Performance Analysis of Order Fulfillment for Low Demand Items in E-tailing

Pallav Chhaochharia, Stephen C Graves
Massachusetts Institute of Technology

Abstract—We study inventory allocation and order fulfillment policies among warehouses for low-demand SKUs at an online retailer. A large e-tailer strategically stocks inventory for SKUs with low demand. The motivations are to provide a wide range of selections and faster customer fulfillment service. We assume the e-tailer has the technological capability to manage and control the inventory globally: all warehouses act as one to serve the global demand simultaneously. The e-tailer will utilize its entire inventory, regardless of location, to serve demand. Thus, given the global demand and an order fulfillment policy, there are trade-offs involving inventory holding costs, transshipment costs, and backordering costs in determining the optimal system inventory level and allocation of inventory to warehouses. For the case of Poisson demand and constant lead time, we develop methods to approximate the key system performance metrics like transshipment, backorders and average system inventory. We then use these results to develop guidelines for inventory stocking and order fulfillment policies for online retailers.

***Index Terms*—Inventory allocation, order fulfillment, low demand SKU, online retailers.**

I. Introduction

A large e-tailer strategically stocks inventory for SKUs with low demand for several reasons. Xu [1] lists the following reasons.

"One motivation is to provide a wide range of selections, since such SKUs actually constitute a significant portion of the total SKUs. The second incentive, of course, is to provide faster customer fulfillment service. The third motivation is to gain a competitive advantage from other online retailer. Suppose that an e-tailer only drop-ships the low-demand SKUs, its drop-shipper who serves many online retailers, may choose to satisfy a competitor's demand. For many of these SKUs, the e-tailer may only stock a handful of inventory units across all warehouses. Inventory planning for low-demand SKUs is challenging because the discrete effect is much more pronounced while the current inventory models often assume all variables are continuous."

Efficient inventory planning and order fulfillment for low-demand SKUs is important in the retailing setting. Often over 90% of a retailer's catalog comprises slow moving SKUs with demand in the range of 0.2 – 0.8 units per week. Therefore, the impact of inventory planning for low-demand SKUs is very significant

Research by Xu [1], focused on the effect of inventory allocation on outbound transportation costs. Her model envisions an e-tailer with several warehouses in the system, with the technological capability to manage and control the inventory globally: all warehouses act as one to serve the global demand simultaneously. Specifically, the e-tailer will utilize its entire inventory, regardless of location, to serve demand. Given that the e-tailer stocks a certain number of units of inventory in the system, Xu studied how best to allocate inventory to warehouses by considering outbound transportation costs from the warehouses to customers. Their approach produced exact solutions for the 2-unit 2-location case, but was not tractable for the general multi-unit multi-location case.

We extend the work done in [1] by developing methods to calculate metrics like transshipment, backorders and average system inventory for special cases of demand distribution over the locations. We use these exact results to develop an approximation for these performance metrics for the case of a general demand distribution over the locations. We then use these results to develop guidelines for inventory stocking and order fulfillment policies for online retailers.

II. Model

We start with the N-Unit N-Location problem for a single item. Suppose the e-tailer decides to stock exactly one unit at each of N warehouses in the system. We want to find methods to estimate key performance metrics like transshipments, backorders and average system inventory for an N-Unit N-Location problem. We start with the following assumptions.

A-1The system demand process is Poisson with rate λ.

Manuscript received November 8, 2006. This work was supported by the Singapore MIT Alliance.

Stephen C Graves is the Abraham J Siegel Professor of Management Science at the Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: sgraves@ mit.edu).

Pallav Chhaochhria is a doctoral student at the Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: pallavc@ mit.edu).

A-2 The demand process is split into N independent processes, 1 to N. With probability α_1, a demand arrival is from region 1; with probability α_2, a demand arrival is from region 2; and so on. The α_i are non-negative and sum to 1.

A-3 The replenishment lead-time for each warehouse is the same constant L.

A-4 The inventory policy is one-for-one replenishment at each warehouse: a replenishment is triggered at each demand epoch.

A-5 Demand is backlogged when there is no on-hand inventory in the system.

In the context of online retailing, the e-tailer can utilize any warehouse or fulfillment center to serve the customer demand. Specifically, a demand is always served by an on-hand inventory unit in the system if there is any; if there are no on-hand inventory units in the system, the demand is served by and triggers replenishment at the warehouse that has the next arriving replenishment. We then have the following assumptions on how the system operates for all stocking scenarios.

