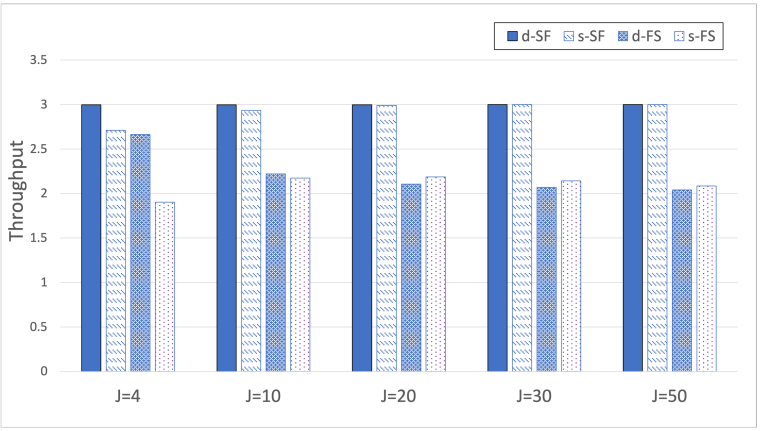


Figure 3: The asymptotic throughput of the deterministic and the stochastic systems for different numbers of stations

that the asymptotic throughput approaches twice the speed of the slower worker as J increases because the line is becoming more like a continuous line and the faster worker upstream is constantly blocked by the slower worker downstream. In contrast, the s-FS setting suffers from severe blocking if J is small. Its asymptotic throughput is lower than 2 when $J = 4$. As J increases, blocking is mitigated and the s-FS setting can achieve a throughput greater than 2. However, as J further increases, the stochastic system resembles the deterministic system, and the throughput drops to 2.

From the above comparison between the stochastic and the deterministic models with discrete work stations, we obtain the following insights. Under the slowest-to-fastest sequence, the deterministic system is more productive than the stochastic system. This suggests that the stochastic service times cause throughput loss for a bucket brigade line with discrete work stations under the slowest-to-fastest sequence. However, under the reverse sequence, the stochastic system may outperform the deterministic system (see Figure 3). For both worker sequences, the performance difference between the stochastic and the deterministic systems gets closer as J increases, which is consistent with the finding of Bartholdi et al. (2001).

Figure 4 shows the asymptotic throughput of each setting for different work-content distributions. We consider three stations and two workers with work speeds 1 and 2. Under each worker sequence, the deterministic system always outperforms the stochastic system.

For the deterministic system, the relative performance of the slowest-to-fastest and the reverse sequences depends on the work-content distribution: If $s_1 = s_2 = 1/3$, both sequences achieve the maximum throughput 3. If $s_1 = 1/3$ and $s_2 = 7/12$, only the slowest-to-fastest

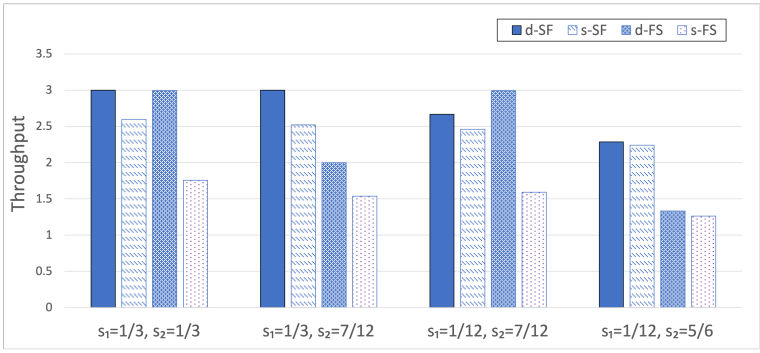


Figure 4: The asymptotic throughput of the deterministic and the stochastic systems for different work-content distributions

sequence (d-SF) achieves the maximum throughput. If $s_1 = 1/12$ and $s_2 = 7/12$, only the fastest-to-slowest sequence (d-FS) achieves the maximum throughput. If $s_1 = 1/12$ and $s_2 = 5/6$, both sequences cannot achieve the maximum throughput. In contrast, for the stochastic system, we show in the following that the slowest-to-fastest sequence (s-SF) always outperforms the reverse sequence (s-FS) under all work-content distributions.

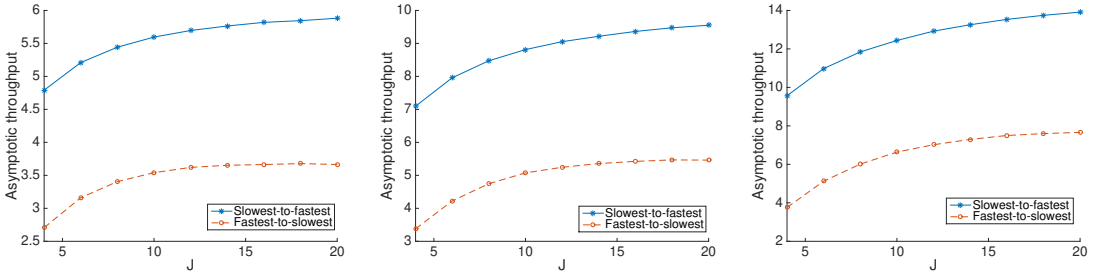
Lemma 3. *For a three-station two-worker line, with $v_1 + v_2$ held constant, the asymptotic throughput of the stochastic bucket brigade system increases with v_2.*

Lemma 3 implies the following corollary.

Corollary 1. *For a three-station two-worker stochastic bucket brigade system, the slowest-to-fastest sequence is always more productive than the fastest-to-slowest sequence.*

Figure 4 and Lim and Yang (2009) show that the slowest-to-fastest sequence can be outperformed by the reverse sequence for a *deterministic* three-station line (see Figure 6(a) of Lim and Yang (2009)). However, for a three-station line with stochastic service times, Corollary 1 shows that the slowest-to-fastest sequence is always more productive.

We further test numerically that whether the result of Corollary 1 continues to hold for a line with more work stations and workers. Figure 5 shows the asymptotic throughput under both worker sequences for 3, 4, and 5 workers, and the number of work stations J varies from 4 to 20. In each graph, we set the work speeds of the I workers equal $1, 2, \ldots, I$. We consider evenly distributed work content such that $s_j = 1/J$, $j = 1, 2, \ldots, J$. We observe that the slowest-to-fastest sequence is always more productive than the reverse sequence in these larger systems with stochastic service times.



