Optimizing Space Utilization in Block Stacking Warehouses

Shahab Derhami, Jeffrey S. Smith and Kevin R. Gue

Industrial & Systems Engineering, Auburn University, Auburn, AL 36849, USA; Industrial Engineering, University of Louisville, Louisville, KY 40292, USA

(Received 00 Month 20XX; accepted 00 Month 20XX)

Block stacking storage is an inexpensive storage system widely used in manufacturing systems where pallets of stock keeping units (SKUs) are stored in a warehouse at the finite production rates. However, determining the optimal lane depth that maximizes space utilization under a finite production rate constraint has not been adequately addressed in the literature and is an open problem. In this research, we propose mathematical models to obtain the optimal lane depth for single and multiple SKUs where the pallet production rates are finite. A simulation model is used to evaluate performance of the proposed models under stochastic uncertainty in the major production parameters and the demand.

Keywords: block stacking; facility layout; warehouse design; space utilization

1. Introduction

Optimizing space utilization has been one of the main goals in designing and operating warehouses (Van den Berg 1999). The U.S. Roadmap for Material Handling and Logistics recognizes low warehouse utilization as one of the main factors that propels companies, associations and governments to employ collaborative warehouses more in the next decade. It also predicts that requests for high speed delivery or same-day delivery forces companies to build their warehouses and distribution centers near major metropolitan area where real estate is very expensive and therefore efficient use of space becomes more important (Gue et al. 2014).

Various approaches from the design to the operational phase of a warehouse have been developed to better utilize storage space. Block stacking is an inexpensive and conventional storage system whose performance depends on the efficient use of space. It is a unit load storage system in which pallets of stock keeping units (SKUs) are stacked on top of one another in lanes on the warehouse floor. Pallets are stacked to the maximum stacking height which depends on the conditions and heights of the pallets, load weights, safety limits, clearance height of the warehouse, and so on. No racking or storage facility is required for this system and it can be employed in any warehouse with wide floor space. This makes it an inexpensive storage system to implement but challenging to manage in terms of space planning.

Two major operating policies that are widely used to manage storage spaces in this system are dedicated and shared storage policies. In the dedicated policy, lanes are dedicated to SKUs and only pallets of the assigned SKU are allowed to be stored in a lane. So, a lane may remain empty until it is replenished with its assigned SKU. On the other hand, lanes are not dedicated to any SKUs in the shared storage policy and they are available to all SKUs once they become empty. This policy utilizes space more efficiently than the former one though the order picking process is generally less efficient since the SKUs storage locations change over time and SKUs are assigned

*Corresponding author. Email: sderhami@auburn.edu
Figure 1. Waste of space in Example 1

The shared policy is operated with allowing or not allowing blockage. When the variety of SKUs is large but their inventory is small, assigning a lane to a single SKU is not justifiable and therefore different SKUs have to be stored in the same lane. In this case, blockage is inevitable and the goal is to arrange SKUs such that the relocation costs are minimized. An example of such a case is the storage space allocation problem in a marine container terminal (Yang and Kim 2006; Petering and Hussein 2013; Jang, Kim, and Kim 2013; Carlo, Vis, and Roodbergen 2014). On the other hand, when the inventories of the SKUs are big enough to justify assigning a lane to a single SKU, no blockage policy is enforced. In this case to avoid blockage and relocation, a lane is dedicated to a SKU once it occupies the first position of the lane. This case mainly occurs in the warehouses located in manufacturing systems or the distribution centers where plenty of pallets of different SKUs are block-stacked. However, this restriction wastes storage spaces in a lane when it is filled or depleted as there will be some unoccupied pallet positions in the lane that are just available to the pallets of the assigned SKU. This effect is termed honeycombing and waste associated with it is incurred to the system until a lane becomes entirely occupied or empty (Hudock 1998).

In addition to honeycombing, aisles also contribute to the overall wasted space. Aisles are required to have access to the lanes but their devoted spaces are not directly used for storage. Warehouse designers aim to minimize these two types of waste to enhance space utilization in the warehouse. The following example shows how waste of storage space is calculated in a lane.

**Example 1.** Consider a batch of 10 pallets of a SKU is stored in a lane of two pallets deep. Pallets are produced at the rate of \((1/5)\) pallets per hour, stacked up to two pallets high and depleted at the rate of \((1/18)\) pallets per hour. Assume that an aisle with width equivalent to two pallets is required to access the lane. Figure 1 shows waste of storage space generated in the inventory cycle time of this SKU. At time zero, waste is zero because an empty lane is available to all SKUs. At time five, the first pallet is stored in the lane and makes the three unoccupied pallet positions in the lane unavailable to the other SKUs. This is the honeycombing waste. Moreover, four pallet positions are dedicated to the aisle to provide accessibility to the lane. So, seven pallet positions are wasted at this time. At times 10 and 15, the next two pallets are stored in the lane and the total number of wasted positions decreases to six and then five pallets, respectively. The first depletion event occurs at time 18 and increases the number of wasted positions to six. This procedure continues until all 10 pallets are produced, stored in the lane, and depleted. The area under the waste plot in Figure 1 is the pallets-time waste of storage space during the inventory cycle time, and dividing it by 175 hours gives the average waste of storage space in the inventory cycle time which is 7.48 pallets.

Storing a batch of pallets in deep lanes increases the honeycombing waste, but the space required for aisles decreases while the reverse is true for the shallow lanes. Hence, a trade-off between the
lane depth and width must be considered to optimize space utilization. This trade-off is shown in Figure 2 by comparing the average waste of storage space generated by storing the SKU described in Example 1 in lanes with one to four pallets deep. In this case, the minimum waste achieved when the SKU is stored in the lanes with three pallets deep.

In this paper, we consider this trade-off from a mathematical point of view and develop models to compute an optimal lane depth that minimizes waste of storage space in the warehouse. Our models are different from those existing in the literature in two major respects. First, the instantaneous pallet storage assumption, which was made in all previous research is relaxed and models are built for finite production rates. Hence, they are more suitable for the warehouses located in manufacturing systems where pallets are stored at finite production rates. Second, our proposed models aim to maximize utilization of the volume instead of the floor space. A review of the previous research on block stacking is provided in the next section.