A-6 When a customer arrives and its closest warehouse has on-hand inventory, then its closest warehouse serves the demand and triggers a replenishment.

A-7 When a customer arrives and its closest warehouse does not have inventory on-hand, the system will assign the demand to another warehouse if there is on-hand inventory elsewhere in the system. A warehouse with on-hand unit is chosen according to an order fulfillment policy, *P,* to serve demand; this assignment triggers replenishment for the chosen warehouse.

A-8 If a customer arrives and the system has no on-hand units, then the policy is to assign the demand to the warehouse with the next arriving unassigned replenishment. This assignment triggers another replenishment for the chosen warehouse.

Note that assumption A-8 is possible because we assume deterministic supply lead-times, so we know exactly when all future replenishments arrive. Also, assumption A-7 and A-8 are analogous to an emergency transshipment.

Notation

D_L : Random variable for the system demand over the lead time. $E[D_L]= \lambda L$

SI : Random variable for the on-hand system inventory $SI = (N - D_L)^+$

We define the following probabilistic events:

$F_{i,j}$: Event that an order from region i is filled immediately from on-hand stock at warehouse j

$B_{i,j}$: Event that an order from region i is backordered and filled subsequently from a replenishment to warehouse j

$A_{i,j}$: Event that an order from region i is filled from warehouse j. Hence, $\boldsymbol{A_{i,j} = F_{i,j} \cup B_{i,j}}$

The system performance metrics such as the fill rate and average inventory can be calculated for general cases if D_L is known.

System fill rate = $1 - \Pr[D_L \geq N]$

Average system inventory = $\sum_{i=1}^{N} i \times \Pr[SI = i]$

However, the amount of transshipment depends on the demand distribution across the regions and the order fulfillment policy in the system. Sections III and IV estimate the transshipment for special cases of balanced demand and extreme demand distribution for a specific order fulfillment policy. We specify a service failure metric and a method to estimate transshipment for general demand distribution in Section V. In Section VI, we compare the performance of various order fulfillment policies. We discuss future research directions and conclusions in Section VII.

III. BALANCED DEMAND CASE

We first analyze the case in which each of the N-warehouses faces a demand rate of λ/N from its local region. We consider the order fulfillment policy as stated in A-6 and A-7, with the feature that if a customer arrives and its closest warehouse does not have inventory on-hand, but one or more of the other warehouses do have inventory on hand, then a warehouse with on-hand unit is randomly chosen (with equal probability) to fill the order. We call this policy P_1.

Result 1:
The probability that an order from region i is filled from warehouse j immediately from stock under the above condition is given by:

$$\Pr\left[F_{i,j}\right]=\sum_{k=1}^{N}\Pr\left[F_{i,j}\mid SI=k\right]\times\Pr\left[SI=k\right]$$

where $\Pr\left[SI=k\right]=\Pr\left[D_L=N-k\right]=\dfrac{e^{-\lambda L}\left(\lambda L\right)^{N-k}}{N-k!}$

$$\Pr\left[F_{i,j}\mid SI=k\right]=\frac{\binom{k-1}{N-1}}{\binom{k}{N}}=\frac{k}{N}\quad for\ i=j$$

$$\Pr\left[F_{i,j}\mid SI=k\right]=\left(1-\frac{k}{N}\right)\frac{1}{N-1}\quad for\ i\neq j$$

The probability that an order from region i is backordered and filled from a replenishment to warehouse j under the above condition is given by:

$$\Pr\left[B_{i,j}\right]=\Pr\left[B_{i,j}\mid SI=0\right]\times\Pr\left[SI=0\right]$$

where $\Pr\left[SI=0\right]=\Pr\left[D_L\geq N\right]$

and $\Pr\left[B_{i,j}\mid SI=0\right]=\dfrac{1}{N}$

The results follow from application of the Total Probability Theorem, and properties of the Poisson process. The key insight used here is that since demand arrival is Poisson with equal rates for each region, for a given level of system inventory, each inventory state is equally likely. For example, if N=3 and SI=2, then the inventory states (1,1,0), (1,0,1) and (0,1,1) are equally likely to occur. Hence, conditioned on SI=2, we can argue that the probability that a demand from region 1 is served from warehouse 1 is $\Pr\left[\boldsymbol{F}_{1,1}\mid \boldsymbol{SI}=2\right]=2/3$, as this will happen for inventory states (1,1,0) and (1,0,1). Similarly, conditioned on SI=2, a demand from region 1 is served by warehouse 2 or 3 only if the inventory state is (0,1,1), where each warehouse has an equal probability; thus, we have $\Pr\left[\boldsymbol{F}_{1,2}\mid \boldsymbol{SI}=2\right]=\Pr\left[\boldsymbol{F}_{1,3}\mid \boldsymbol{SI}=2\right]=1/6$.