(a) $I = 3$ (b) $I = 4$ (c) $I = 5$

Figure 5: The asymptotic throughput for different numbers of workers

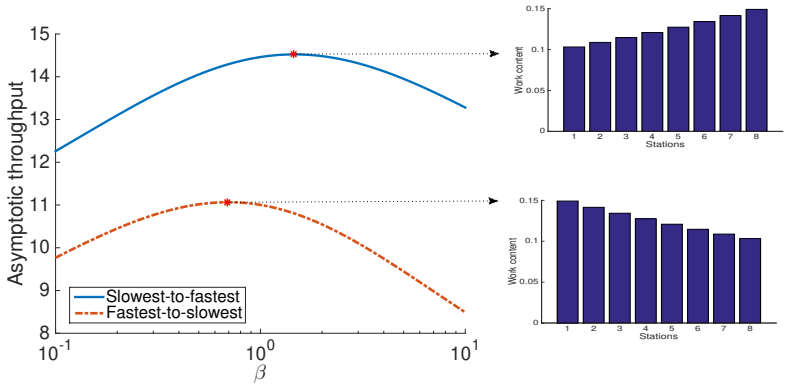
As suggested by Figures 3 and 4, if the number of stations is small, the throughput difference between the stochastic and the deterministic systems can be *quite significant* (the gap is up to 47%). Furthermore, Corollary 1 and Figure 5 show that the slowest-to-fastest sequence is more productive than the reverse sequence for the stochastic system. However, this result may not hold for a deterministic system (see Figure 6(a) of Lim and Yang (2009)). Thus, it is important to study the stochastic bucket brigade model especially for a line with a small number of stations.

5.3 Maximizing the throughput

Theorem 2 allows us to maximize the asymptotic throughput of the stochastic bucket brigade by optimizing the expected work contents $s_1, \ldots, s_J$, and the sequence of the workers. For illustration purposes, we assume $s_j/s_{j-1} = \lambda$, $j = 2, \ldots, J$. We define $\beta = s_J/s_1 = \lambda^{J-1}$. If $\beta > 1$, the work content s_j is increasing in j. If $\beta = 1$, the work content is evenly distributed over the stations. If $\beta < 1$, the work content s_j is decreasing in j.

Figure 6(a) shows the asymptotic throughput of a line with $J = 8$ stations and $I = 4$ workers. We assume the work speeds of the workers are 3, 4, 5 and 6. The slowest-to-fastest sequence (solid line) always outperforms the reverse sequence (dashed line). Under each worker sequence, an *overly skewed* work-content distribution over the stations (when β is too small or too large) is not productive because of severe blocking. As a result, for each worker sequence in Figure 6(a), the asymptotic throughput is unimodal in β. For the slowest-to-fastest sequence, the asymptotic throughput reaches the maximum when $\beta = 1.45$, corresponding to the top graph of Figure 6(b) where the work content s_j increases with j. For the fastest-to-slowest sequence, the maximum asymptotic throughput occurs at $\beta = 0.69$, which corresponds to the bottom graph of Figure 6(b) where the work content decreases from upstream to downstream.

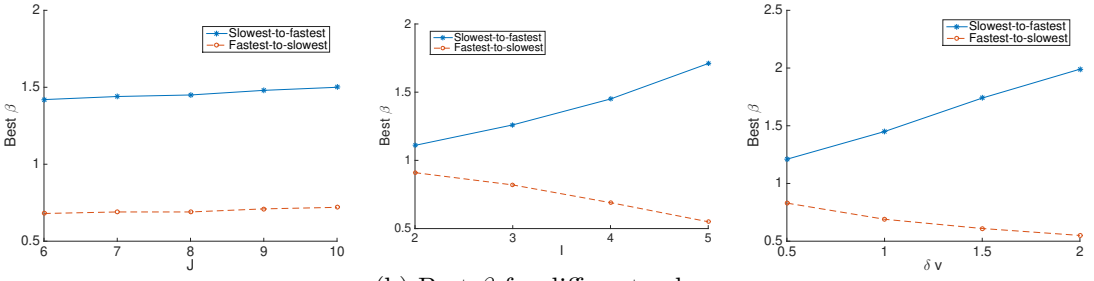
Figure 6(b) suggests that *to maximize the asymptotic throughput, more work content should be assigned to the stations near the faster workers.*



(a) (b)

Figure 6: (a) Asymptotic throughput of a line with eight stations and four workers (b) Work-content distributions with $\beta = 1.45$ (top) and $\beta = 0.69$ (bottom)

To test the robustness of the above results, we vary the number of stations, the number of workers, and the difference in the work speeds of the workers. In Figure 7(a), we consider the same team of four workers in Figure 6. We vary the number of stations J from 6 to 10. For each J, we identify the best β for both the slowest-to-fastest and the reverse sequences. We find that the best β for the slowest-to-fastest sequence is always greater than 1, implying that the work content s_j increases with j. In contrast, the best β for the reverse sequence is always less than 1, implying that s_j is decreasing in j.



(a) Best β for different values of J (b) Best β for different values of I (c) Best β for different values of δv

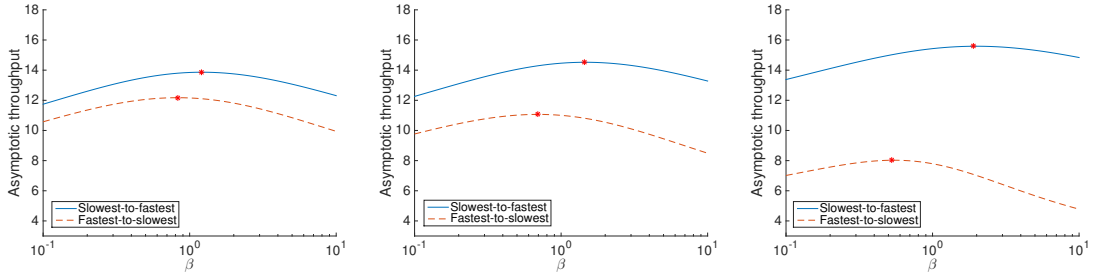
Figure 7: Best β for maximizing the asymptotic throughput under different parameter settings

Figure 7(b) shows the best β by varying the number of workers I from 2 to 5. To be comparable with Figure 6, we consider eight stations and assume $V \equiv \sum_{i=1}^{I} v_i = 18$, where $v_i = V/I - (I+1)/2 + i$ for the slowest-to-fastest sequence, and $v_i = V/I + (I+1)/2 - i$ for the fastest-to-slowest sequence, for $i = 1, \ldots, I$. Thus, $\delta v \equiv |v_i - v_{i-1}| = 1$, $i = 2, \ldots, I$, for

both sequences. We find that for each I, the best β for the slowest-to-fastest sequence is always greater than 1. Furthermore, the best β increases with I (the skewness of the work-content distribution increases with I) because the gap between the work speeds of the first and the last workers increases with I. In contrast, the best β for the fastest-to-slowest sequence is always less than 1, and the best β decreases with I because of the same reason.

Figure 7(c) shows the best β by varying the work-speed difference δv. We consider eight stations and four workers with $\sum_{i=1}^{4} v_i = 18$. We find that the best β for the slowest-to-fastest sequence is always greater than 1. The best β increases with δv because the gap between the work speeds of the first and the last workers increases with δv. On the other hand, the best β for the fastest-to-slowest sequence is always less than 1 and it decreases with δv because of the same reason.

To further investigate the impact of δv, Figure 8 shows the asymptotic throughput of each worker sequence for $\delta v = 0.5$, 1, and 2, with $J = 8$, $I = 4$, and $\sum_{i=1}^{I} v_i = 18$. We see that the gap between the asymptotic throughputs of the two sequences becomes larger as δv gets larger.