2. Related research

Various studies have investigated designing the layout of a warehouse (Gu, Goetschalckx, and McGinnis 2010). Most of them considered designing the layout with respect to the construction and maintenance costs (Ashayeri, Gelders, and Wassenhove 1985; Rosenblatt, Roll, and Zyser 1993; Pazour and Meller 2011), material handling costs in order picking process (Gue and Meller 2009; Pohl, Meller, and Gue 2009; Ömer Öztürköğlu, Gue, and Meller 2012, 2014; Ramtin and Pazour 2014), products allocation (Heragu et al. 2005; Ramtin and Pazour 2015) and few of them considered objectives pertinent to the space utilization (Gue 2006). Extensive reviews on different approaches used to design different storage systems are found in Gu, Goetschalckx, and McGinnis (2010); Baker and Canessa (2009); Gu, Goetschalckx, and McGinnis (2007); Klose and Drexel (2005); Rouwenhorst et al. (2000). Studies that investigated space utilization in the block stacking storage systems are reviewed in the following.

To the best of our knowledge, Kind (1975) was the first person who considered the trade-off between the lane depth and width in the block stacking storage and proposed a model to approximately find the optimal lane depth that minimizes waste of storage space. He proposed this approximation for a single SKU whose batch of pallets are instantaneously stored in a warehouse. Marsh (1979) developed a simulation model to evaluate different storage and operating policies for block-stacking. Later, Matson (1982) developed a more accurate version of Kind’s approximation (Kind 1975) for a single SKU and extended it to obtain the optimal common lane depth for multiple SKUs. Her models aim to maximize utilization of the floor space (area).

Goetschalckx and Ratliff (1991) showed that if a batch of pallets of a SKU is allowed to be stored in lanes with unlimited different depths, then the optimal lane depths follow a continuous
triangular pattern. They developed a continuous and discrete approximations to obtain the optimal multiple lane depths considering limited and unlimited lane depths. They compared their proposed models with scenarios like equal lane depth models developed by Matson (1982) and two extreme heuristics in which a batch of pallets is stored in the lanes whose depths are equal to one or to the batch size, respectively. They concluded space utilization is “relatively insensitive” to the lane depth and all heuristic methods except the two extreme cases provide comparatively equivalent results in terms of accuracy and computational complexity. However, their approaches, especially the one that assumes unlimited multiple lane depths, are not practical for multiple SKUs.

Larson, March, and Kusiak (1997) proposed a heuristic approach to design the layout of a block stacking warehouse where the objectives are maximizing space utilization and minimizing transportation costs. Their class-based storage approach classifies SKUs based on the throughput to the required storage space ratio, ranks classes based on their average ratios, and constructs and dedicates the storage regions to the classes considering their ranks and required storage spaces. The algorithm considers honeycombing, fluctuations in the inventory level and the maximum stacking height to determine storage medium (i.e., racks or floor stacking) for a SKU, and assumes randomized storage policy among the classes (storage zones).

Bartholdi (2014) developed Matson’s model to optimize volume utilization instead of the floor utilization. He suggested that maximizing the volume utilization is the better objective in the current modern warehouse because volume within a warehouse is worth as much as floor space in today’s modern warehouses.

All aforementioned studies assumed pallets of SKUs are instantaneously stored in a warehouse. In practice, this case occurs in a distribution center where trucks quickly unload pallets of SKUs, and hence it appears realistic to assume infinite arrival rate for incoming pallets. However, this assumption cannot be justified in the warehouses located in manufacturing systems where pallets of SKUs are stored in the warehouse at finite production rates. In such a warehouse, waste of storage space is generated either when a lane is filled or depleted. However, the traditional models are not capable of taking the first part into account and therefore do not correctly compute the optimal lane depth for such cases. We address this drawback in this paper by developing models to determine the optimal lane depth for both single and multiple SKUs under the finite production rate constraint.

3. Optimal lane depth

Minimizing the waste of storage space maximizes space utilization. So, one can obtain the optimal lane depth that maximizes the space utilization by deriving a mathematical expression to calculate waste of storage space and then finding the optimal lane depth that minimizes this waste expression. The waste of storage space is obtained by calculating three types of waste:

1. Waste of storage space caused by honeycombing, $W_H$.
2. Waste of storage space dedicated to the aisles, $W_A$.
3. Waste of unoccupied storage space on top of the occupied lanes, $W_U$.

$W_U$ is incurred as the result of different stacking and pallet heights for different SKUs. This waste is not computed in the single SKU models as it does not affect the optimal lane depth in that cases. Figure 3 shows total waste of storage space and its components with respect to the optimal lane depth. The relation between the total waste of storage space and lane depth is analogous to the relation between the total cost and order quantity in the Economic Order Quantity (EOQ) model.

In general, the number of SKUs stored in a warehouse is too numerous to assign all SKUs to their optimal lane depths and assort all the lane depths in the warehouse. To overcome this issue, the optimal common lane depth is computed — one that minimizes total waste for multiple SKUs.
The models in this section are derived by assuming that a batch of \( Q \) pallets of a SKU is produced (or unloaded to a warehouse) at the deterministic production rate of \( P \) pallets per unit of time and block-stacked in the lanes of \( x \) pallets deep to the height of \( z \) pallets. Pallets are depleted at the deterministic rate of \( \lambda \) pallets per unit of time and aisles with \( a \) pallets width are required to access the lanes. Table 1 describes the notation used in the following models. The following assumptions are made for all models in this paper:

1. Lanes are accessible from one side and as a result, they are depleted based on last-In-first-out (LIFO).
2. Partially occupied lanes are prioritized to be depleted first. This helps to utilize space more efficiently because unlike the fully occupied lanes which incur only accessibility wastes to the system, a partially occupied lane generates both honeycombing and accessibility waste. Thus, the longer it remains incomplete, the more storage space is wasted. Another advantage of this policy is that these lanes are depleted faster than the fully occupied lanes and consequently, their devoted spaces are released sooner.
3. Fully occupied lanes are depleted in an arbitrary order. This is because the order of depleting such lanes does not affect waste of storage space.
4. The production, if exists, and the depletion quantities are one pallet at a time and the storage, and depletion times are assumed to be zero.
5. No safety stock is kept in the warehouse. Production stops once all pallets of a batch are produced and it is restarted when the inventory of the SKU becomes zero.