A customer order is back-ordered if and only if none of the warehouses in the system has inventory, i.e, SI=0. Again, since demand is Poisson with equal rates for each region, when the system inventory is 0, each warehouse is equally likely to have the next arriving unassigned replenishment unit. Thus:

$$\Pr\left[B_{i,j}\mid SI=0\right]=\frac{1}{N}$$

With the above results, we can find the probability a demand in region i is served by a transshipment from warehouse j: $\Pr[F_{i,j}]+\Pr[B_{i,j}]$ for $i\neq j$.

IV. EXTREME DEMAND CASE

We now suppose that all demand originates from one region, e.g., $\alpha_1=1$*, while* $\alpha_j=0$ *for j=2,..N*

We now analyze the case where one of the N-warehouses faces a demand rate of λ from its local region, while the other warehouses do not face any demand but still carry inventory. The order fulfillment policy for this analysis is P_1, as described in Section III.

Consider the demand arrival process with $\alpha_1=1$. We define a renewal as occurring whenever an order is filled by warehouse 1 either immediately from stock or as a backorder. We define the inter-renewal interval (M_t) as the number of demands that occur between renewal epochs. Then the counting process that looks at the number of orders served by warehouse 1, is a renewal process, and M_t are IID RVs for renewals occurring at t.

Result 2:
The probability that an order from region 1 is filled from warehouse 1 immediately from stock under the above condition is given by:

$$\Pr\left[F_{1,1}\right]=\Pr\left[F_{1,1}\mid A_{1,1}\right]\times\Pr\left[A_{1,1}\right]$$

where

$$\Pr\left[F_{1,1}\mid A_{1,1}\right]=\Pr\left[D_L<N\right]$$

$$\Pr\left[A_{1,1}\right]=\frac{1}{1+E\left[M\right]}$$

$$E\left[M\right]=\sum_{k=0}^{N-1}k\times\Pr\left[D_L=k\right]+\left(N-1\right)\Pr\left[D_L\geq N\right]$$

The probability that an order from region 1 is backordered and filled from a replenishment to warehouse 1 under the above condition is given by:

$$\Pr\left[B_{1,1}\right]=\Pr\left[B_{1,1}\mid A_{1,1}\right]\times\Pr\left[A_{1,1}\right]$$

where

$$\Pr\left[B_{1,1}\mid A_{1,1}\right]=1-\Pr\left[F_{1,1}\mid A_{1,1}\right]=1-\Pr\left[D_L<N\right]$$

$$\Pr\left[A_{1,1}\right]=\frac{1}{1+E\left[M\right]}$$

$$E\left[M\right]=\sum_{k=0}^{N-1}k\times\Pr\left[D_L=k\right]+\left(N-1\right)\Pr\left[D_L\geq N\right]$$

The results follow from application of the Total Probability Theorem and Renewal-Reward Theory [2].
Define a reward function, R(n) = 1 if order n is served by warehouse 1. Then, by the Key Renewal Theorem,

$$\lim_{n\to\infty} E[R(t)] = \frac{E[R(n)]}{\overline{X}}$$

where $\overline{X}$ is the average inter-renewal interval, $\overline{X} = 1 + E(M)$

However, E[R(t)] is the expected rate of reward accumulation, which in this model, is the probability of an order being assigned to warehouse 1 to be fulfilled either immediately from stock or from a replenishment unit when it arrives. Hence, the result follows.

For the other warehouses, we can show that *for* $j \neq 1$:

$$\Pr\left[F_{1,j}\right] = \frac{\Pr\left[D_L < N\right] - \Pr\left[F_{1,1}\right]}{N-1}$$

$$\Pr\left[B_{1,j}\right] = \frac{\left(1 - \Pr\left[D_L < N\right]\right) - \Pr\left[B_{1,1}\right]}{N-1}$$

As explanation, we note that the fill rate from the non-local warehouses to serve demand in region 1 equals the system fill rate, net of the fill rate from warehouse 1; by symmetry, we then divide the total fill rate associated with the non-local warehouses equally across these N-1 warehouses. Similarly we find the probability that a non-local warehouse serves a backordered demand from region 1.

With the above results, we can find the probability a demand in region i is served by a transshipment from warehouse j: $\Pr[F_{i,j}]+\Pr[B_{i,j}]$ for $i \neq j$.

