PERFORMANCE OF BUCKET BRIGADES WHEN WORK IS STOCHASTIC

John J. Bartholdi, III
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0205 USA
john.bartholdi@isye.gatech.edu

Donald D. Eisenstein
Graduate School of Business
The University of Chicago
Chicago, Illinois 60637 USA
don.eisenstein@gsb.uchicago.edu

Robert D. Foley
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0205 USA
rfoley@isye.gatech.edu

February 13, 1998; revised March 22, 2000

Abstract

"Bucket brigades" are a way of sharing work on a flow line that results in the spontaneous emergence of balance and consequent high throughput. All this happens without a work-content model or traditional assembly-line balancing technology. Here we show that bucket brigades can be effective even in the presence of variability in the work content. In addition, we report confirmation at the national distribution center of a major chain retailer, which experienced a 34% increase in productivity after the workers began picking orders by bucket brigade.

Key words: *flow line, assembly line, work-sharing, bucket brigade, self-organizing systems, dynamical systems*

"Bucket brigades" are a way of coördinating workers who progressively assemble a product along a flow line. Each worker follows this simple rule: "Carry work forward from station to station until someone takes over your work; then go back for more". When the last worker completes a product, he walks back upstream and takes over the work of his predecessor, who then walks back and takes over the work of his predecessor, and so on, until the first worker begins a new product at the start of the line. No unattended work-in-process (WIP) is allowed in the system.

Workers are not restricted to any subset of stations; rather each is to carry his work as far toward completion as possible, except that workers may not pass one another. This means that, at least in principle, a worker might catch up to his successor and be blocked; the bucket brigade rule requires that the blocked worker remain idle until the station is available. (As we shall see, the art of implementing a successful bucket brigade is to make such events unlikely.)

The final requirement of bucket brigades is that the workers be sequenced from slowest to fastest along the direction of material flow. When these requirements are met, work is paced by the fastest worker, who triggers each successive series of walk-backs. The result is a pure pull system.

Bucket brigades are distinguished from similar work-sharing protocols, such as the Toyota Sewing System (TSS), by insisting on the total abolishment of any *a priori* work assignment or zones that might restrict the movement of the workers; and by requiring that the workers be sequenced from slowest to fastest along the direction of material flow.

The distinctive and valuable feature of bucket brigades is that they are self-balancing; that is, a balanced partition of the work will emerge spontaneously, which reduces the need for traditional industrial engineering technologies of time-motion studies, work-content models, and assembly-line balancing. Moreover, under quite general conditions the emergent balance results in the maximum possible rate of production. Finally, the simplicity of bucket brigades makes them easy to implement and so to realize these benefits.

Bartholdi and Eisenstein (1996) analyzed the performance of bucket brigades performing high-volume assembly of a mature product, for which a deterministic model of work content was appropriate. Here we extend this analysis to a stochastic model of work content and show that the dynamics and production rate will be similar to those of the deterministic model when there is "sufficient work" distributed among "sufficiently many" work stations. We also report confirmation of the practical value of this at the national distribution center of a major chain retailer, where the products are customer orders, which are "assembled" by order-pickers. Because customer orders vary in hard-to-predict ways, their work content may be imagined to be stochastic. After converting to bucket brigades, the order-pickers realized a 34% increase in productivity, and similar successes have subsequently been achieved in other distribution centers.

1 Bucket Brigades

The simplest model of the dynamics of bucket brigades is based on the following assumptions.

Assumption 1 (Insignificant Walking Time). *The total time to assemble a product is significantly greater than the time to walk the length of the flow line. Therefore all hand-offs occur, for all practical purposes, simultaneously, synchronized by item-completions of the last worker.*

Assumption 2 (Total Ordering of Workers by Velocity). *Each worker $i = 1, \ldots, n$ is characterized by a distinct, constant work velocity v_i.*

Assumption 3 (Smoothness and Predictability of Work). *The nominal work content of the product is a constant (which we normalize to 1); and the work content is spread continuously and uniformly along the flow line.*

We call this the "Normative Model" because it represents ideal conditions sufficient to guarantee that bucket brigades achieve the maximum possible throughput. For the Normative Model a variation of a result from Bartholdi and Eisenstein (1996) applies: Because of Assumption 1, Smoothness and Predictability of Work, we can model the work content as the unit interval $[0,1]$. Consider the moment at which the k-th item is completed and worker i takes over the item being assembled by worker $i-1$. Let $x_i^{(k)}$ represent the fraction of work completed for that item at that moment.

Theorem 1 (Self-balancing).

$$\lim_{k\to\infty} x_i^{(k)} = \frac{\sum_{j=1}^{i-1} v_j}{\sum_{j=1}^{n} v_j} \quad \text{for } 1 < i \leq n.$$

This means that when workers are sequenced from slowest to fastest worker i comes to repeatedly execute the interval of work content

$$\left[\frac{\sum_{j=1}^{i-1} v_j}{\sum_{j=1}^{n} v_j}, \frac{\sum_{j=1}^{i} v_j}{\sum_{j=1}^{n} v_j}\right]$$

and the production rate of the flow line increases to $\sum_{j=1}^{n} v_j$, the largest possible.

Proof. At the moment of handoff coinciding with completion of the k-th item, the clock time separating workers i and $i+1$ is

$$t_i^{(k)} = \frac{x_{i+1}^{(k)} - x_i^{(k)}}{v_i};$$

and the next item will be completed after time

$$t_n^{(k)} = \frac{1 - x_n^{(k)}}{v_n}.$$

After completion of the $(k+1)$-st item the clock-time separating adjacent workers becomes

$$\begin{aligned} t_i^{(k+1)} &= \frac{x_{i+1}^{(k+1)} - x_i^{(k+1)}}{v_i} \\ &= \frac{\left(x_i^{(k)} + v_i t_n^{(k)}\right) - \left(x_{i-1}^{(k)} + v_{i-1} t_n^{(k)}\right)}{v_i} \\ &= \left(\frac{v_{i-1}}{v_i}\right) t_{i-1}^{(k)} + \left(1 - \frac{v_{i-1}}{v_i}\right) t_n^{(k)}. \end{aligned}$$

The workers are sequenced from slowest to fastest, so we may interpret these equations as describing a finite state Markov Chain with transition matrix

$$A = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ v_1/v_2 & 0 & \dots & 0 & 1 - v_1/v_2 \\ 0 & v_2/v_3 & \dots & 0 & 1 - v_2/v_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & v_{n-1}/v_n & 1 - v_{n-1}/v_n \end{bmatrix}$$

with $t^{(k+1)} = At^{(k)} = A^{k+1}t^{(0)}$. This Markov Chain is irreducible and, because $v_{i-1} < v_i$ for all i, aperiodic. Therefore, by basic results about Markov chains, A^k converges to a matrix, each row of which is

$$\left(\frac{v_1}{\sum_j v_j}, \frac{v_2}{\sum_j v_j}, \dots, \frac{v_n}{\sum_j v_j}\right).$$