3.1 \( P = \infty \), **Demand is constant**

In this case, pallets of SKUs are instantaneously stored in a warehouse and depleted at a finite rate. So, the production rate is considered to be infinite. Figure 4 shows the inventory level of such a SKU over its inventory cycle time. Kind (1975) proposed the following formula to estimate the optimal lane depth in this case:

\[
x^* = \sqrt{\frac{Qa}{z}} - \frac{a}{2}
\]

Figure 3. Total waste of storage space

However, he did not provide any derivation for his formula. Later, Matson (1982) developed another approximation for the optimal lane depth. We derive the optimal lane depth for this case by providing a correction on the model proposed by Matson (1982) and also considering volume utilization instead of the floor utilization as proposed by Bartholdi (2014). The correction herein is
Table 1. Table of notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>production rate (in units of pallet/time)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>depletion rate (in units of pallet/time)</td>
</tr>
<tr>
<td>$x$</td>
<td>lane depth (in units of pallet)</td>
</tr>
<tr>
<td>$x^*$</td>
<td>optimal lane depth for single SKU (in units of pallet)</td>
</tr>
<tr>
<td>$x_c^*$</td>
<td>optimal common lane depth for multiple SKUs (in units of pallet)</td>
</tr>
<tr>
<td>$z$</td>
<td>stackable height (in units of pallet)</td>
</tr>
<tr>
<td>$Q$</td>
<td>production (arrival) batch quantity (in units of pallet)</td>
</tr>
<tr>
<td>$H$</td>
<td>maximum inventory level (approximation)</td>
</tr>
<tr>
<td>$K$</td>
<td>maximum number of lanes required for storage (approximation)</td>
</tr>
<tr>
<td>$a$</td>
<td>aisle width (in units of pallets)</td>
</tr>
<tr>
<td>$h$</td>
<td>height of a pallet of a SKU (in units of distance i.e., inch, cm)</td>
</tr>
<tr>
<td>$e$</td>
<td>clear height of the warehouse (in units of pallet)</td>
</tr>
<tr>
<td>$n$</td>
<td>number of SKUs</td>
</tr>
<tr>
<td>$T$</td>
<td>inventory cycle time</td>
</tr>
<tr>
<td>$O_T$</td>
<td>occupied space-time in the inventory cycle time</td>
</tr>
<tr>
<td>$U$</td>
<td>space utilization (single SKU)</td>
</tr>
<tr>
<td>$U_c$</td>
<td>space utilization for a common lane depth (multiple SKUs)</td>
</tr>
<tr>
<td>$W_H$</td>
<td>waste of storage space caused by honeycombing</td>
</tr>
<tr>
<td>$W_A$</td>
<td>waste of storage space dedicated to the aisles</td>
</tr>
<tr>
<td>$W_U$</td>
<td>waste of unoccupied storage space on top of the occupied lanes</td>
</tr>
<tr>
<td>$W$</td>
<td>average waste of storage space (single SKU)</td>
</tr>
<tr>
<td>$W_c$</td>
<td>average waste of storage space for a common lane depth (multiple SKUs)</td>
</tr>
</tbody>
</table>

Figure 4. Changes in the inventory of a SKU stored instantaneously, $P = \infty$

Correcting the approach used to calculate total space dedicated to the aisles. However, applying this correction does not change the original lane depth model proposed by Matson (1982) for the single SKU, but the model for multiple SKUs changes as the result of optimizing volume utilization.

3.1.1 Optimal lane depth for a single SKU

The number of lanes required for storage is $\lceil Q/zx \rceil$, where $\lfloor x \rfloor$ is the smallest integer not less than $x$. Relaxing the integrality restriction, the approximate number of lanes would be

$$K \approx \frac{Q}{zx}$$

Assume a fully occupied lane is being depleted at the rate of $\lambda$ pallets per unit of time. Once
the first pallet is depleted, the lane will have one unoccupied but unavailable position to the other
SKUs. This waste of storage space remains in the lane for the time period of \((1/\lambda)\). Then the second
pallet is depleted and two pallet positions are wasted for the same amount of time. This waste is
rendered and accumulated until the last pallet is depleted at which \((zx - 1)\) pallet positions are
unoccupied. Total pallets-time wasted in a lane as a result of honeycombing is

\[
\left(\frac{1}{\lambda}\right)(1 + 2 + \cdots + (zx - 1)) \tag{3}
\]

Multiplying (3) by approximate number of lanes gives

\[
W_H \approx \left(\frac{1}{\lambda}\right)\left(\frac{Q(zx - 1)}{2}\right) \tag{4}
\]

\(W_A\) is calculated by computing total time that lanes are occupied and require accessibility. Herein, in accordance with assumption (2) in section (3), the last lane, which is most likely a
partially occupied lane, is depleted first and the remaining lanes are depleted in an arbitrary order.
The lane that is depleted at the last remains occupied until the whole batch is gone. So, this lane
remains occupied for \((1/\lambda)(Q)\) time periods. The lane depleted before this lane becomes entirely
empty \((1/\lambda)(zx)\) time periods before the latter one. Thus, it remains occupied for \((1/\lambda)(Q - zx)\)
time periods. This process is applied to the remaining lanes, and the total time that all lanes are
occupied is

\[
\left(\frac{1}{\lambda}\right)(Q + (Q - zx) + (Q - 2zx) + \cdots + (Q - Kzx)) \tag{5}
\]