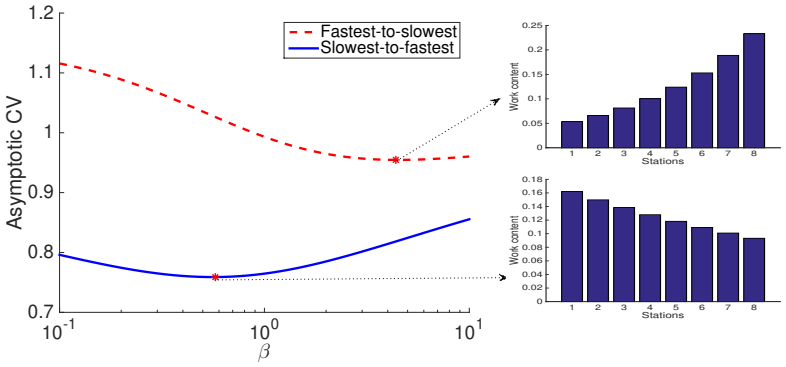
(a) $\delta v = 0.5$ (b) $\delta v = 1$ (c) $\delta v = 2$

Figure 8: The gap between the two sequences' asymptotic throughputs becomes larger as δv increases

5.4 Minimizing the CV of the inter-completion time

Theorem 3 enables us to minimize the CV of the inter-completion time by optimizing the work-content distribution over the stations. A small CV of the inter-completion time ensures a more predictable output process of the line, which facilitates planning of the downstream processes of the supply chain. Moreover, a small CV of the inter-completion time implies that the hand-offs between any two consecutive workers occur at close to the same point upon each reset. As a

result, the portion of tasks that each worker performs tends to be repeated. The system can benefit from the workers' faster learning that leads to higher productivity. Based on the same definition of $\beta = s_J/s_1 = \lambda^{J-1}$ as in Section 5.3, Figure 9(a) shows the asymptotic CV of the inter-completion time under the same setting as in Figure 6(a). The slowest-to-fastest sequence (solid line) always has a smaller asymptotic CV of the inter-completion time than the reverse sequence (dashed line). Moreover, for each worker sequence the CV is unimodal in β. This suggests that an overly skewed work-content distribution (when β is too small or too large) causes severe blocking, resulting in a higher CV of the inter-completion time. For the fastest-to-slowest sequence, the asymptotic CV of the inter-completion time reaches the minimum when $\beta = 4.37$. This corresponds to the top graph of Figure 9(b), where the work content s_j increases with j. For the slowest-to-fastest sequence, the minimum asymptotic CV of the inter-completion time occurs at $\beta = 0.58$, corresponding to the bottom graph of Figure 9(b), where the work content decreases from upstream to downstream. Figure 9(b) suggests that *we can minimize the CV of the inter-completion time by allocating more work content to the stations near the slower workers.*

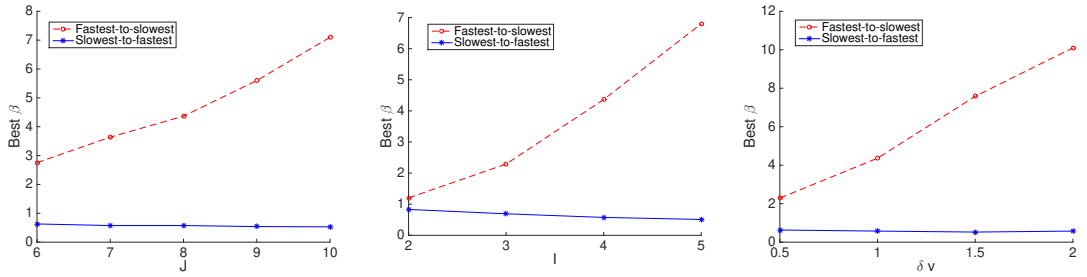


(a) (b)

Figure 9: (a) Asymptotic CV of the inter-completion time of a line with 8 stations and 4 workers (b) Work-content distributions with $\beta = 4.37$ (top) and $\beta = 0.58$ (bottom)

To test the robustness of the above results, we vary the number of stations, the number of workers, and the difference in the work speeds of the workers. In Figure 10(a), we consider the same team of four workers as in Figure 9. We vary the number of stations J from 6 to 10. For each J, we identify the best β for both the slowest-to-fastest and the reverse sequences. We find that the best β for the slowest-to-fastest sequence is always less than 1, and the best β decreases with J. In contrast, the best β for the fastest-to-slowest sequence is always greater

than 1, and it increases with J. Similarly, Figure 10(b) identifies the best β by varying the number of workers I from 2 to 5. The setting is the same as Figure 7(b). Figure 10(c) shows the best β by varying the work-speed difference δv. The setting is the same as Figure 7(c). We find similar behavior of the best β as I or δv increases.



(a) Best β for different values of J (b) Best β for different values of I (c) Best β for different values of δv

Figure 10: Best β for minimizing the asymptotic CV of the inter-completion time under different parameter settings

Combining the results in Sections 5.3 and 5.4, we obtain the following managerial insights. Either maximizing the asymptotic throughput or minimizing the asymptotic CV of the inter-completion time, the slowest-to-fastest sequence always outperforms the reverse sequence for the stochastic bucket brigade. Furthermore, to maximize the asymptotic throughput, more work content should be assigned to the stations near the faster workers (see Figure 6(b)). However, to minimize the asymptotic CV of the inter-completion time, more work content should be assigned to the stations near the slower workers (see Figure 9(b)).

Note that the advantages of a small CV of the inter-completion time are important in some settings, but not in others. For example, achieving a small CV may be important for an assembly line where tasks vary along the line, so that each worker can learn faster from repeatedly working on the same portion of the tasks. However, it may not be as important in other settings such as order-picking in a warehouse where workers simply pick items from a pick-to-light system. In such a setting, it is possible that no significant learning or speed-up is gained by a worker picking from the same section of the warehouse. In general, maximizing the throughput is a primary objective of the management, whereas the CV of the inter-completion time may play a secondary role in some settings where workers are able to learn if their task range is limited.

6 Case II: The work speeds depend only on the workers and the stations

In this section, we assume the work speed of each worker depends on the worker and the stations, but is independent of the jobs such that $v_{i,j}^{(k)} = v_{i,j}$, for all k, $i = 1, \ldots, I$, $j = 1, \ldots, J$. It is challenging to analyze this case because a worker i may not dominate another worker i' along the entire line. That is, worker i is faster at some stations, but slower at other stations than worker i'. In this situation, it is not clear how we should sequence the workers along the line.

As the work speed of each worker may vary across the stations, we define the *average work speed* of worker i along the entire line as

$$\nu_i = \frac{\sum_{j=1}^{J} f_{i,j} s_j}{\sum_{j=1}^{J} f_{i,j} s_j / v_{i,j}}, \tag{6}$$

where $f_{i,j}$ represents the probability of worker i finishing his job at station j between two consecutive resets under the stationary distribution $\boldsymbol{\pi}$. Note that if $v_{i,j} = v_i$, we have $\nu_i = v_i$. Recall that h_i^n denotes the ith component of $\boldsymbol{h}^n$. For convenience, we set $h_0^n = 1$ and $h_I^n = J+1$. Let $f_{i,j}^{(k)}$ denote the probability of worker i finishing his job at station j between the $(k-1)$st and the kth resets. The following lemma determines $f_{i,j}^{(k)}$ and $f_{i,j}$.