The convergence of the $t_i^{(k)}$ and $x_i^{(k)}$ and the specific claims follow by simple algebra. □

Assumption 1, Insignificant Walking Time, seems uncontroversial. (We use its extreme form, instantaneous walk-backs, for convenience.) Assumption 2, Total Ordering of Workers by Velocity, holds for unskilled work or whenever workers have similar training (Bartholdi and Eisenstein, 1996). Assumption 3, Smoothness and Predictability of Work, tends to hold for mature technologies because management and engineering continually strive to remove variation from work and to eliminate bottlenecks. However, in some important economic contexts, such as order-picking in a warehouse, this last assumption is unreliable. Therefore, the object of this paper is to explore the behavior of bucket brigade lines when this assumption is modified to allow randomness in the amount and location of work.

Under the Normative Model bucket brigades achieve the maximum possible throughput; furthermore, in real life bucket brigades have performed with remarkable efficiency in a range of commercial applications, many of which are described on our web page at `www.isye.gatech.edu/faculty/John_Bartholdi/bucket-brigades`. Why is it that bucket brigades perform well even when the strong assumptions about the nature of work content do not hold? Here we prove that the conclusions of Theorem 1 continue to apply in a useful sense even when there is randomness in the work content.

2 A stochastic model of work

Consider the behavior of bucket brigades in which Assumption 3 (Smoothness and Predictability of Work) is replaced by the following stochastic model.

Assumption 3′ (iid, exponentially distributed work.) *Let the work to assemble a product consist of m discrete task primitives at m successive work stations. The nominal work-content at each station is independent and follows an exponential distribution with common mean normalized to* 1.

This means that the time required for the i-th worker to complete a task follows an exponential distribution with mean $1/v_i$. We will prove that, as the number of stations increases, the moment-to-moment behavior of the stochastic line will increasingly resemble that of the Normative Model. Moreover, this resemblance will assert itself with great uniformity.

Increasing the number of stations may also be taken to model the partition of tasks into subtasks.

One may interpret our conclusion in the following way: Imagine a video of workers operating according to the Normative Model and another video of the same workers operating under the Stochastic Model (with work rescaled to be comparable). The two copies of the workers begin at the same starting positions relative to the total (expected) work content. Then our claim is that the two videos will become indistinguishable as the number of stations increases in the stochastic model; and so all measurements of the two lines, including the instants at which each successive item is completed, become similar. Therefore the stochastic line ever more resembles the deterministic line, which Theorem 1 has shown to achieve the maximum production rate. Furthermore, as we show by both simulation and by case study, this similarity asserts itself for few enough stations to be of practical benefit.

Our analysis is conservative in assuming work that is exponentially distributed. This means that there will be greater variance at each work station than one would expect to find in practice. (It is hard to imagine an economically viable production process in which a partially-completed task had no memory of the work invested in it!) This unrealistically large variance reduces the throughput of bucket brigades because it increases the chances of blocking.

To compare the stochastic and deterministic models, we will build and analyze a more detailed model of the deterministic system. Where Theorem 1 considered a series of "snapshots" of the workers taken immediately after walkback, our new model, the *fluid model*, is more like a video in that it captures not just the system state after walkbacks, but the dynamics of the bucket brigade in continuous time.

3 The fluid model

Here we model the evolution of the bucket brigade in continuous time. Assume there are n workers located on the unit interval $[0,1)$. Worker i moves to the right with speed $0 < v_i < \infty$ unless blocked by worker $i+1$. When worker n reaches the end of the unit interval, a part is completed, and the workers instantaneously reset (walk back to get more work). Let $\bar{X} : [0,\infty) \to E$ be the function which gives the location of the n workers at any arbitrary time $t \geq 0$. In particular, $\bar{X}_i(t)$ denotes the location of worker i at time t. Note that at reset times $\bar{X}$ is not well-defined because the workers instantaneously move from one location to another. To avoid this ambiguity, we select $\bar{X}$ so that it is right-continuous at all $t \in [0,\infty)$ and has limits from the left at all $t \in (0,\infty)$. Thus, none of the workers are ever at one, and $\bar{X}(t) \in E$ where $E \equiv \{(\ell_1,\ldots,\ell_n)|0 \leq \ell_1 \leq \ell_2 \leq \cdots \leq \ell_n < 1\}$. Let E^- denote the closure of E; thus, $E^- \equiv \{(\ell_1,\ldots,\ell_n)|0 \leq \ell_1 \leq \ell_2 \leq \cdots \leq \ell_n \leq 1\}$. Define $\bar{R} : E^- \setminus E \to E$ to be the reset function such that $\bar{R}(x(t-)) = x(t)$ where $x(t-) = \lim_{s\uparrow t} x(s)$. In particular, if $x(t-) \in E^- \setminus E$, then $x(t-)$ is of the form $(x_1(t-),\ldots,x_i(t-),1,\ldots,1)$ for some $i < n$ with $x_i(t-) < 1$. In this case, $\bar{R}(x(t-)) = (0,\ldots,0,x_1(t-),\ldots,x_i(t-))$ and $m-i$ parts were finished at time t.

At this point, it will be convenient to define several classes of functions. Let $D_E[0,\infty)$ be the space of all E-valued functions on $[0,\infty)$ that are right-continuous with limits from the left (RCLL). Thus,

$\bar{X} \in D_E[0,\infty)$. Similarly, we let $D_{\mathbb{R}^n_+}[0,\infty)$ denote the space of RCLL $\mathbb{R}^n_+$-valued functions, and $C_{\mathbb{R}^n_+}[0,\infty)$ the space of continuous, $\mathbb{R}^n_+$-valued functions on $[0,\infty)$.

The following result gives the equations describing the fluid model of the system, where $\bar{X}_i(t)$ represents the location of the worker i at time t, $\bar{T}_i(t)$ represents the amount of time that worker i was productive during the interval $[0,t]$, $\bar{I}_i(t)$ the amount of time worker i was idle (blocked by worker $i+1$) in $[0,t]$, and $\bar{S}_i(t)$ represents the total distance walked back (instantaneously) by worker i during the interval $[0,t]$.