Each aisle is shared between two lanes located on both sides of it, so half of an aisle volume is
dedicated to each lane. It is equal to \((az/2)\) pallets. This follows that

\[
W_A \approx \left(\frac{1}{\lambda}\right)\left(Q(K + 1) - \left(\frac{K(K + 1)}{2}\right)zx\right)\left(\frac{az}{2}\right) \tag{6}
\]

\(O_T\) is given by

\[
O_T \approx \left(\frac{1}{\lambda}\right)(Q + (Q - 1) + \cdots + 1) \\
\approx \left(\frac{1}{\lambda}\right)\left(\frac{Q(Q + 1)}{2}\right) \tag{7}
\]

and subsequently, space utilization is obtained by

\[
U \approx \frac{O_T}{O_T + W_A + W_H} \\
\approx \frac{2x(Q + 1)}{(2x + a)(Q + zx)} \tag{8}
\]

Finally, the average waste of storage space is obtained by summing (4) and (6) and dividing the
result by \(T\) which is \((Q/\lambda)\). That is,

\[
W \approx \left(\frac{1}{4zx}\right)(Qa - 2x + zx(2x + a)) \tag{9}
\]
Proposition 1. The optimal lane depth to block-stack a SKU whose batch of $Q$ pallets is instantaneously stored in a warehouse is

$$x^* \approx \sqrt{\frac{Qa}{2z}} \quad (10)$$

Proof. The optimal lane depth is obtained by taking the derivative of (9) with respect to $x$, set it equal to zero and solve for $x$.

From a practical point of view, the optimal lane depth should be an integer. To obtain an integer lane depth, the two nearest integers to $x^*$ are obtained by rounding $x^*$ up and down and then evaluating (9) at these values. This method can be used to obtain integer solutions for all further propositions.

3.1.2 Common optimal lane depth for multiple SKUs

$W_U$ needs to be computed for the multiple SKUs model. If the clear height of a warehouse in units of pallets of a particular SKU is denoted by $e$, then $(e - z)x$ is the total pallet positions wasted on top of a lane (stack) for the period of time that the lane is occupied by that SKU. The total time that all lanes are occupied is given by (5). It follows that

$$W_U \approx \left( \frac{1}{\lambda} \right) \left( Q(K + 1) - \frac{K(K + 1)}{2} z x \right) (e - z) x \quad (11)$$

Taking the clear height of the warehouse into account changes the space that a lane requires for accessibility to $(ae/2)$. So, expression (6) is updated accordingly.

Denote the height of a pallet of SKU $i$ by $h_i$. Different SKUs may have different pallet heights and consequently be stackable to different heights. To take this into account, all waste expressions are scaled by multiplying to a factor of $h_i$. Total waste of storage space for each SKU is obtained by aggregating all three types of waste. Denote the least common multiple of the inventory cycle times of all SKUs by $T_{\text{LCM}}$. Since SKUs have different inventory cycle times, $W_c$ is determined by calculating the total waste that all SKUs generate in $T_{\text{LCM}}$ and then dividing the result by $T_{\text{LCM}}$. The number of inventory turns of SKU $i$ in $T_{\text{LCM}}$ is $T_{\text{LCM}}(\lambda_i/Q_i)$, and multiplying it by the total waste that SKU $i$ generates in its inventory cycle time gives the total waste that SKU $i$ generates in the $T_{\text{LCM}}$. Summing these wastes for all SKUs and dividing the result by the $T_{\text{LCM}}$ terms to be canceled out from the expression. Therefore from a mathematical point of view, $W_c$ is obtained by summing $W$s for all SKUs.

$$W_c \approx \sum_{i=1}^{n} \left( \frac{h_i}{4z_i x} \right) (Q_i e_i (2x + a) + z_i x (2e_i x + a e_i - 2Q_i - 2)) \quad (12)$$

$U_c$ is calculated similarly by computing the occupied and wasted space-time in $T_{\text{LCM}}$. That is,

$$U_c \approx \frac{\sum_{i=1}^{n} \left( \frac{\lambda_i}{Q_i} \right) (O_i^T)}{\sum_{i=1}^{n} \left( \frac{\lambda_i}{Q_i} \right) (O_i^T + W_i^A + W_i^H)}$$

$$\approx \frac{2x \sum_{i=1}^{n} h_i (Q_i + 1)}{(2x + a) \sum_{i=1}^{n} e_i h_i \left( \frac{Q_i}{z_i} + x \right)} \quad (13)$$
where $W_H^i$, $W_A^i$ and $O_T^i$ are obtained for SKU $i$ by (4), (6) and (7), respectively. Note that $T_{LCM}$ is canceled out in this expression too.

**Proposition 2.** The optimal common lane depth to block-stack $n$ SKUs whose batches of pallets are instantaneously stored in a warehouse is

$$x_c^* \approx \sqrt{\frac{a \sum_{i=1}^{n} \left( \frac{e_i h_i}{z_i} \right) Q_i}{2 \sum_{i=1}^{n} e_i h_i}}$$ (14)

**Proof.** Differentiating (12) with respect to $x$, setting it equal to zero and solving for $x$ gives the result. 

### 3.2 $P > \lambda$, demand is constant

This is a prevalent case in manufacturing systems where pallets of SKUs are produced at finite rates, stored in the warehouse and depleted at finite rates. Figure 5 shows changes in the inventory level of a SKU in this system. Period $T_1$ is the production phase in which $Q$ pallets of the SKU are produced at the rate $P$ and stored in the warehouse. Since the demand is constant, pallets are depleted at the rate $\lambda$ in this period. To simplify calculations, we assume that lanes are filled at the rate $(P - \lambda)$. Production stops at the end of $T_1$ when the inventory reaches its maximum level. Then, the inventory starts decreasing in $T_2$ at the rate $\lambda$. Since the demand is constant, the inventory cycle time is $(Q/\lambda)$, the maximum on-hand inventory, $H$, is $(Q - \lfloor Q\lambda/P \rfloor)$, and the maximum number of lanes required for storage is $(\lceil H/zx \rceil)$ where $\lfloor x \rfloor$ is the largest integer not greater than $x$. Relaxing the integrality restriction results in the following approximations:

$$H \approx \frac{Q(P - \lambda)}{P}$$ (15)

$$K \approx \frac{Q(P - \lambda)}{Pzx}$$ (16)