Lemma 4. *The probability of worker i finishing his job at station j between the $(k-1)$st and the kth resets can be obtained as*

$$f_{i,j}^{(k)} = \sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_i^n \geq j+1}} \pi_n^{(k)} - \sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_{i-1}^n \geq j+1}} \pi_n^{(k-1)}. \tag{7}$$

As k approaches infinity, the probability of worker i finishing his job at station j between two consecutive resets under the stationary distribution $\boldsymbol{\pi}$ can be obtained as

$$f_{i,j} = \sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_i^n \geq j+1}} \pi_n - \sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_{i-1}^n \geq j+1}} \pi_n. \tag{8}$$

Define the expected blocked time of worker i as the expected total time that worker i is blocked between two consecutive resets. Let B_i denote the *asymptotic expected blocked time* of worker i as K approaches infinity. Since worker I is never blocked, we have $B_I = 0$. Recall that the inter-completion time $Y(k) = T^{(k)} - T^{(k-1)}$. Let $Y_\infty = \lim_{k\to\infty} E[Y(k)]$ denote the asymptotic inter-completion time of jobs. According to Theorem 2, $Y_\infty = \boldsymbol{\pi} \boldsymbol{z} = 1/\rho_\infty$.

Lemma 5. *The asymptotic expected blocked time of worker i can be expressed as*

$$B_i = Y_\infty - \sum_{j=1}^{J} \frac{f_{i,j}}{\mu_{i,j}}, \quad i = 1, \ldots, I-1. \tag{9}$$

The following theorem expresses the asymptotic throughput of the line as a function of the average work speeds and the asymptotic expected blocked times of the workers.

Theorem 4. *If the work speeds of the workers are independent of the jobs, the asymptotic throughput of a stochastic bucket brigade can be expressed as*

$$\rho_\infty = \sum_{i=1}^{I} \frac{Y_\infty - B_i}{Y_\infty} \nu_i. \tag{10}$$

Equation (10) can be interpreted as follows. The term $\frac{Y_\infty - B_i}{Y_\infty}$ represents the fraction of the inter-completion time that worker i is not blocked on the line, and $\frac{Y_\infty - B_i}{Y_\infty}\nu_i$ represents the effective production rate of worker i without being blocked. The asymptotic throughput of the line is the sum of the effective production rates of all the workers.

7 Conclusion

The literature of bucket brigades with stochastic service times is very limited. To the best of our understanding, Bartholdi et al. (2001) is the only paper that analytically studies bucket brigades on discrete work stations with stochastic service times. The authors assume that the work speed of each worker at each station depends only on the worker. In this paper, we extend the work of Bartholdi et al. (2001) by assuming that the work speed of each worker at each station on a job depends not only on the worker, but also on the station and the job. Thus, the workers may not dominate each other at all the stations, and their work speeds may change with the jobs. We consider a bucket brigade line with J stations and I workers. We assume the time duration for each worker to finish a job at each station is exponentially distributed with a rate that depends on the station's expected work content and the work speed of the worker.

By observing the Markov property of the hand-off stations and analyzing the transition from one hand-off station vector to the next, we are able to derive the system's average throughput (Theorem 1) and the CV of the inter-completion time (Lemma 2). We prove that *if the work speeds of the workers are independent of the jobs, then the probability distribution of the hand-off station vector will converge to a unique stationary distribution as the number of jobs approaches*

infinity (Theorem 2). Furthermore, the average throughput and the CV of the inter-completion time converge to a constant that depends on the stationary distribution (Theorems 2 and 3).

We first study a case where each worker's work speeds depend only on the worker. For a two-worker system in which one worker has a larger speed than the other at all the stations, we find that the probability distribution of the hand-off station is analogous to the behavior of the deterministic model. If the workers are sequenced from slowest to fastest, the probability distribution has a peak in the middle of the line. It is interesting to note that this result is consistent with that of the deterministic model in which the hand-offs converge to a fixed location (Bartholdi and Eisenstein, 1996a). However, if the workers are sequenced from fastest to slowest, the probability distribution has two peaks at the two ends of the line. This result is analogous to that of the deterministic model studied by Bartholdi et al. (1999) in which the hand-offs converge to a 2-cycle.

We further compare the throughput of our stochastic model with the deterministic model with discrete work stations. Under the slowest-to-fastest sequence, the deterministic system is more productive than the stochastic system. This suggests that the stochastic service times cause throughput loss for a bucket brigade line with discrete work stations under the slowest-to-fastest sequence. However, under the reverse sequence, the stochastic system may outperform the deterministic system. For both worker sequences, the performance difference between the stochastic and the deterministic systems gets closer as the number of stations J increases, which is consistent with the finding of Bartholdi et al. (2001).

However, if the number of stations is small, the throughput difference between the stochastic and the deterministic systems can be *quite significant* (the gap is up to 47%). Furthermore, for a stochastic system, the slowest-to-fastest sequence is more productive than the reverse sequence (see Corollary 1 and Figure 5). It is worth noting that this result may not hold for a deterministic system (see Figure 6(a) of Lim and Yang (2009)). Thus, it is worthwhile and important to study the stochastic bucket brigade model especially for a line with a small number of stations.

To maximize the asymptotic throughput, the manager can optimize both the worker sequence and the work-content distribution over the stations. We demonstrate that the slowest-to-fastest sequence always outperforms the reverse sequence (Figure 6(a)). Furthermore, under the slowest-to-fastest sequence, the asymptotic throughput is maximized if the expected work

content of the stations increases in the direction of the production flow. In contrast, under the fastest-to-slowest sequence, the asymptotic throughput is maximized if the work content decreases from upstream to downstream. These results also hold when we vary the number of stations, the number of workers, or the difference in the work speeds of the workers.

In terms of minimizing the asymptotic CV of the inter-completion time, we also find that the slowest-to-fastest sequence always outperforms the reverse sequence. For the fastest-to-slowest sequence, the asymptotic CV is the minimum when the expected work content increases from upstream to downstream. However, for the slowest-to-fastest sequence, the asymptotic CV reaches the minimum when the expected work content decreases in the direction of the production flow. The above results are robust when we vary the number of stations, the number of workers, or the difference in the work speeds of the workers.

From the above results, we obtain the following important managerial insights. *Either maximizing the asymptotic throughput or minimizing the asymptotic CV of the inter-completion time, the slowest-to-fastest sequence always outperforms the reverse sequence for the stochastic bucket brigade. Furthermore, to maximize the asymptotic throughput, more work content should be assigned to the stations near the faster workers (Figure 6(b)). In contrast, to minimize the asymptotic CV of the inter-completion time, more work content should be assigned to the stations near the slower workers (Figure 9(b)).*

We then analyze another case where the work speeds depend on the workers and the stations. Given that the workers may not dominate each other along the entire line, we define the average work speed of each worker as a weighted average of his work speeds at all the stations (Equation (6)). We also derive the asymptotic expected blocked time of each worker (Equation (9)). The asymptotic throughput of the stochastic bucket brigade can be expressed as a function of the average work speeds and the asymptotic expected blocked times of the workers, and can be interpreted as the sum of the effective production rates of all the workers (Theorem 4).