V. Interpolation Method

For other cases of demand, exact solutions could not be found either by similar methods or using the method in [1]. We expect that $\Pr[F_{i,i}]$ to be monotonically decreasing with α_i for a given system demand rate, λ (provided that demand in all other regions remains proportional, and the order fulfillment policy in place is P_1). Thus, we propose to approximate the $\Pr[F_{i,i}]$ for other cases of demand using some form of monotonic interpolation Using the known results for $\alpha_i = 1$ and $\alpha_i = 1/N$, we considered a linear interpolation and exponential interpolation approximation for $\Pr[F_{i,i}]$ using the values of $\Pr[F_{i,i}]$ for the balanced and extreme demand distribution cases. We compared this approximation with Monte-Carlo simulation results for $\Pr[F_{i,i}]$. A sample case for a 4-unit 4-warehouse scenario with λ=1 and L=3 is shown below:

TABLE I: Comparison of Simulation and Interpolation Results

α1 - Region 1 orders filled from WH1 stock	*1.0*	*0.80*	*0.60*	*0.50*	*0.40*	*0.30*	*0.25*	*0.20*	*0.10*	*0.05*
Simulation	0.196	0.220	0.250	0.272	0.291	0.316	0.328	0.339	0.370	0.376
Linear approx	0.195	0.231	0.267	0.285	0.303	0.321	0.330	0.339	0.357	0.366
Exp approx	0.195	0.224	0.258	0.277	0.297	0.318	0.330	0.342	0.367	0.380

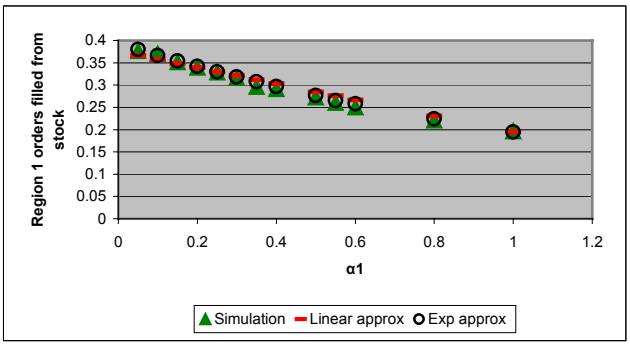

Fig 1: Comparison of Simulation and Interpolation Results

We observe that the exponential approximation performs slightly better than the linear approximation. In general, the approximations are both reasonably good, within 5% of the simulation results. Furthermore, the error seems systematic with the approximation overestimating $\Pr[F_{i,i}]$ when $\alpha_i \in [1/N, 1]$ and underestimating $\Pr[F_{i,i}]$ when $\alpha_i \in [0, 1/N]$.

However, this approximation method does not account for the effects due to the demand distribution across the warehouses. For instance, consider a 3-unit 3-location scenario. If $\alpha_1 = 0.33, \alpha_2 = 0.67, \alpha_3 = 0$, then the $\Pr\left[\boldsymbol{F}_{1,1} \mid \alpha_1 = 0.33, \alpha_2 = 0.67, \alpha_3 = 0\right]$ is clearly not equal to $\Pr\left[\boldsymbol{F}_{1,1} \mid \alpha_1 = 0.33, \alpha_2 = 0.33, \alpha_3 = 0.33\right]$. We expect that this approximation performs bests when local demand faced at the other warehouses is equal. If demand at other warehouses is not equal, then under order fulfillment policy P_1, the approximation overestimates the probability of local order fulfillment from stock.

We define another performance metric, Service Failure, as the probability that an order is not filled immediately by its local warehouse.

Service Failure for region i, $SF_i = 1 - \Pr[F_{i,i}]$

$$\text{Service Failure for system, } SF \cong 1 - \sum_{i=1}^{N} \alpha_i \times \Pr[F_{i,i}]$$

We can estimate the Service Failure for the system quite accurately using the above formula despite the errors in estimating $\Pr[F_{i,i}]$. This is due to the cancellation of the systematic errors in the approximation of $\Pr[F_{i,i}]$ as some $\alpha_i \in [1/N, 1]$ while other $\alpha_i \in [0, 1/N]$.