#### 3.2.1 Optimal lane depth for a single SKU

$W_H$ is generated in two phases, $T_1$ and $T_2$. First the former is calculated. Once a pallet is stored in an empty lane, $(zx - 1)$ pallet positions are wasted in that lane for $1/(P - \lambda)$ time periods. This is
the time that it takes until the next unoccupied position is filled. Then, \((z x - 2)\) pallet positions are wasted for the same period of time. This waste is generated and accumulated until the last unoccupied pallet position in the lane is filled. So, the total honeycombing waste in a single lane in \(T_1\) is

\[
\left( \frac{1}{P - \lambda} \right) ((z x - 1) + (z x - 2) + \cdots + (z x - (z x - 1)))
\]

(17)

Multiplying (17) by \(K\) gives

\[
W_{H_{T_1}} \approx \left( \frac{z x - 1}{2} \right) \left( \frac{Q}{P} \right)
\]

(18)

\(W_{H_{T_2}}\) is calculated similar to (3). Multiplying (3) by \(K\) results in

\[
W_{H_{T_2}} \approx \left( \frac{z x - 1}{2\lambda} \right) \left( \frac{Q(P - \lambda)}{P} \right)
\]

(19)

\(W_A\) is calculated by approximating the total time that lanes are occupied in \(T_1\) and \(T_2\). Once the first pallet is stored in the first lane in \(T_1\), this lane remains occupied until all \((H - 1)\) positions are filled. Since the inventory increases at the rate \((P - \lambda)\), this lane remains occupied for \((H - 1)/(P - \lambda)\) time periods. The second lane is used when the first one becomes fully occupied. It means \((H - z x)\) pallet positions are remained to be filled. Once the first pallet is stored in the second lane, this lane remains occupied until the remaining \((H - 1 - z x)\) pallet positions are filled. That is, it remains occupied for \((H - 1 - z x)/(P - \lambda)\) time periods. Thus, that the total time that all lanes are occupied in \(T_1\) is

\[
\left( \frac{1}{P - \lambda} \right) ((H - 1) + (H - 1 - z x) + (H - 1 - 2 z x) + \cdots + (H - K z x))
\]

(20)

The total time that lanes are occupied in \(T_2\) is calculated similar to (5) but herein \(H\) pallets are depleted, therefore \(Q\) is substituted with \(H\) in (5). Given that the dedicated aisle space to a lane is \((a z/2)\),

\[
W_A \approx \left( \frac{1}{P - \lambda} + \frac{1}{\lambda} \right) \left( H(K + 1) - z x \left( \frac{K(1 + K)}{2} \right) \right) - \frac{K}{P - \lambda} \left( \frac{a z}{2} \right)
\]

(21)

\(O_T\) is computed in \(T_1\) and \(T_2\) by

\[
O_T \approx \left( \frac{1}{P - \lambda} \right) (1 + 2 + \cdots + (H - 1)) + \left( \frac{1}{\lambda} \right) (H + (H - 1) + \cdots + 1)
\]

(22)

It follows that

\[
U \approx \frac{2x(Q(P - \lambda) + P - 2\lambda)}{(2x + a)(Q(P - \lambda) + P z x - 2\lambda)}
\]

(23)

The average waste of storage space is obtained by accumulating all types of waste and dividing the result by \(T\), which is \((Q/\lambda)\). That is,

\[
W \approx \left( \frac{1}{4Pz} \right) (2Pz(x z - 1) + aP(Q + z x) - a\lambda(Q + 2))
\]

(24)
Figure 6 compares space utilization and the average waste of storage space with respect to lane depth. At $x^*$, the space utilization and the average waste of storage space reach their maximum and minimum values, respectively.

**Proposition 3.** The optimal lane depth to block-stack a SKU whose batch of pallets is produced at the rate $P$ and depleted at the rate $\lambda$, where $P > \lambda$, is

$$x^* \approx \sqrt{\frac{a(Q(P - \lambda) - 2\lambda)}{2zP}} \tag{25}$$

**Proof.** It is proved by taking the derivative of (24) with respect to $x$, setting it equal to zero and solving for $x$. \qed

### 3.2.2 Common optimal lane depth for multiple SKUs

$W_U$ is calculated by multiplying the total time that lanes are occupied in $T_1$ and $T_2$ by the unoccupied volume on top of lanes which is $(e - z)x$ pallets. $W_A$ is computed similar to (21), except that the aisle space that a lane requires for accessibility changes to $(ae/2)$. Scaling all waste expressions by multiplying them by $h_i$, summing them, and dividing the result by $T$ results in $W$ for SKU $i$. $W_c$ is obtained by aggregating the $W$s for all SKUs. That is,

$$W_c \approx \sum_{i=1}^{n} \left( \frac{h_i}{4P_iz_i} \right) (P_i(Q_i(e_i(2x + a) + z_i(x(2e_ix - 2Q_i + ae_i - 2)) - \lambda(Q_i + 2)(2x(e_i - z_i) + ae_i)))$$

$U_c$ is obtained as described in (13). That is,

$$U_c \approx \frac{2\pi \sum_{i=1}^{n} \left( \frac{h_i}{P_i} \right) (Q_i(P_i - \lambda_i) - 2\lambda_i + P_i)}{(2x + a) \sum_{i=1}^{n} e_ih_i(x + \left( \frac{1}{P_iz_i} \right) (Q_i(P_i - \lambda_i) - 2\lambda_i))} \tag{27}$$

**Proposition 4.** The optimal common lane depth to block-stack $n$ SKUs whose batches of pallets are produced at the rate $P$ and depleted at the rate $\lambda$, where $P > \lambda$, is

$$x^*_c \approx \sqrt{\frac{a \sum_{i=1}^{n} \left( \frac{e_ih_i}{z_iP_i} \right) (Q_i(P_i - \lambda_i) - 2\lambda_i)}{2 \sum_{i=1}^{n} e_ih_i}} \tag{28}$$

11
Figure 7. Changes in the inventory of a SKU stored at the rate $P$ and depleted at the rate $\lambda$, where $\lambda > P$

**Proof.** Differentiating (26) with respect to $x$, setting the result equal to zero and solving for $x$, proves the proposition.