Our methodology can be generalized to a case where the work speeds depend on the workers, the stations, and the jobs. It is also worth noting that in our stochastic bucket brigade model, we consider exponential service times that have a coefficient of variation equal to 1 (Hopp and Spearman, 2008). In practice, the service time of a worker at a station may not be as variable. Choosing between the deterministic model and the stochastic model with exponential service times for practical purposes becomes an interesting future research direction.

Acknowledgments

The authors thank the senior editor and the two anonymous referees for their valuable comments that have substantially improved the paper. Portions of this paper have been presented in the National University of Singapore, 2019; INFORMS Annual Meeting, Seattle, USA, 2019; MSOM Conference, Singapore, 2019; Mostly OM Workshop, Shenzhen, China, 2019; and POMS-HK International Conference, Hong Kong, China, 2019. The authors thank the audiences for many insightful comments and stimulating questions. The second and last authors were supported by the Research Grants Council of Hong Kong [Grant 15501920]. The third author was supported by the Start-up Grant of Nanyang Technological University. The last author is grateful for the generous support from the Lee Kong Chian School of Business, Singapore Management University under the MPA Research Fellowship.

References

Ahn, H.-S., I. Duenyas, R. Zhang. 2004. Optimal control of a flexible server. *Adv. Appl. Prob.* **36**(1): 139–170.

Andradóttir, S., H. Ayhan. 2005. Throughput maximization for tandem lines with two stations and flexible servers. *Oper. Res.* **53**(3): 516–531.

Andradóttir, S., H. Ayhan, D.G. Down. 2001. Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Manage. Sci.* **47**(10): 1421–1439.

Andradóttir, S., H. Ayhan, D.G. Down. 2003. Dynamic server allocation for queueing networks with flexible servers. *Oper. Res.* **51**(6): 952–968.

Andradóttir, S., H. Ayhan, D.G. Down. 2007a. Compensating for failures with flexible servers. *Oper. Res.* **55**(4): 753–768.

Andradóttir, S., H. Ayhan, D.G. Down. 2007b. Dynamic assignment of dedicated and flexible servers in tandem lines. *Prob. Eng. Inform. Sci.* **21**(4): 497–538.

Armbruster, D., E.S. Gel. 2006. Bucket brigades revisited: Are they always effective? *Eur. J. Oper. Res.* **172**(1) 213–229.

Armbruster, D., E.S. Gel, J. Murakami. 2007. Bucket brigades with worker learning. *Eur. J. Oper. Res.* **176**(1) 264–274.

Armony, M., C.W. Chan, B. Zhu. 2018. Critical care capacity management: Understanding the role of a step down unit. *Production Oper. Management* **27**(5) 859-883.

Bartholdi, J.J. III, L.A. Bunimovich, D.D. Eisenstein. 1999. Dynamics of two- and three-worker "bucket brigade" production lines. *Oper. Res.* **47**(3) 488–491.

Bartholdi, J.J. III, D.D. Eisenstein. 1996a. A production line that balances itself. *Oper. Res.* **44**(1) 21–34.

Bartholdi, J.J. III, D.D. Eisenstein. 1996b. The bucket brigade web page. `https://www2.isye.gatech.edu/~jjb/bucket-brigades.html`, accessed on Oct 15, 2018.

Bartholdi, J.J. III, D.D. Eisenstein. 2005. Using bucket brigades to migrate from craft manufacturing to assembly lines. *Manufacturing Service Oper. Management* **7**(2) 121–129.

Bartholdi, J.J. III, D.D. Eisenstein, R.D. Foley. 2001. Performance of bucket brigades when work is stochastic. *Oper. Res.* **49**(5) 710–719.

Bartholdi, J.J. III, D.D. Eisenstein, Y.F. Lim. 2006. Bucket brigades on in-tree assembly networks. *Eur. J. Oper. Res.* **168**(3) 870–879.

Bartholdi, J.J. III, D.D. Eisenstein, Y.F. Lim. 2009. Deterministic chaos in a model of discrete manufacturing. *Naval Res. Logist.* **56**(4) 293–299.

Bartholdi, J.J. III, D.D. Eisenstein, Y.F. Lim. 2010. Self-organizing logistics systems. *Ann. Rev. in Control* **34**(1) 111–117.

Bartholdi, J.J. III, S.T. Hackman. 2019. *Warehouse and Distribution Science.* `http://www.warehouse-science.com/`. Accessed November 02, 2020.

Bell, S.L., R.J. Williams. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Ann. Appl. Prob.* **11**(3): 608–649.

Bukchin, Y., E. Hanany, E. Khmelnitsky. 2018. Bucket brigade with stochastic worker pace. *IISE Transactions* **50**(12): 1027–1042.

Duenyas, I., D. Gupta, T.L. Olsen. 1998. Control of a single server queueing system with setups. *Oper. Res.* **46**(2): 218–230.

Farrar, T.M. 1993. Optimal use of an extra server in a two station tandem queueing network. *IEEE Trans. Autom. Control* **38**(8): 1296–1299.

Harrison, J.M., M.J. López. 1999. Heavy traffic resource pooling in parallel-server systems. *Queue. Sys.* **33**(4): 339–368.

Hopp, W.J., E. Tekin, M.P. Van Oyen. 2004. Benefits of skill chaining in serial production lines with cross-trained workers. *Manage. Sci.* **50**(1): 83–98.

Hopp, W.J., M.P. Van Oyen. 2004. Agile workforce evaluation: A framework for cross-training and coordination. *IIE Trans.* **36**(10): 919–940.

Hopp, W.J., M.L. Spearman. 2008. Factory Physics: Foundations of Manufacturing Management, 3rd Edition. Waveland Press, Inc.

Iravani, S.M.R., M.J.M. Posner, J.M. Buzacott. 1997. A two-stage tandem queue attended by a moving server with holding and switching costs. *Queue. Syst.* **26**(3-4): 203–228.

Isik, T., S. Andradottir, H. Ayhan. 2016. Optimal control of queueing systems with non-collaborating servers. *Queueing Syst.* **84**(1) 79-110.

Kaufman, D.L., H.-S. Ahn, M.E. Lewis. 2005. On the introduction of an agile, temporary workforce into a tandem queueing system. *Queue. Syst.* **51**(1-2): 135–171.

Kirkizlar, E., S. Andradóttir, H. Ayhan. 2010. Robustness of efficient server assignment policies to service time distributions in finite-buffered lines. *Naval Res. Logist.* **57**(6): 563–582.

Kirkizlar, E., S. Andradottir, H. Ayhan. 2012. Flexible servers in understaffed tandem lines. *Production Oper. Management* **21**(4) 761-777.

Kirkizlar, E., S. Andradottir, H. Ayhan. 2014. Profit maximization in flexible serial queueing networks. *Queueing Syst.* **77**(4) 427-464.