Lemma 1. *Given a starting position $x \in E$ and vector of speeds v, there exists a unique triple $(\bar{X},\bar{T},\bar{S})$ where $\bar{X} \in D_E[0,\infty)$, $\bar{T} \in C_{\mathbb{R}^n_+}[0,\infty)$, and $\bar{S}$ is a pure jump process in $D_{\mathbb{R}^n_+}[0,\infty)$ that satisfy for $i = 1,\ldots,n$ and $t > 0$:*

$$\bar{X}_i(t) = x_i + v_i\bar{T}_i(t) - \bar{S}_i(t) \tag{1}$$

$$\bar{X}(t) \in E \tag{2}$$

$$\bar{T}_i(0) = 0 \text{ and } \bar{T}_i(t) \text{ is non-decreasing in } t \tag{3}$$

$$\bar{I}_i(t) = t - \bar{T}_i(t) \tag{4}$$

$$\int_0^\infty \mathbf{1}\left(\bar{X}_{i+1}(t) > \bar{X}_i(t)\right) d\bar{I}_i(t) = 0 \tag{5}$$

$$\bar{S}_i(t) = \int_0^t \mathbf{1}\left(\bar{X}_n(s-) = 1\right)\left(\bar{X}_i(s-) - \bar{R}_i(\bar{X}(s-)\right) d\bar{N}(s) \tag{6}$$

where $\mathbf{1}(A)$ is the indicator function of A, $\bar{N}$ is counting measure, and $\bar{X}_{n+1}(t) \equiv 1$ so $\bar{I}_n(t) = 0$.

Proof. See Appendix. ∎

4 Stochastic dynamics

Assume m stations labeled 0 through $m-1$, with the nominal length of time to process a job at station j being exponentially distributed with mean one. Worker i works at velocity v_i; so the time for worker i at each station is exponentially distributed with rate v_i. For each of the n workers let $X_i^m(t)$ denote the location of worker i at time t, and $X^m(t)$ be the column vector of these locations. We explicitly carry the number of stations m as part of the notation since in the next section we allow the number of stations to vary. One difference between our stochastic and deterministic models is that in the stochastic model a slower worker may catch up to a faster worker and be at the same station. If there are several workers at a station, only the highest numbered among them will be allowed to work, while the others must remain idle.

Because the durations of work are independent, identically distributed exponential random variables, the number of movements of the ith worker behaves like a Poisson process with rate v_i unless the worker is blocked. Let $N(t)$ be a vector of independent Poisson processes with rates $(v_1, \ldots, v_n)$. Let T^m be the vector of the amounts of time that each worker is productive during the interval $[0, t]$; let $I_i^m(t)$ be the amounts of time that worker i is idle (blocked) during $[0, t]$, and let $S_i^m(t)$ be the total distance that worker i has walked during resets. Then the following set of equations, which hold for $i = 1, \ldots, n$ and $t > 0$, uniquely define these processes:

$$X_i^m(t) = X_i^m(0) + N_i(T_i^m(t)) - S_i^m(t) \tag{7}$$

$$X^m(t) \in E^m \tag{8}$$

$$T_i^m(0) = 0 \text{ and } T_i^m(t) \text{ is non-decreasing} \tag{9}$$

$$I_i^m(t) = t - T_i^m(t) \tag{10}$$

$$\int_0^\infty \mathbf{1}\left(X_{i+1}^m(t) > X_i^m(t)\right) dI_i^m(t) = 0 \tag{11}$$

$$S_i^m(t) = \int_0^t \mathbf{1}\left(X_n^m(s-) = m-1\right) D_i^m(X^m(s-)) dN_n(s) \tag{12}$$

where $E^m = \{i \in \mathbb{Z}_+^n | 0 \le i_1 \le i_2 \cdots i_n \le m-1\}$, $X_{n+1}(t) \equiv m$,

$$D^m(X(s-)) = X(s-) - R^m(X(s-), \tag{13}$$

$$R^m(X(s-)) = (0, X_1(s-), \ldots, X_{n-1}(s-)), \tag{14}$$

and the stochastic integral in Expression (11) is a sample path integral; cf. Chapter 6 of Wong and Hajek (1985).

5 Convergence of the stochastic model to the fluid model

We would like to compare the behavior of the stochastic model and the continuous deterministic model, but their state spaces are quite different: $\{0, \ldots, m-1\}$ vs. $[0, 1)$ and their time scales are quite different: one worker working with velocity 1 takes 1 unit of time to produce an item in the deterministic model, but the expected time in the stochastic model is m units of time. However, we can directly compare the two in a reasonable way by rescaling the stochastic model, in effect speeding up time while reducing resolution by a factor of m. Define $\tilde{X}^m(t) \equiv X^m(mt)/m$, and $\tilde{X}^m \equiv \{\tilde{X}^m(t), t \ge 0\}$. Then rewriting Expressions (7)–(12)

under this rescaling, we obtain $\tilde{S}_i^m(t) = S_i^m(mt)/m$, $\tilde{I}_i^m(t) = I_i^m(mt)/m$, $\tilde{T}_i^m(t) = T_i^m(mt)/m$ and

$$\tilde{X}_i^m(t) = \tilde{X}^m(0) + N_i(\tilde{T}_i^m(mt))/m - \tilde{S}_i^m(t) \tag{15}$$

$$\tilde{X}^m(t) \in E \tag{16}$$

$$\tilde{T}_i^m(0) = 0 \text{ and } \tilde{T}_i^m(t) \text{ is non-decreasing} \tag{17}$$

$$\tilde{I}_i^m(t) = t - \tilde{T}_i^m(t) \tag{18}$$

$$\int_0^\infty \mathbf{1}\left(\tilde{X}_{i+1}^m(t) > \tilde{X}_i^m(t)\right) d\tilde{I}_i(t) = 0 \tag{19}$$

$$\tilde{S}_i^m(t) = \int_0^t \mathbf{1}\left(\tilde{X}_n^m(s-) = (m-1)/m\right) D_i^m(X^m(s-))dN_n(ms), \tag{20}$$

where $D^m(\tilde{X}^m(s-)) = \tilde{X}^m(s-) - R^m(\tilde{X}^m(s-))$.

We show that the rescaled stochastic model converges in a certain sense to the deterministic fluid model as the number of stations m increases. The following limits hold almost surely as $m \to \infty$. We use $X \overset{u.o.c.}{\longrightarrow} Y$ or $X(t) \overset{u.o.c.}{\longrightarrow} Y(t)$ to denote uniform convergence over compact sets (u.o.c.); that is, $X(t) \to Y(t)$ uniformly for t restricted to compact sets. It is well known that $N(mt)/m \overset{u.o.c.}{\longrightarrow} vt$. Unfortunately our rescaled model $\tilde{X}^m$ does not converge u.o.c. to $\bar{X}$ because the two processes may reset at slightly different times. Consequently we must resort to a weaker metric that considers the two processes to be close if they jump approximately the same distance at approximately the same time. Let $X \overset{J1}{\longrightarrow} Y$ denote convergence in the Skorohod $J1$ topology (Skorohod, 1956).