### 3.3 $P < \lambda$, demand is not constant

In this case, the demand rate is higher than the production rate, but it is not constant. So, the production strategy shall be make-to-stock in order to catch up with the demand. We assume that the demand is known and the production is started soon enough to make sufficient stock in order to catch up with the demand. Figure 7 shows changes in the inventory of a SKU in this system. In period $T_1$, the SKU is produced and stored in the warehouse. No depletion occurs in this period and lanes are filled at the rate $P$. The depletion starts in period $T_2$, while the SKU is still produced and stored in the warehouse. In this period, the demand is fulfilled by the stock and production together. Since the demand rate is higher than the production rate, lanes are depleted at the rate $(\lambda - P)$ in $T_2$. An example of this case is seasonal products.

Inventory reaches its maximum level at the end of $T_1$; therefore, $T_1$ is $(H/P)$. The entire production batch is depleted in $T_2$; hence, $T_2$ is $(Q/\lambda)$ and the inventory cycle time, $T$, is equal to $(Q/P)$. The maximum number of lanes required for storage is $([H/zx])$ where $H$ is $(Q - \lfloor QP/\lambda \rfloor)$. Relaxing the integrality restriction results in the following approximations:

$$H \approx \frac{Q(\lambda - P)}{\lambda} \tag{29}$$

$$K \approx \frac{Q(\lambda - P)}{\lambda zx} \tag{30}$$

One should notice that (29) could also be obtained by substituting the values of $T$, $T_1$ and $T_2$ in $T = T_1 + T_2$.

### 3.3.1 Optimal lane depth for a single SKU

$W_H$ in $T_1$ and $T_2$ is calculated by replacing filling rates in (17) with $P$ and depletion rates in (3) with $(\lambda - P)$. $W_A$ and $O_T$ are similarly calculated by updating the new filling and depletion rates in (21) and (22), respectively. It follows that

$$U \approx \frac{2x(Q(\lambda - P) + 2P - \lambda)}{(2x + a)(Q(\lambda - P) + \lambda zx + 2P - 2\lambda)} \tag{31}$$
\[ W \approx \frac{1}{4\lambda x} (2\lambda x (zx - 1) + a\lambda (Q + zx - 2) - aP(Q - 2)) \] (32)

**Proposition 5.** The optimal lane depth to block-stack a SKU whose batch of pallets is produced at the rate \( P \) and depleted at the rate \( \lambda \), where \( P < \lambda \), is

\[ x^* \approx \sqrt{\frac{a(Q - 2)(\lambda - P)}{2z\lambda}} \] (33)

**Proof.** It is proven by taking the derivative of (32) with respect to \( x \), setting it equal to zero and solving for \( x \). \( \Box \)

3.3.2 Common optimal lane depth for multiple SKUs

\( W_U \) is calculated as described in section (3.2.2) using the new filling and depletion rates. \( W_A \) is computed as described for the single SKU model. Herein, the aisle space that a lane requires for accessibility changes to \( (ez/2) \). To take the pallet heights into account, all waste expressions are scaled by multiplying by \( h_i \). Summing the \( W \)s for all SKUs results in the average waste of storage space for a common lane depth.

\[ W_c \approx \sum_{i=1}^{n} \left( \frac{h_i}{4\lambda_i z_i x} \right) (\lambda_i(e_i(Q_i - 2)(2x + a) + z_i(x(2e_i x + ae_i - 2Q_i + 2)) - P_i(Q_i - 2)(2x(e_i - z_i) + ae_i)) \] (34)

It follows that

\[ U_c \approx \frac{2x \sum_{i=1}^{n} \left( \frac{h_i}{z_i} \right) (Q_i(\lambda_i - P_i) + 2P_i - \lambda_i)}{(2x + a) \sum_{i=1}^{n} e_i h_i (x + \left( \frac{1}{\lambda_i z_i} \right) (Q_i(\lambda_i - P_i) + 2P_i - 2\lambda_i))} \] (35)

**Proposition 6.** The optimal common lane depth to block-stack \( n \) SKUs whose batches of pallets are produced at the rate \( P \) and depleted at the rate \( \lambda \), where \( P < \lambda \), is

\[ x_c^* \approx \sqrt{\frac{a \sum_{i=1}^{n} \left( \frac{e_i h_i}{z_i \lambda_i} \right) (Q_i - 2)(\lambda_i - P_i)}{2 \sum_{i=1}^{n} e_i h_i}} \] (36)

**Proof.** Differentiating (34) with respect to \( x \), setting the result equal to zero and solving for \( x \), proves it. \( \Box \)

4. Experimental framework

The experimental framework is designed as follows: First, we describe the simulation model used to evaluate performance of the proposed models. Next, the test problem sets are described and then accuracy of the proposed models is evaluated by the simulation model. Finally, the finite and infinite production rate models are compared with respect to the optimal lane depth and space utilization.
Algorithm 1 Pseudo-code of the simulation

```plaintext
Initialize parameters \((P_i, \lambda_i, Q_i, z_i, x, a, n)\)
Schedule first events for all SKUs
while (simulation time not terminated) do
  Find the earliest event
  if (the event is production (arrival)) then
    if (a partially occupied lane is available for the SKU) then
      Assign the SKU to the lane
    else
      Dedicated a new lane to the SKU
  else
    if (a partially occupied lane is available for the SKU) then
      Deplete the SKU from the lane
    else
      Deplete the SKU from a fully occupied lane
  Update the lane occupancies
  Update \(W_H, W_A, W_U\) and \(O_T\)
  Schedule the next event for the SKU
return \(W_c\) and \(U_c\)
```