Lim, Y.F. 2011. Cellular bucket brigades. *Oper. Res.* **59**(6) 1539–1545.

Lim, Y.F. 2012. Order-picking by cellular bucket brigades: A case study. R. Manzini, ed. *Warehousing in The Global Supply Chain*. Springer-Verlag, London, 71–85.

Lim, Y.F. 2017. Performance of cellular bucket brigades with hand-off times. *Production Oper. Management* **26**(10), 1915–1923.

Lim, Y.F., Y. Wu. 2014. Cellular bucket brigades on U-lines with discrete work stations. *Production Oper. Management* **23**(7), 1113–1128.

Lim, Y.F., K.K. Yang. 2009. Maximizing throughput of bucket brigades on discrete work stations. *Production Oper. Management* **18**(1) 48–59.

Mazur, J.E., R. Hastie. 1978. Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin* **85**(6) 1256–1274.

Muñoz, L.F., J.R. Villalobos. 2002. Work allocation strategies for serial assembly lines under high labour turnover. *International Journal of Production Research* **40**(8) 1835–1852.

Mandelbaum, A., A.L. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* **52**(6): 836–855.

Rosberg, Z., P.P. Varaiya, J.C. Walrand. 1982. Optimal control of service in tandem queues. *IEEE Trans. Autom. Control* **27**(3): 600–609.

Villalobos, J.R., F. Estrada, L.F. Muñoz, L. Mar. 1999a. Bucket brigade: A new way to boost production. *Twin Plant News* **14**(12) 57–61.

Villalobos, J.R., L.F. Muñoz, L. Mar. 1999b. Assembly line designs that reduce the impact of personnel turnover. *Proc. of IIE Solutions Conference*, Phoenix, AZ.

Webster, S., R.A. Ruben, K.K. Yang. 2012. Impact of storage assignment decisions on a bucket brigade order picking line. *Production Oper. Management* **21**(2) 276–290.

Williams, R. J. 2000. On dynamic scheduling of a parallel server system with complete resource pooling. McDonald, D. R. and Turner, S. R. E., eds. *Analysis of Communication Networks: Call Centers, Traffic and Performance*. American Mathematical Society, Toronto, 49–71.

A Online supplement

A.1 Proof of Theorem 1

The expected makespan for K jobs can be obtained by summing up $E\left[T^{(k)} - T^{(k-1)}\right]$ from $k = 1$ to $k = K$:

$$\begin{aligned} E\left[T^{(K)}\right] &= \sum_{k=1}^{K} E\left[T^{(k)} - T^{(k-1)}\right] \\ &= \sum_{k=1}^{K} \boldsymbol{\pi}^{(k-1)} \boldsymbol{z}^{(k)} \\ &= \sum_{k=1}^{K} \boldsymbol{\pi}^{(0)} \dots \boldsymbol{P}^{(k-1)} \boldsymbol{z}^{(k)}. \end{aligned}$$

So we have:

$$\rho(K) = K \Big/ E\left[T^{(K)}\right] = K \Bigg/ \sum_{k=1}^{K} \boldsymbol{\pi}^{(0)} \boldsymbol{P}^{(1)} \cdots \boldsymbol{P}^{(k-1)} \boldsymbol{z}^{(k)} .$$

□

A.2 Proof of Theorem 2

If the processing rates of the workers are independent of the jobs, then from Lemma 1, $p^{(k)}_{\boldsymbol{h},\boldsymbol{h}'} = p_{\boldsymbol{h},\boldsymbol{h}'}$, for $\boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}$. This implies that $\{\boldsymbol{H}^{(k)}, k = 0, 1, 2, \dots\}$ is an irreducible aperiodic homogeneous Markov chain with a finite number of states. Since there is only a finite number of states and all of them communicate, all the states are positive recurrent. Recall that we set $\boldsymbol{\pi}^{(0)} = (1, 0, ..., 0)$. According to the definition of $\pi^{(k)}_n$, we have

$$\begin{aligned} \lim_{k\to\infty} \pi^{(k)}_n &= \lim_{k\to\infty} Pr\left\{\boldsymbol{H}^{(k)} = \boldsymbol{h}^n\right\} \\ &= \lim_{k\to\infty} \sum_{m=1}^{|\mathcal{H}|} \pi^{(0)}_n Pr\left\{\boldsymbol{H}^{(k)} = \boldsymbol{h}^n | \boldsymbol{H}^{(0)} = \boldsymbol{h}^m\right\} \\ &= \lim_{k\to\infty} Pr\left\{\boldsymbol{H}^{(k)} = \boldsymbol{h}^n | \boldsymbol{H}^{(0)} = \boldsymbol{h}^1\right\} \\ &= \pi_n. \end{aligned}$$

The last equality is due to Theorem 4.3.3 of Ross (1996), which also guarantees that $\boldsymbol{\pi} = \left(\pi_1, \dots, \pi_{|\mathcal{H}|}\right)$ is the only stationary distribution of the Markov chain. This stationary distribution can be obtained by solving the equations $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\boldsymbol{e} = 1$.

Thus, we have

$$\begin{aligned}\lim_{K\to\infty}\frac{1}{\rho(K)} &= \lim_{K\to\infty}\frac{E\left[T^{(K)}\right]}{K}\\ &= \lim_{K\to\infty}\frac{\sum_{k=1}^{K}E\left[T^{(k)}-T^{(k-1)}\right]}{K}\\ &= \lim_{K\to\infty}\frac{\sum_{k=1}^{K}\boldsymbol{\pi}^{(k-1)}\boldsymbol{z}}{K}\\ &= \boldsymbol{\pi}\boldsymbol{z}.\end{aligned}$$

The last equality holds because $\boldsymbol{\pi}^{(k-1)}\boldsymbol{z} \to \boldsymbol{\pi}\boldsymbol{z}$ as k goes to infinity. As a result, the Cesaro mean $\frac{\sum_{k=1}^{K}\boldsymbol{\pi}^{(k-1)}\boldsymbol{z}}{K}$ also converges to $\boldsymbol{\pi}\boldsymbol{z}$. (The proof of the convergence of the Cesaro mean can be found in Demidovich (1989).) Therefore, $\lim_{K\to\infty}\rho(K) = 1/(\boldsymbol{\pi}\boldsymbol{z})$. □

A.3 Proof of Lemma 2

We denote the cumulative distribution function (CDF) of $Y(k)$ as $F^{(k)}(\cdot)$. Let $Y_n(k)$ denote the inter-completion time between the $(k-1)$st and the kth resets, conditioned on $\boldsymbol{H}^{(k-1)} = \boldsymbol{h}^n$. Let $F_n^{(k)}(\cdot)$ denote the CDF of $Y_n(k)$. Recall that $\boldsymbol{\pi}^{(k-1)} = \left(\pi_1^{(k-1)},\ldots,\pi_{|\mathcal{H}|}^{(k-1)}\right)$ is the probability distribution of the $(k-1)$st hand-off station vector $\boldsymbol{H}^{(k-1)}$. The CDF of $Y(k)$ can be expressed as a mixture distribution $F^{(k)}(\cdot) = \sum_{n=1}^{|\mathcal{H}|}\pi_n^{(k-1)}F_n^{(k)}(\cdot)$. According to Proposition 5.2 of Ross (2010), the variance of $Y(k)$ can be obtained as follows:

$$\mathrm{Var}\left(Y(k)\right) = \sum_{n=1}^{|\mathcal{H}|}\pi_n^{(k-1)}\mathrm{Var}\left(Y_n(k)\right) + \sum_{n=1}^{|\mathcal{H}|}\pi_n^{(k-1)}\left(E\left[Y_n(k)\right]\right)^2 - \left(\sum_{n=1}^{|\mathcal{H}|}\pi_n^{(k-1)}E\left[Y_n(k)\right]\right)^2. \quad (11)$$