Theorem 2. *If $0 < v_1 \leq v_2 \leq \cdots \leq v_n < \infty$ and the rescaled starting positions of the workers $\tilde{X}^m(0) \to x = (x_1, \ldots, x_n)$ with $0 \leq x_1 < x_2 < \cdots < x_n < 1$ then $\tilde{X}^m \overset{J1}{\longrightarrow} \bar{X}$ and $\tilde{T}^m(t) \overset{u.o.c.}{\longrightarrow} \bar{T}(t) \equiv et$ where e is the n-dimensional vector of one's.*

Proof. See Appendix. □

6 The effectiveness of bucket brigades in practice

6.1 Order-picking in a distribution warehouse

In the stores of chain retailers, space for inventory is expensive, so the distribution centers (DC's) supporting them replenish stock-keeping units (sku's) frequently and in small, less-than-caseload amounts. This means that a typical store orders many sku's, but small numbers of each, so that picking these orders is labor intensive. Often a DC employs hundreds of order-pickers.

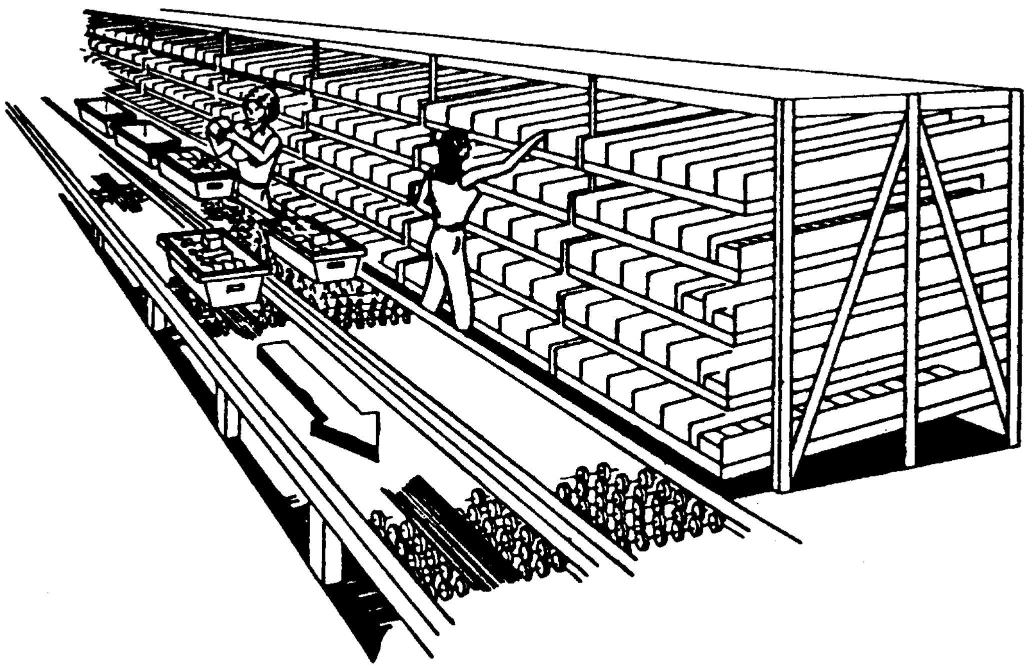

Figure 1: A team picking from an aisle of flow rack to a conveyor (from "Warehouse Modernization and Layout Planning Guide", Department of the Navy, Naval Supply Systems Command, NAVSUP Publication 529, March 1985, p 8-17). The "passive" conveyor (closer to the pickers) holds partially completed orders. The powered "take-away" conveyor transports completed orders to the shipping department.

Under these circumstances, the fast-moving sku's are generally picked from flow rack, as illustrated in Figure 1. Flow rack is arranged in aisles, through which runs a unidirectional conveyor. The racks are divided into *bays*, and within each bay are tilted shelves with rollers to slide the cases forward.

An *order* is a list of sku's for a single customer together with quantities to be picked. Workers assemble each order progressively along the aisle, putting the sku's into *totes* (cartons), which travel together. Workers keep the orders in sequence so they arrive at the shipping dock in reverse order of delivery.

Because broken-case order-picking is so labor-intensive, managers naturally want to keep all pickers busy. Standard practice is to adopt an assembly-line model, partitioning the bays into contiguous sections called *zones* and then restricting each picker to work within her[1] zone. The picker in the first zone begins a new order by opening a tote and sliding it along the passive lane of the conveyor while picking the sku's for that order. On reaching the end of her zone, she leaves the order for the next worker and returns to the beginning of her zone for more work. Each worker remains in her zone, moving totes forward while picking, and possibly standing idle if there is no work in her zone. The last picker pushes the totes of a completed

[1] In our experience most pickers are female.

order onto a powered conveyor, which takes them to the shipping department. The idea, like that of an assembly line, is that all workers will presumably remain busy if their zones have approximately the same total work. This style of order-picking is called *sequential zone-picking*. (For more about order-picking protocols, see "The warehouse manager's guide to effective order picking", Monograph M-8, Tompkins Associates, Inc., Raleigh, NC.)

Under zone-picking each assembly-line must be balanced one or more times a day. To support this the DC must maintain a model of work content on which to base the zones. But the work-content model will always be wrong, despite the effort invested in it, because of issues like the following.

- Work-content models ignore speeds of the workers because their identities will not be known until work begins. Instead, work-content models are based on the notion of a mythical "standard worker". However, it is common, in our experience, for people to differ in work velocity by a factor of three or four, in part because of the use of temporary labor to match large seasonalities in business. Consequently, the rigid zones of an assembly line underutilize the faster workers, while frustrating the slower workers, who, under pressure to keep up, may introduce errors.

- The work-content model attempts to balance only the *total* work accomplished, but fails to maintain balance from order to order.

- There are more factors that determine work content than can be economically modeled: In addition to the number and locations of the sku's to be picked, work content is also determined by heights of the locations (at waist level or inconveniently high?), weight and shape of the sku's (heavy? hard to handle?), and so on. Moreover, such models cannot account for inevitable disruptions such as disposing of an empty case, opening a new case, sealing a full tote, pulling stalled cases to the front of the flow rack, and so on.

Because of these inaccuracies, the work-content model will be wrong and so the assembly-line will not be balanced. This is why zone-picking requires constant supervision — but is imbalanced nonetheless. The cost is reduced pick rates due to underutilized pickers. Furthermore, imbalances cause congestion because the length of the conveyor strictly limits the work-in-process; and congestion further reduces the effective pick rate by making it harder to put product in the right totes.