Table 2. Parameters of the triangular distributions used in the simulation

<table>
<thead>
<tr>
<th>Variables</th>
<th>Low</th>
<th>Mode</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production time</td>
<td>(\frac{1}{p}(1 - 0.3))</td>
<td>(\frac{1}{p})</td>
<td>(\frac{1}{p}(1 + 0.3))</td>
</tr>
<tr>
<td>Demand inter-arrival time</td>
<td>(\frac{1}{\lambda}(1 - 0.5))</td>
<td>(\frac{1}{\lambda})</td>
<td>(\frac{1}{\lambda}(1 + 0.5))</td>
</tr>
<tr>
<td>Batch size</td>
<td>(Q(1 - 0.3))</td>
<td>(Q)</td>
<td>(Q(1 + 0.3))</td>
</tr>
</tbody>
</table>

4.1 Simulation model

The pseudo-code of the simulation model used for evaluation (an event-oriented model written in Python) is shown in Algorithm 1. The main goal of the experimental analysis is to evaluate accuracy of the proposed models for the real world situations where stochastic variations exist among the major production factors and demand. For this reason, the simulation model utilizes random variables for the production times, demand inter-arrival times, and the batch sizes as presented in Table 2.

To compute the performance metrics under the stochastic variations, the simulation model is run for lane depths from 5 to 50 pallets deep and replicated 40 times for each lane depth. The average waste of storage space is computed for each replication and the average of results within the replications is recorded for each lane depth. Finally the lane depth that generated the minimum average waste of space is reported as the optimal lane depth. Common random numbers were used for all lane depths in the same replication to ensure that randomness does not interfere in selecting the optimal lane depth. However, the number of replications must be determined such that the confidence intervals on the average waste of space obtained for any two consecutive lane depths do not overlap. Our experiments showed that 40 replications sufficiently narrows the confidence intervals such that this condition is met. Also, they showed that the space utilization converges faster when the warm-up period is set to 10 percent of the simulation period.

Our objective is to evaluate the long time performance of the system, therefore the simulation period must be defined long enough to cover sufficient numbers of inventory cycles for each SKU in the test problems. Considering the randomly generated test problems, the inventory cycle times in
our test problems vary from less than a month to higher than 6 months. Thus, we set the simulation period to five years for both single and the multiple SKU test problems.

4.2 Analyzing accuracy of the models

The performance of the models was analyzed on randomly generated test problems for single and multiple SKUs. The test problems were designed as described in the following:

- **Single SKU:** A repository of 1000 SKUs were randomly generated for each of the two finite production rate models. Uniform random distributions with the parameters shown in Table 3 were used to generate the SKUs. The following are the non-random parameters used in generating the SKUs:
  - monthly holding cost: \((\text{pallet cost}) \times 0.3/12\)
  - setup cost: \((\text{pallet cost}) \times 5\)
  - warehouse clear height: 25 feet
  - \(Q\): computed by the EOQ model (Nahmias 2005)

  The SKU repository for the infinite production rate model was created by duplicating the SKU repository generated for the finite production rate with \(P > \lambda\) and discarding the \(\lambda\)s.

  The simulation model was run for the 1000 SKUs in each repository one by one and the results were compared with the outputs of the relevant models.

- **Multiple SKUs:** For each of the three cases investigated in this paper, three test sets were designed for 10, 50 and 100 SKU test problems. Each test set consists of 30 different test problems whose SKUs were randomly chosen from the relevant SKU repository. For all multiple SKU test problems, aisle width was set to three pallets.

  The accuracy of the proposed models in estimating the optimal lane depth and space utilization was evaluated by calculating the Mean Absolute Percentage Error (MAPE) between the model estimations and the simulation results. The MAPE for the optimal lane depth is calculated as

\[
MAPE_{x^*} = \frac{1}{N} \sum_{i=1}^{N} \frac{|x^S_i - x^*_i|}{x^S_i} \tag{37}
\]

where \(x^S_i\) and \(x^*_i\) are the optimal lane depth obtained by the simulation and the proposed models for the \(i\)th test problem, respectively, and \(N\) is the number of the test problems, which is equal to 1000 and 30 for the single and multiple SKU test sets, respectively. The MAPE for space utilization
Figure 8. The MAPE for the optimal lane depth and space utilization, $P = \infty$

Figure 9. The MAPE for the optimal lane depth and space utilization, $P > \lambda$

Figure 10. The MAPE for the optimal lane depth and space utilization, $P < \lambda$

is obtained by

$$MAPE_U = \frac{1}{(N \times 45)} \sum_{i=1}^{N} \sum_{j=5}^{50} \left| \frac{\bar{U}_{ij} - U_{ij}}{\bar{U}_{ij}} \right|$$  \hspace{1cm} (38)

where $\bar{U}_{ij}$ is the average space utilization within the 40 replications of simulation for the lane depth $j$ in test problem $i$ and $U_{ij}$ is the space utilization estimated by the relevant model for the lane depth $j$ in test problem $i$. Figures 8, 9 and 10 show the $MAPE_x$ and $MAPE_U$ for the three investigated cases. The following observations are obtained from the experimental study:
The \( \text{MAPE}_U \) and \( \text{MAPE}_{x^*} \) are less than 2.6 and 8 percent for all three cases for all four test sets, respectively. This shows that despite the high variations applied to the \( P \), \( \lambda \), and \( Q \) in the simulations, the models accurately estimated both the optimal lane depth and space utilization in all three cases and their performance are robust under the presence of stochastic variations.