Note that $Y_n(k)$ is the sum of the processing times of worker I on job k from station h_{I-1}^n to station J. Since these processing times are independent exponential random variables, we have

$$E\left[Y_n(k)\right] = \sum_{j=h_{I-1}^n}^{J}E\left[Z_{I,j}^{(k)}\right] = \sum_{j=h_{I-1}^n}^{J}1/\mu_{I,j}^{(k)}, \quad (12)$$

and

$$Var\left(Y_n(k)\right) = \sum_{j=h_{I-1}^n}^{J}Var\left(Z_{I,j}^{(k)}\right) = \sum_{j=h_{I-1}^n}^{J}\left(1/\mu_{I,j}^{(k)}\right)^2. \quad (13)$$

Substituting Equations (12) and (13) into Equation (11), we obtain Lemma 2. □

A.4 Proof of Lemma 3

We assume $v_1 + v_2 = c$, where c is a constant. We have

$$\begin{aligned}
1/\rho_\infty \;=\; & \big(v_1^7 s_2^2 s_3^5 + 4v_1^6 v_2 s_1 s_2^2 s_3^4 + 2v_1^6 v_2 s_1 s_2 s_3^5 + 2v_1^6 v_2 s_2^3 s_3^4 + 2v_1^4 v_2^3 s_1^3 s_2^2 s_3^2 + \\
& 3v_1^4 v_2^3 s_1^2 s_2^3 s_3^2 + 3v_1^4 v_2^4 s_1^2 s_2^2 s_3^3 + 6v_1^4 v_2^2 s_1^2 s_2 s_3^4 + v_1^4 v_2^2 s_1^2 s_3^5 + \\
& 2v_1^4 v_2^2 s_1 s_2^4 s_3^2 + 5v_1^4 v_2^2 s_1 s_2^3 s_3^3 + 4v_1^4 v_2^2 s_1 s_2^2 s_3^4 + v_1^4 v_2^2 s_2^5 s_3^2 + \\
& 2v_1^3 v_2^3 s_1^3 s_3^4 + 2v_1^3 v_2^3 s_1^2 s_2^4 s_3 + 6v_1^3 v_2^3 s_1^2 s_2^3 s_3^2 + 8v_1^3 v_2^3 s_1^2 s_2^2 s_3^3 + \\
& 2v_1^3 v_2^3 s_1^2 s_2 s_3^4 + 2v_1^3 v_2^3 s_1 s_2^5 s_3 + 2v_1^3 v_2^3 s_1 s_2^4 s_3^2 + v_1^2 v_2^4 s_1^5 s_3^2 + \\
& 2v_1^2 v_2^4 s_1^4 s_2^2 s_3 + 3v_1^2 v_2^4 s_1^3 s_2^3 s_3 + 3v_1^2 v_2^4 s_1^3 s_2^2 s_3^2 + 3v_1^2 v_2^4 s_1^3 s_2 s_3^3 + \\
& v_1^2 v_2^4 s_1^2 s_2^5 + 4v_1^2 v_2^4 s_1^2 s_2^4 s_3 + 2v_1 v_2^5 s_1^5 s_2 s_3 + 2v_1 v_2^5 s_1^3 s_2^4 + v_2^6 s_1^5 s_2^2\big) / \\
& [7v_2\left(v_1^3 s_2 s_3^2 + 2v_1^2 v_2 s_1 s_2 s_3 + v_1^2 v_2 s_1 s_3^2 + v_1^2 v_2 s_2^2 s_3 + v_1 v_2^2 s_1^2 s_3 + v_1 v_2^2 s_1 s_2^2 + \right. \\
& \left. v_1 v_2^2 s_1 s_2 s_3 + v_2^3 s_1^2 s_2\right)].
\end{aligned}$$

Taking the derivative of $1/\rho_\infty$ with respect to v_1, we have

$$\begin{aligned}
\frac{d(1/\rho_\infty)}{dv_1} \;=\; & \big(v_1^6 s_2^2 s_3^5 + 4v_1^5 v_2 s_1 s_2^2 s_3^4 + 2v_1^5 v_2 s_1 s_2 s_3^5 + 2v_1^5 v_2 s_2^3 s_3^4 + 2v_1^4 v_2^2 s_1^3 s_2^2 s_3^2 + \\
& 3v_1^4 v_2^2 s_1^2 s_2^3 s_3^2 + 3v_1^4 v_2^2 s_1^2 s_2^2 s_3^3 + 6v_1^4 v_2^2 s_1^2 s_2 s_3^4 + v_1^4 v_2^2 s_1^2 s_3^5 + \\
& 2v_1^4 v_2^2 s_1 s_2^4 s_3^2 + 5v_1^4 v_2^2 s_1 s_2^3 s_3^3 + 4v_1^4 v_2^2 s_1 s_2^2 s_3^4 + v_1^4 v_2^2 s_2^5 s_3^2 + \\
& 2v_1^3 v_2^3 s_1^4 s_2 s_3^2 + 2v_1^3 v_2^3 s_1^3 s_2^3 s_3 + 2v_1^3 v_2^3 s_1^3 s_2^2 s_3^2 + 2v_1^3 v_2^3 s_1^3 s_2 s_3^3 + \\
& 2v_1^3 v_2^3 s_1^3 s_3^4 + 2v_1^3 v_2^3 s_1^2 s_2^4 s_3 + 6v_1^3 v_2^3 s_1^2 s_2^3 s_3^2 + 8v_1^3 v_2^3 s_1^2 s_2^2 s_3^3 + \\
& 2v_1^3 v_2^3 s_1^2 s_2 s_3^4 + 2v_1^3 v_2^3 s_1 s_2^5 s_3 + 2v_1^3 v_2^3 s_1 s_2^4 s_3^2 + v_1^2 v_2^4 s_1^5 s_3^2 + \\
& 2v_1^2 v_2^4 s_1^4 s_2^2 s_3 + 3v_1^2 v_2^4 s_1^3 s_2^3 s_3 + 3v_1^2 v_2^4 s_1^3 s_2^2 s_3^2 + 3v_1^2 v_2^4 s_1^3 s_2 s_3^3 + \\
& v_1^2 v_2^4 s_1^2 s_2^5 + 4v_1^2 v_2^4 s_1^2 s_2^4 s_3 + 2v_1 v_2^5 s_1^5 s_2 s_3 + 2v_1 v_2^5 s_1^3 s_2^4 + v_2^6 s_1^5 s_2^2\big) / \\
& [v_2^2\left(v_1^3 s_2 s_3^2 + 2v_1^2 v_2 s_1 s_2 s_3 + v_1^2 v_2 s_1 s_3^2 + v_1^2 v_2 s_2^2 s_3 + v_1 v_2^2 s_1^2 s_3 + v_1 v_2^2 s_1 s_2^2 + \right. \\
& \left. v_1 v_2^2 s_1 s_2 s_3 + v_2^3 s_1^2 s_2\right)^2].
\end{aligned}$$