Bucket brigades seem to be an ideal solution to this problem because they restrict WIP and dynamically balance themselves to achieve high production rates, all without the need of a work-content model.

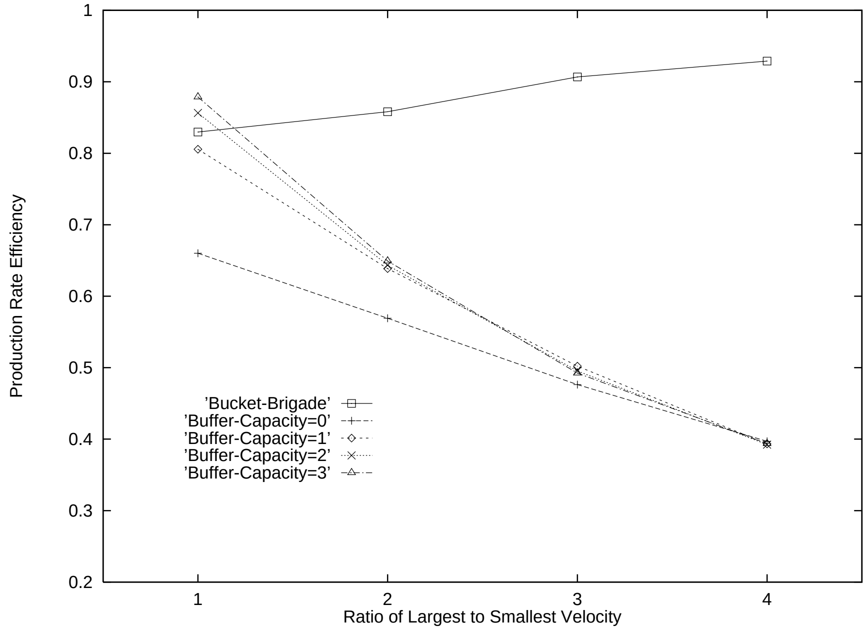

Figure 2: Production rate efficiency decreased with increasing difference in worker velocities for zoned lines; but bucket brigades remained highly productive. The maximum possible value of production rate efficiency is 1, which is achieved by bucket brigades under the deterministic model of work content.

6.2 Bucket brigades vs. zone picking

Our Stochastic Model provides a good description of order-picking in a distribution warehouse. In particular:

1. Walking time in a high volume DC is at least an order of magnitude less than picking time.

2. Workers proceed at different velocities and may be ranked from slowest to fastest because the same skill pertains all along the line. Indeed, many DC's track the individual pick rate of workers and base part of their pay on this; and, in any event, everyone on the floor knows who is faster and who slower.

3. The work at a "station" (storage location) varies from order to order, which suggests a stochastic model of work.

Finally, because the number of stations (storage locations) is much greater than the number of workers (by several orders of magnitude), Theorems 1 and 2 suggest that bucket brigades can be very effective in coördinating work among order-pickers.

We tested this in both simulations and in a commercial distribution center. Figure 2 shows typical simulation results comparing a bucket brigade, with workers sequenced from slowest to fastest, to zone-picking that allows up to 0, 1, 2, or 3 units of WIP to build between adjacent zones. Each line has 5 workers and 20 work stations, with work at each station independently following an exponential distribution. The velocities of the team are spread uniformly with the ratio of the velocity of the fastest to slowest worker varying along the x-axis from 1 (all workers identical) to 4 (the last worker is 4 times the velocity of the first), which are representative of our observations in practice. To make the comparison meaningful, we imposed a constraint that the sum of the velocities of all worker remain constant, so that each team had the same inherent productive capacity.

As is common in industry, we balanced zones based on a common work standard. Then the workers were sequenced as closely as possible to adhere to the "bowl" phenomenon (Hillier and Boling, 1979). In addition, we granted a special advantage exclusively to the simulated zone-picking by not penalizing it for accumulation of WIP, which in real life slows throughput by creating opportunities to put product in the wrong tote.

In this family of simulations we measured *production rate efficiency*, the realized production rate divided by the maximum possible rate, which is the sum of the velocities of the workers. The largest possible value of production rate efficiency is 1 and this is achieved by bucket brigades under the deterministic model (cf. Theorem 1). When all workers were identical (which, of course, is never the case in the real world) the production rate efficiency of the simulated bucket brigade was similar to that of zone picking that allows WIP between stations. But as the velocities of the workers were allowed to become distinct, as one invariably finds in the real world, then bucket brigades were more productive. This is because bucket brigades spontaneously and continually adjust to account for variances in the system, including variances in the velocities of workers and in the amount and location of work.

6.3 Experience at Revco Drug Stores, Inc.

The strongest proof of the effectiveness of bucket brigades when work varies comes from practice.

We implemented order-picking by bucket brigade at the national distribution center of Revco Drugs, Inc., which supports over two thousand retail outlets. A key advantage of bucket brigades is simplicity, so that implementation required less than an hour, with no special equipment and no changes to the warehouse management system or related operations. This made it easy to experiment one morning on a single aisle that had previously been using sequential zone-picking. We described the idea to the workers in fifteen minutes, sequenced them from slowest to fastest, and watched them work.

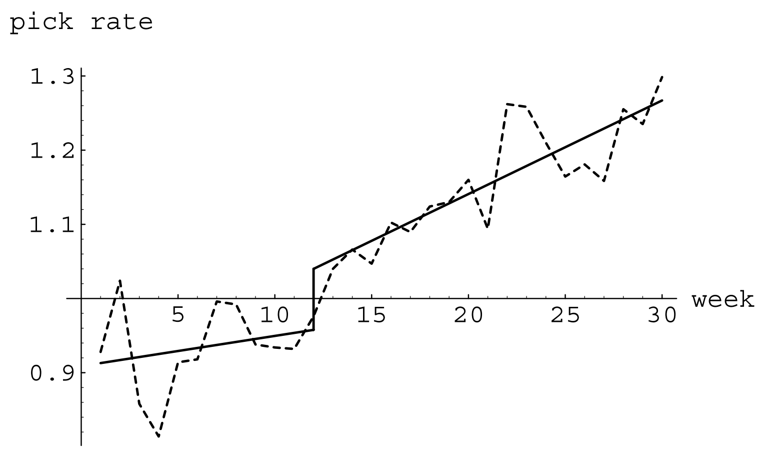

Figure 3: Average pick rate as a fraction of the work-standard. Zone-picking was replaced by bucket brigade in week 12. (The solid lines represent best fits to weekly average pick rates before and after introduction of the bucket brigade protocol.)

The most striking benefit of order-picking by bucket brigade was the increase in pick rates, which reached sustained levels of 34% greater than the previous historical averages under zone-picking, while simultaneously reducing management intervention (Figure 3). This was achieved at essentially no cost, and in particular, with no change to the product layout, equipment, or control system (except to render parts of the latter unnecessary).