- Performance of the models are consistent and do not depend on the size of the problems (number of SKUs). The \( \text{MAPE}_U \) varies less than one percent as the number of SKUs increases from 10 to 100 SKUs. This variation is less than three percent for the \( \text{MAPE}_{x^*} \).
- Single SKU models estimated both performance metrics more accurately than the multiple SKU models, though the difference is relatively small. The \( \text{MAPE}_U \) and \( \text{MAPE}_{x^*} \) are less than 0.5 and 1 percent for these models, respectively.

### 4.3 Finite vs. infinite production rate model

The infinite production rate model is the only existing model in the literature that addresses optimal lane depth in block stacking warehouses (Matson 1982). Employing this model in the warehouses located in manufacturing systems is equivalent to disregarding the production flow to the warehouse and assuming instantaneous SKU arrivals. In this section, we show the disadvantages of disregarding the production rates in such warehouses. We compare the infinite production rate model with the finite model on the same test problems using the simulation model to compute space utilization for the optimal lane depths obtained by the two models. The experimental study in this section is designed as follows:

First, we test the single SKU models for different \( \lambda/P \) ratios. From a mathematical point of
Table 4. Finite vs. infinite production rate models

<table>
<thead>
<tr>
<th>Metric</th>
<th>10 SKUs</th>
<th></th>
<th>50 SKUs</th>
<th></th>
<th>100 SKUs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
<td>Avg.</td>
<td>Min</td>
<td>Max</td>
<td>Avg.</td>
</tr>
<tr>
<td>Improvement achieved in space</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>utilization</td>
<td>0.36%</td>
<td>3.67%</td>
<td>2.09%</td>
<td>0.87%</td>
<td>3.25%</td>
<td>1.94%</td>
</tr>
<tr>
<td>Optimal lane depth (finite</td>
<td>14</td>
<td>21</td>
<td>17.8</td>
<td>15</td>
<td>19</td>
<td>17.60</td>
</tr>
<tr>
<td>model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal lane depth (infinite</td>
<td>24</td>
<td>40</td>
<td>30.00</td>
<td>26</td>
<td>34</td>
<td>29.90</td>
</tr>
<tr>
<td>model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

view, the infinite production rate model is a special case of the finite model where \( P > \lambda \). That is, (10) can be alternatively obtained by substituting \( P = \infty \) in (25). This experiment aims to determine when the infinite model is capable of estimating optimal lane depth for a finite production rate problem with a relatively small error. Then, performance of both models are examined for multiple SKUs with respect to space utilization. Single and multiple SKU test problems are designed for this experiment as described in the following:

- **Single SKU:** 10 different test problems each consisting of 1000 randomly generated SKUs were generated similar to the SKU repositories in section 4.2 except that here, the demand rates were generated such that the \((\lambda/P)\) ratios for all SKUs in the first, second, ..., and the tenth test problems were in the ranges 1-10 percent, 11-20 percent,..., 91-99 percent, respectively.

- **Multiple SKUs:** Two SKU repositories were randomly generated as described in section 4.2. Herein, the demand rates were generated such that \((\lambda/P)\) ratios were between 5 to 30 percent for the SKUs in the first repository and between 70 to 95 percent for the ones in the second repository. Then, three test sets were designed each containing 30 test problems for 10, 50 and 100 SKUs, respectively. For each test problem, 70 percent of its SKUs were randomly chosen from the first SKU repository and the remaining were randomly chosen from the second repository. This setup aims to preserve diversity among the SKUs in the test problems.

Figure 11 and 12 compare the \( MAPE_U \) and \( MAPE_{x^*} \) for the single SKU models for different \((\lambda/P)\) ratios. The infinite production rate model obtained relatively small errors when the ratio is less than 10 percent. As the ratio increases, both the \( MAPE_U \) and the \( MAPE_{x^*} \) drastically increase for the infinite model. On the contrary, the finite model obtained relatively small and consistent errors for all ranges.

Table 4 shows the test results for the multiple SKU test problems. On average, the optimal lane depths obtained by the finite production rate model resulted in about two percent higher space utilization than the ones obtained by the infinite model. This improvement is almost consistent among the three test sets. On the other hand, the infinite model obtained much deeper lanes than the finite model. The average of the optimal lane depths estimated by the infinite model is more than 68 percent deeper than the average of the ones obtained by the finite model. Deep lanes form a warehouse layout that has few cross aisles and therefore, the transportation costs increase in the warehouse. So, the layout designed by the finite model is more flexible and also incurs less transportation costs. However, quantifying this cost is out of the scope of this paper and is an open problem for a future research.

The optimal lane depths obtained by the finite model achieved higher space utilization in all test problems. This improvement increases when the production rates are closer to the demand rates, like just-in-time manufacturing systems where the production lines are in balance with the demand.
5. Conclusion

In this paper, we developed mathematical models to obtain the optimal lane depth for single and multiple SKUs in block stacking storage systems under finite production rate constraints. The following cases were studied: infinite production rate, finite production rate where the production rate is higher than the demand rate, and less than the demand rate. A simulation model was developed to evaluate performance of the models for the real world situation where uncertainty exists in production and demand. That is, the assumptions of constant, deterministic production and demand were relaxed in the simulation. An experimental study was carried out on the randomly generated test problems for single and multiple SKUs. The experimental analysis shows that although the models were developed based on the assumption of deterministic production and demand, they are robust and accurate under uncertainty.

We evaluated the finite production rate model with the only existing model in the literature (infinite production rate model) on the same test problems and highlighted the advantages achieved in the space utilization and transportation costs by taking the production rates into consideration. The optimal lane depth obtained by the finite model led to higher space utilization in all test problems. They are nearly half as deep as the ones obtained by the infinite production rate model. This implies, they form a flexible layout that contains more cross aisles and as a result less transportation costs are incurred to the system.

Our proposed models accurately estimate the lane depth that enhances space utilization in block stacking warehouses in manufacturing systems. However, it is important to note that our model does not consider safety stock or transportation costs which could influence the results in practice. Considering both space utilization and Transportation costs in finding the optimal lane depth seems a substantial problem for future research.

References