We approximate the probability of a backorder filled by its local warehouse, $\Pr[B_{i,i}]$, as being almost equal for each warehouse in the system, then:

$$\Pr[B_{i,i}] \approx \frac{1}{N}\Pr[D_L \geq N]$$

Thus, we can estimate the probability of transshipment for each region and for the system as:

$$\boldsymbol{TS}_i = \boldsymbol{SF}_i - \Pr[\boldsymbol{B}_{i,i}]$$

$$\boldsymbol{TS} = \sum_{i=1}^{N} \alpha_i \times \boldsymbol{TS}_i$$

We compared these estimates of Service Failure and Transshipment for the system with Monte-Carlo simulation results obtained over different Fill Rates, system configurations and demand distribution spreads. We find a generally good fit with errors below 5% for the most part.

TABLE II: SERVICE FAILURE APPROXIMATION ERROR PERCENTAGE (RELATIVE TO SIMULATION RESULTS)

	Demand Distribution Spread Low			Demand Distribution Spread Medium			Demand Distribution Spread High		
Fill Rate	3U-3L	4U-4L	5U-5L	3U-3L	4U-4L	5U-5L	3U-3L	4U-4L	5U-5L
99%	0.14	-0.08	-0.36	-0.06	-0.95	-2.12	-0.70	-2.63	-5.18
96%	0.14	-0.06	-0.21	0.02	-0.75	-1.63	-0.56	-2.18	-4.04
90%	0.19	0.01	-0.16	0.08	-0.51	-1.16	-0.38	-1.58	-2.94
80%	0.16	0.04	-0.07	0.12	-0.30	-0.73	-0.27	-1.17	-1.93
70%	0.14	0.06	-0.01	0.10	-0.20	-0.47	-0.15	-0.71	-1.30

Statistics taken over 100 simulation runs of 500k orders

TABLE III: TRANSSHIPMENT APPROXIMATION ERROR PERCENTAGE (RELATIVE TO SIMULATION RESULTS)

	Demand Distribution Spread Low			Demand Distribution Spread Medium			Demand Distribution Spread High		
Fill Rate	3U-3L	4U-4L	5U-5L	3U-3L	4U-4L	5U-5L	3U-3L	4U-4L	5U-5L
99%	0.02	-0.12	-0.37	-0.16	-0.95	-2.10	-0.70	-2.59	-5.13
96%	-0.02	-0.14	-0.25	-0.09	-0.76	-1.61	-0.51	-2.09	-3.94
90%	0.00	-0.08	-0.20	-0.05	-0.52	-1.15	-0.30	-1.47	-2.82
80%	-0.01	-0.06	-0.11	0.01	-0.33	-0.71	-0.16	-1.06	-1.79
70%	0.00	-0.03	-0.05	-0.01	-0.20	-0.46	-0.07	-0.60	-1.18

Statistics taken over 100 simulation runs of 500k orders

VI. ORDER FULFILLMENT POLICIES

We have only considered the order fulfillment policy, P_1, in the previous cases due to its tractability. However, we find that other order fulfillment policies can perform better than P_1 in terms of reducing the Transshipments in the system, without affecting the Average System Inventory and Backorders.

We consider two other types of Order Fulfillment policies that have the same rules as stated in A-6 and A-7, but now with a different fulfillment policy when a customer arrives and its closest warehouse does not have inventory on-hand. When one or more of the other warehouses do have inventory on hand, then a warehouse with an on-hand unit is chosen by a certain rule to fill the order. In the first type of policy, say P_2, the warehouse is randomly chosen but with higher probabilities for warehouses facing lower local demand rates. In the second type of policy, say P_3, the warehouse is chosen from a priority list that orders the warehouses according to their local demand rates, with lower local demand having a higher priority. Results from Monte-Carlo simulation indicate that policy P_3 has the best system-wide performance in terms of total transshipments but results in more variability in the local performance measures. Overall, there was not a significant difference in system performance between these policies.

VII. CONCLUSION

We have extended the work done in [1] by developing methods to calculate performance metrics like transshipment, backorders and average system inventory for special cases of the demand distribution across the locations. We have also developed approximations for these metrics in the case of general demand distribution across the locations for N-units N-locations. Comparing these approximations with Monte-Carlo simulation results indicate that these are good estimates. We are currently using these performance metrics to develop guidelines for inventory stocking and order fulfillment policies for online retailers, i.e, given a certain system demand rate and lead time, we intend to develop guidelines for the optimal inventory holding configuration across the warehouses and the optimal order fulfillment policy to service demand.

REFERENCES

[1] P. Xu, “Order Fulfillment in Online Retailing: What goes where,” *Doctoral Dissertation at MIT*, pp. 89–125, September 2005.

[2] R. G. Gallager, "Discrete Stochastic Processes," pp. 57-81, Kluwer Academic Publishers, 1996