Picking by bucket brigade produced additional benefits, including the following.

- Spontaneous (re)balance of the work has freed management time. Previously each aisle was monitored by a manager who adjusted zones within the aisle to correct the inevitable intermittent imbalances and the resulting congestion or starvation. This level of supervision is no longer necessary because adjustments are spontaneous and continual.

  Furthermore, differences in work rates are now visible and so it has become easier to recognize problems. For example, at Revco each bay contains comparable amounts of work for each order and so, under the bucket brigade protocol, each worker tends to visit a length of aisle proportional to her pick rate. In one case, an unusually slow worker at the first position was repeatedly "pushed back" by her faster teammates: She was unable to pick quickly enough ever to leave the first bay of flow rack and so her teammates asked that she be removed. It was apparent to all that they could pick as fast without her and they preferred to split the incentive pay $n-1$ ways. Under zone-picking such imbalances were harder to recognize because they could be hidden by work-in-process.

- The synchronization of multiple aisles has become easier. A manager can now monitor the progress of an aisle by simply checking what order any worker is picking. Under zone-picking it was difficult to know the status of an aisle because of the considerable and variable work-in-process.

  It has also become easier to move workers to maintain the balance among aisles. Under zone-picking, when one picker was moved, work was interrupted while management redefined the zones in each aisle; but under bucket brigades, the pickers in each aisle spontaneously adjust to account for the new configuration.

- A bucket brigade is extensible. For example, at Revco there was a worker picking from carousels immediately upstream from one aisle; and she occasionally got ahead of the workers in that aisle. Under zone-picking she had to cease working until the congestion was cleared. Now she simply joins the bucket brigade to help them pick. After they have caught up, she returns to the carousels at the next walkback.

- Reduced levels of work-in-process increased the accuracy of order-picking. Because the number of totes on the conveyor is strictly controlled, there is no congestion and workers rarely put sku's in the wrong totes.

- The pickers claim to be more satisfied because they prefer working in teams, with clear instructions about where to go and when. Furthermore, the simplified and regularized movements mean that temporary workers can become productive more quickly.

- The expense and inaccuracy of a work-content model can be avoided. Revco had calculated zones several times a day based on a sophisticated, computer-hosted model of work content and advance knowledge of all 100,000 pieces to be picked that day. With bucket brigades, Revco can abandon its detailed work-content model and get better balance and higher productivity nonetheless.

Revco has subsequently implemented bucket brigades in all its regional warehouses, involving hundreds of order-pickers, all of whom had previously picked by zone. As of this writing they have been successfully using bucket brigades for over four years.

7 Extensions and open problems

In the statement and proof of Lemma 1 we have followed the Normative Model, in which the velocity of worker i is a constant, v_i, unless blocked by the worker immediately downstream. However, the queueing equivalence used in our proof allows extending the result to more general models in which the velocity of each worker may be either state dependent (dependent on the locations of the workers, as in Bartholdi and Eisenstein, 1996), or time dependent. (Discussions of state- and/or time-dependent dynamic complementarity may be found in Appendices 2 and 3 of Pats, 1995).

A result similar to Theorem 2 holds when workers are sequenced other than slowest to fastest. However, the Skorohod $J1$ topology does not allow several jumps accumulating at the same point in time, which could occur if $v_i > v_{i+1}$ or if $v_i = v_{i+1}$ and $x_i = x_{i+1}$ (in which case the conditions at the end of the proof fail, with Θ_k^m/m and Θ_{k+1}^m/m both converging to the same time θ_j). Therefore, to show a similar result for any sequence of workers (not just slowest to fastest), we would need to use an even weaker topology than the Skorohod J1. However, the model with instantaneous movement and $v_1 \leq v_2 \leq \cdots \leq v_n$ is the most interesting, so we have presented the analysis for this case only.

Other researchers have considered stochastic models of work-sharing on a flow line; but all assume that the workers are identical in velocity (Bischak, 1996; Zavadlav, McClain, and Thomas, 1996). We believe that it is unrealistic to assume that workers proceed at identical speeds when the workers are humans. However, assuming the workers are identical can be useful: In particular, it gives a case in which the throughput is easily computed and should be a lower bound on achievable throughput for heterogeneous workers. If the n workers have identical velocities, then the columns of the generator also sum to zero. Hence, the stationary distribution has all states equally likely, as when the transition matrix of a Markov chain is doubly stochastic. The throughput is simply the proportion of states with worker n at the last machine times the velocity of a single worker. If we scale the velocities $v_i = m/n$, then the system throughput would be one if there were no blocking. The actual system throughput is simply

$$\frac{\binom{n+m-2}{n-1}}{\binom{n+m-1}{n}} \frac{m}{n} = \frac{m}{n+m-1}. \tag{21}$$

This expression appears in Bischak (1996) except that the velocities of the workers are μ instead of m/n. Bischak derived the result by showing an equivalence between bucket brigades and cyclic queues when worker's velocities are equal.

Note that the throughput achieves the maximum possible rate of 1 when there is $n = 1$ worker. Also note that throughput increases with the number of machines m, but decreases with the number of workers n. Of course, this is under the assumption that increasing the number of workers does not increase their combined work rate m, but simply splits it evenly over more workers. Thus, the result implies that it is better to have fewer workers with the same combined speed than many. As the lower bound suggests, we would not expect bucket brigades to work particularly well when there is a small number of machines and a relatively large number of workers. For example, if there are two workers and three machines, the lower bound guarantees a throughput of only 3/4. Of course, it is even worse if there are three workers and two machines since at least one of the workers is always blocked and the lower bound drops to 1/2. However, if m is large relative to n, bucket brigades should function well, as they have in practice.

Expresion 21 also gives an upper bound on the fraction of production rate lost due to blocking:

$$1 - \frac{m}{n+m-1} = \frac{n-1}{m+n-1} \tag{22}$$

The preceding discussion alludes to two interesting open problems for the Stochastic Model. The first problem is which arrangement of workers is optimal in the bucket brigade? The intuitively obvious answer is slowest to fastest, but this is unproven. The second problem is whether assuming workers are homogeneous (that is, their combined speed is divided evenly among the workers) provides a lower bound on the production rate of the workers arranged slowest to fastest. Again the intuitively obvious answer is yes, but this result is also unproven. These two results together with the simple expression (21) would combine to give a useful lower bound on throughput of any bucket brigade with m machines and n workers sequenced slowest to fastest.