Since $d(1/\rho_\infty)/dv_1 > 0$, as v_2 increases (v_1 decreases), $1/\rho_\infty$ decreases, which means ρ_∞ increases. □

A.5 Proof of Lemma 4

Recall that $f_{i,j}^{(k)}$ denote the probability of worker i finishing his job at station j between the $(k-1)$st and kth resets. For convenience, we set $H_0^{(k)} = 1$, and $H_I^{(k)} = J+1$, for all k. We have

$$\begin{aligned} f_{i,j}^{(k)} &= Pr\left\{H_{i-1}^{(k-1)} \leq j, H_i^{(k)} \geq j+1\right\} \\ &= Pr\left\{H_i^{(k)} \geq j+1\right\} - Pr\left\{H_{i-1}^{(k-1)} \geq j+1, H_i^{(k)} \geq j+1\right\} \\ &= Pr\left\{H_i^{(k)} \geq j+1\right\} - Pr\left\{H_{i-1}^{(k-1)} \geq j+1\right\} \\ &= \sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_i^n \geq j+1}} \pi_n^{(k)} - \sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_{i-1}^n \geq j+1}} \pi_n^{(k-1)}. \end{aligned} \tag{14}$$

Note that the third equality holds because $H_{i-1}^{(k-1)} \geq j+1$ implies $H_i^{(k)} \geq j+1$. Equation (8) in Lemma 4 is obtained by setting k to infinity in Equation (14). □

A.6 Proof of Lemma 5

Recall that $Z_{i,j}^{(k)}$ denotes the time duration for worker i to process job k at station j, which follows an exponential distribution with rate $\mu_{i,j}^{(k)} = \left(s_j / v_{i,j}^{(k)}\right)^{-1} = (s_j/v_{i,j})^{-1} = \mu_{i,j}$. Let $F(\cdot)$ denote the CDF of $Z_{i,j}^{(k)}$. Define $D_{i,j}^{(k)}$ as the time duration from the time point when worker i starts working at station j on job $k+I-i$ after the $(k-1)$st reset to the time point $T^{(k)}$. Between the $(k-1)$st and kth resets, we set $D_{i,j}^{(k)} = 0$ if worker i does not work at station j on job $k+I-i$. Let $G^{(k)}(\cdot)$ denote the CDF of $D_{i,j}^{(k)}$. Between the $(k-1)$st and kth resets, the time duration that worker i spends working at station j on job $k+I-i$ is $\min\left(Z_{i,j}^{(k+I-i)}, D_{i,j}^{(k)}\right)$. Condition on $D_{i,j}^{(k)} = t$, we have

$$\begin{aligned} E\left[\min\left(Z_{i,j}^{(k+I-i)}, D_{i,j}^{(k)}\right) | D_{i,j}^{(k)} = t\right] &= \int_0^t x\mu_{i,j}e^{-\mu_{i,j}x}dx + t\int_t^\infty \mu_{i,j}e^{-\mu_{i,j}x}dx \\ &= -te^{-\mu_{i,j}t} + \int_0^t e^{-\mu_{i,j}x}dx + te^{-\mu_{i,j}t} \\ &= \frac{1}{\mu_{i,j}}\int_0^t \mu_{i,j}e^{-\mu_{i,j}x}dx \\ &= \frac{1}{\mu_{i,j}}F(t). \end{aligned}$$

So, we have

$$\begin{aligned}
E\left[\min\left(Z_{i,j}^{(k+I-i)}, D_{i,j}^{(k)}\right)\right] =& E_{D_{i,j}^{(k)}}\left[E\left[\min\left(Z_{i,j}^{(k+I-i)}, D_{i,j}^{(k)}\right)|D_{i,j}^{(k)}\right]\right] \\
=& E_{D_{i,j}^{(k)}}\left[F\left(D_{i,j}^{(k)}\right)\right] \\
=& \frac{1}{\mu_{i,j}}\int_0^\infty F\left(D_{i,j}^{(k)}\right)\ dG^{(k)}\left(D_{i,j}^{(k)}\right) \\
=& f_{i,j}^{(k)}/\mu_{i,j},
\end{aligned}$$

where $f_{i,j}^{(k)}$ denotes the probability of worker i finishing his job at station j between the $(k-1)$st and kth resets. The expected inter-completion time $E[Y(k)]$ between the $(k-1)$st and kth resets equals the sum of $\sum_{j=1}^{J} f_{i,j}^{(k)}/\mu_{i,j}$ and the expected blocked time of worker i between the $(k-1)$st and kth resets. Setting k to infinity, we obtain Equation (9). □

A.7 Proof of Theorem 4

We can obtain the following equation

$$\begin{aligned}
\sum_{i=1}^{I}\nu_i\left(Y_\infty - B_i\right) =& \sum_{i=1}^{I}\left(\frac{\sum_{j=1}^{J} f_{i,j}s_j}{\sum_{j=1}^{J} f_{i,j}s_j/v_{i,j}}\left(Y_\infty - B_i\right)\right) \\
=& \sum_{i=1}^{I}\left(\frac{\sum_{j=1}^{J} f_{i,j}s_j}{\sum_{j=1}^{J} f_{i,j}/\mu_{i,j}}\left(Y_\infty - B_i\right)\right) \\
=& \sum_{i=1}^{I}\left(\sum_{j=1}^{J} f_{i,j}s_j\right) \\
=& \sum_{j=1}^{J}\left(s_j\sum_{i=1}^{I} f_{i,j}\right) \\
=& \sum_{j=1}^{J}\left(s_j\sum_{i=1}^{I}\left(\sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_i^n \geq j+1}}\pi_n - \sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_{i-1}^n \geq j+1}}\pi_n\right)\right) \\
=& \sum_{j=1}^{J}\left(s_j\left(\sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_I^n \geq j+1}}\pi_n - \sum_{\substack{n=1,\ldots,|\mathcal{H}| \\ h_0^n \geq j+1}}\pi_n\right)\right) \\
=& \sum_{j=1}^{J}\left(s_j \cdot 1\right) \\
=& 1.
\end{aligned} \tag{15}$$

The third equality is due to Equation (9). The fifth equality is due to Equation (8). Substituting $1 = \rho_\infty Y_\infty$ into Equation (15), we have Theorem 4. □

References

Demidovich, B. 1989. *Problems in Mathematical Analysis*. Mir Publishers, Moscow.

Ross, S.M. 1996. *Stochastic Processes, Second Edition*. John Wiley & Sons, Inc., New York.

Ross, S.M. 2010. *A First Course in Probability, Eighth Edition*. Pearson Education, Inc., Upper Saddle River.