When work-content is "sufficiently variable", bucket brigades could, in principle, be out-performed by a policy that allowed instantaneous resequencing of the workers. For example, if at some instant worker 1 was far behind worker 2, and worker $n-1$ was close to worker n, it would be better to swap workers 1 and $n-1$ to decrease the likelihood of blocking in the near future. It is an interesting control problem to determine which sequence of workers is optimal at each instant; however, it is unlikely that an optimal policy for this model would be worth implementing in most real world situations.

8 Conclusions

The main benefits of bucket brigades are increased production rate, reduced dependence of work-content models, and simplified management. These benefits were so substantial at Revco that other, initial concerns, such as whether brigade members might shirk or free-ride, were dismissed as second order effects at best.

Bucket brigades can be more productive than traditional assembly lines for a number of reasons: First, bucket brigades constantly and spontaneously seek balance; and second, balance is based on the actual, realized work content, and the particular workers — and not mere estimates of work content based on standardized workers. Furthermore, bucket brigades can achieve high production rate without resorting to high work-in-process because they absorb variance in the work by moving the workers where the work is. Of course the strongest "proof" of the effectiveness of bucket brigades is experience across a range of commercial applications, one of which we have reported here. Others may be found at our web site `www.isye.gatech.edu/faculty/John_Bartholdi/bucket-brigades`.

Our work may be seen to lie within two current streams of thought. Most immediately, it is a special case of *dynamic line-balancing*, wherein an intelligent controller adjusts the allocation of work in real time (for example, Ostolaza, Thomas, and McClain, 1990). For bucket brigades the allocation occurs spontaneously, which has the considerable advantage of requiring no controller at all. Furthermore, as of this writing, bucket brigades are unique in that local adjustments (worker movement) have been proved to lead to global balance.

The second stream of thought into which our work fits is the hosting of computational processes on analog devices. In our case the assembly-line is the computer of its own allocation of work. It might be said that we program this computer by sequencing the workers from slowest to fastest. There is no need to measure and input data because the work content is read directly by the doing of it. The output is the balance.

Acknowledgments

We thank the many people who contributed to the success of the implementation at Revco, including Dave Cole, Jim Rollins, Joe Jernigan, Ron Kelly, and especially Victor Lee and Pam Hinkle. We also thank Dave Wolfe of TransTech Consulting, Inc. for sharing his considerable experience and insights in order-picking; and we thank Leonid Bunimovich and Jim Dai for helpful technical discussions. In addition we thank two anonymous referees for their helpful suggestions.

We are grateful to the many others with whom we have collaborated in implementing bucket brigades, especially those at Anderson Merchandisers, Blockbuster Music, Dell Computers, Ford Motor Company, Mitsubishi Consumer Electronics, and QSP Distribution, Inc.

We appreciate the support of the Office of Naval Research through grant #N00014-89-J-1571 (Bartholdi); and the Graduate School of Business at the University of Chicago (Eisenstein).

Portions of this paper have been presented at the Massachusetts Institute of Technology (March 1996); the University of Chicago (May 1996); the national meeting of the Warehousing Educational and Research Council (May 1996); INFORMS-San Diego (May 1997); American Wholesale Marketers Association Convention, Houston (February 1998); Instituto Tecnologico y de Estudios Superiores de Monterrey, Toluca, Mexico (March 1998); IIE National Meeting, Banff, Canada (May 1998); The North American Warehousing and Distribution Exposition and Conference (September 1998); Ford Customer Service Division (October 1998); the University of California at Berkeley (November 1998); the Wharton School (December 1998); and several industry short courses. We thank the audiences for many stimulating questions.

References

[1] J. J. Bartholdi, III, L. A. Bunimovich, and D. D. Eisenstein (1999). "Dynamics of 2- and 3-worker 'bucket brigade' production lines", *Operations Research* **47**(3):488–491.

[2] J. J. Bartholdi, III and D. D. Eisenstein (1996). "A production line that balances itself", *Operations Research* **44**(1).

[3] D. P. Bischak (1996). "Performance of a manufacturing module with moving workers", *IIE Transactions* **28**(9).

[4] J. G. Dai (1995). "On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models", *The Annals of Applied Probability* **5**(1):49–77.

[5] S. N. Ethier and T. G. Kurtz (1986). *Markov Processes: Characterizations and Convergence,* Wiley.

[6] F. S. Hillier and R. W. Boling (1979). "On the optimal allocation of work in symmetrically unbalanced production line systems with variable operation times", *Management Science* **25**(8).

[7] J. Ostolaza, J. O. McClain, and L. J. Thomas (1990). "The use of dynamic (state-dependent) assembly-line balancing to improve throughput", *Journal of Manufacturing and Operations Management* **3**:105–133.

[8] G. Pats (1995). *State Dependent Queueing Networks: Approximations and Applications,* Ph. D. Thesis, The Technion – Israel Institute of Technology.

[9] J. M. Harrison, and M. I. Reiman (1981). "Reflected brownian motion on an orthant", *The Annals of Probability,* **9**, 302–308.

[10] A. V. Skorohod (1956). "Limit theorems for stochastic processes", *Theory of Probability and Its Applications,* **1**, 261–290.

[11] E. Wong and B. Hajek (1985). *Stochastic Processes in Engineering Systems,* Springer-Verlag.

[12] E. Zavadlav, J. O. McClain, and L. J. Thomas (1996). "Self buffering, self balancing, self flushing production lines", *Management Science* **42**(8):1151–1164.

A Proof of Lemma 1

Proof. To solve the set of equations we first consider a related problem in which the workers are not constrained to $[0, 1)$, but instead are allowed to continue moving to the right on $[0, \infty)$ without ever resetting. This process $(\hat{X}, \hat{T})$ will be the solution to the following set of equations, which hold for $i = 1, \ldots, n$ and $t > 0$:

$$\hat{X}_i(t) = x_i + v_i \hat{T}_i(t) \tag{23}$$

$$\hat{X}(t) \in \hat{E} \tag{24}$$

$$\hat{T}_i(0) = 0 \text{ and } \hat{T}_i(t) \text{ is non-decreasing} \tag{25}$$

$$\hat{I}_i(t) = t - \hat{T}_i(t) \tag{26}$$

$$\int_0^\infty \mathbf{1}\left(\hat{X}_{i+1}(t) > \hat{X}_i(t)\right) d\hat{I}_i(t) = 0 \tag{27}$$

where $\hat{E} \equiv \{(\ell_1, \ldots, \ell_n) | 0 \leq \ell_1 \leq \ell_2 \leq \cdots \leq \ell_n < \infty\}$.

If we define $Q_i(t) \equiv \hat{X}_{i+1}(t) - \hat{X}_i(t)$, we can view $(Q_1(t), \ldots, Q_{n-1}(t))$ as the vector of amounts of fluid in a deterministic fluid queueing system consisting of $n - 1$ servers in tandem. Fluid arrives continuously to the last queue, $Q_{n-1}(t)$, at rate v_n. The i-th queue pumps fluid to the $(i-1)$-st queue at rate v_i as long as fluid is present. Fluid pumped out of the first queue is lost from the system. Rewriting equations (23)–(27) in terms of $(Q_1(t), \ldots, Q_{n-1}(t))$ yields a special case of the dynamic complementarity problem (DCP) discussed in Dai (1995), Pats (1995), and Harrison and Reiman (1981). From Theorem 1 in Harrison and Reiman (1981), there exists a unique solution to these rewritten equations even when they are restricted to $t \in [0, M]$ for any positive M; hence, there exists a unique solution $(\hat{X}, \hat{T}, \hat{S})$ to (23)–(27). Let $\theta_1 \equiv \inf\{t \geq 0 | \hat{X}_n(t) = 1\}$, which will be the first reset time. Thus, for $t \in [0, \theta_1)$, we must have $\bar{X}(t) = \hat{X}(t)$, $\bar{T}(t) = \hat{T}(t)$ and $\bar{S}(t) = 0$.

Now assume that $(\bar{X}, \bar{T}, \bar{S})$ is uniquely defined for $t \in [0, \theta_k)$. Redefine and reconstruct $(\hat{X}, \hat{T})$, except use $R(\bar{X}(\theta_k-))$ as the starting position of the workers. This must be the starting point due to (1) and (6). Define $\theta_{k+1} = \theta_k + \inf\{t \geq 0 | \hat{X}_n(t) = 1\}$. Note that $R(\bar{X}(\theta_k-)) \in E$ and $\theta_{k+1} > \theta_k$. For $t \in [\theta_k, \theta_{k+1})$, define $\bar{X}(t) = \hat{X}(t)$, $\bar{T}(t) = \hat{T}(t) + \bar{T}(\theta_k)$, and $\bar{S}(t) = \bar{X}(\theta_k-) - R(\bar{X}(\theta_k-)) + \bar{S}(\theta_k)$. Note that θ_k is not necessarily the k-th reset time because more than one reset may occur (more than one item may be produced) at θ_k. Continuing in this fashion, we will have constructed $(\bar{X}, \bar{T}, \bar{S})$ which is the only solution to (1)–(6) provided $\theta_k \to \infty$. However, this follows since the number of items produced in any interval of length t is bounded above by $n + t/(v_1 + \cdots + v_n)$. □

B Proof of Theorem 2

Proof. Let Θ_1^m be the first reset time in the stochastic model, and let Ψ^m be the first time that that one of the workers is blocked in the stochastic model. In the rescaled model, the first reset time is Θ_1^m/m and the first time blocking occurs is at Ψ^m/m. During the interval $[0, \min[\Theta_1^m/m, \Psi^m/m])$, we have $\tilde{X}^m(t) = \tilde{X}^m(0) + N((mt))/m$. Since $N(mt)/m \stackrel{u.o.c.}{\longrightarrow} vt$ and $X^m(0) \to x$, we have

$$\tilde{X}^m(t) \to x + vt \text{ uniformly for } 0 \leq t < \min[M, \lim_{m\to\infty} \min[\Theta_1^m/m, \Psi^m/m]], \tag{28}$$

where M is any positive finite constant. Under the assumptions that $x_i < x_{i+1}$ and $v_i \leq v_{i+1}$, the ith and $i+1$st coordinates of $x + vt$ differ by at least $x_{i+1} - x_i$ for all $t \geq 0$. Hence, $\Pr\{\Theta_1^m/m < \Psi^m/m\} \to 1$ and $\Theta_1^m/m \to \theta_1$, where θ_1 is the first reset time of the fluid model as used in the proof of Lemma 1. To see that $\Pr\{\Theta_1^m/m < \Psi^m/m\} \to 1$, note that if $\Psi^m/m \leq \Theta_1^m/m$, then two workers are at the same location, i.e., have zero separation, at time Ψ^m/m. If $\Pr\{\Theta_1^m/m < \Psi^m/m\} \to 1 - \epsilon$ with $\epsilon > 0$, then (28) would not hold. Thus, $\tilde{T}^m(t) \to et$ for $t \in [0, \theta_1)$ and $\tilde{X}^m(\Theta_1^m/m-) \to \bar{X}(\theta_1-)$; hence,

$$\tilde{X}^m(\Theta_1^m/m) = (0, \tilde{X}_1^m(\Theta_1^m/m-), \dots, \tilde{X}_{n-1}^m(\Theta_1^m/m-)) \to \bar{X}(\theta) = R(\bar{X}(\theta_1-)).$$

Furthermore, $\bar{X}(\theta_1)$ has the property that $0 \le \bar{X}_1(\theta_1) < \bar{X}_2(\theta_1) < \cdots < \bar{X}_n(\theta_1) < 1$. Hence, we can repeat the argument on the next interval of time $[\theta_1, \theta_2)$. Since $\theta_k < \theta_{k+1}$ and since for any time t, the number of resets is at most $n + t/(v_1 + \cdots + v_n)$, we can see that $\tilde{X}^m(t) \to \bar{X}(t)$ for $t \notin \{\theta_1, \theta_2, \dots\}$, $\Theta_k^m/m \to \theta_k$, $\tilde{T}^m(t) \to et$, $\tilde{X}^m(\Theta_k^m/m-) \to \bar{X}(\theta_k-)$, and $\tilde{X}^m(\Theta_k^m/m) \to \bar{X}(\theta_k)$.

To show $\tilde{X}^m \xrightarrow{J1} \bar{X}$, we use Proposition 6.5 in Chapter 3 of Ethier and Kurtz (1986). In our case, it suffices to show that whenever $\{t_m\} \subset [0, \infty)$, $t \ge 0$, and $t_m \to t$ the following conditions hold, where r is the sup norm:

- $\min(r(\tilde{X}^m(t_m), \bar{X}(t)), r(\tilde{X}^m(t_m), \bar{X}(t-)) \to 0$.
- If $r(\tilde{X}^m(t_m), \bar{X}(t)) \to 0$, $s_m \ge t_m$ for each m, and $s_m \to t$, then $r(\tilde{X}^m(s_m), \bar{X}(t)) \to 0$.
- If $r(\tilde{X}^m(t_m), \bar{X}(t-)) \to 0$, $0 \le s_m \le t_m$ for each m, and $s_m \to t$, then $r(\tilde{X}^m(s_m), \bar{X}(t)) \to 0$.

These conditions are easy to show after noticing that there exists $\Delta > 0$ such that $\theta_{k+1} - \theta_k \ge \Delta$ for $k = 1, 2, \dots$. □