Base-Stock Control, Kanban Control, and a Network of Inventory Queues

(Extended Abstract)

David D. Yao
Dept. of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027, USA
<yao@ieor.columbia.edu>

Abstract

We survey several production-inventory control mechanisms and their relations with queueing models, both of finite and infinite buffer capacities. These relations suggest natural ways to adapt the decomposition techniques of queueing networks to analyze networks of inventory queues.

1 Base-Stock Control

A widely studied inventory control mechanism is the *base-stock control*. It works as follows: Whenever the *inventory position* — defined as on-hand inventory minus backorder plus outstanding orders (orders that were placed but have not yet arrived) — drops to below R, a constant parameter, place an order to bring the inventory position back to R. This way, every demand will trigger the placement of an order, such that the overall inventory position is always maintained at a constant level, R, which is called the *base-stock level*.

The base-stock control model is known to relate to a queueing model as follows. Let each unit of demand correspond to a job arrival to the queue. Each unit of demand, on arrival, is supplied by the on-hand inventory, or, in the stock-out situation, joins a back-order queue. Regardless, however, a replenishment/production order is triggered.

Let the outstanding orders correspond to the jobs in the queueing system. In a pure inventory system, each of these orders will materialize (or, "arrive" — become available for supplying demand) after a lead time. This corresponds to a queueing system with an infinite number of servers, so that any job will be served immediately on arrival; and hence, the overall cycle time of each job in the system is simply its service time, which corresponds to the lead time of orders in the inventory system.

Let N be the number of jobs in the queueing system (in steady state). This, as explained above, represents the number of outstanding orders. If $N < R$, then there are $R - N$ units of on-hand inventory. If $N > R$, then we know there are $N - R$ units of backorders: R units of demand have been supplied with on-hand inventory, while the remaining $N - R$ units are backordered. Hence, with I and B denoting the on-hand inventory and the backorders, respectively, and with $[x]^+$ denoting $\max\{x, 0\}$, we have,

$$I = [R - N]^+, \qquad B = [N - R]^+. \tag{1}$$

Note the above implies

$$I - B = R - N,$$

and $I \cdot B = 0$, i.e., I and B cannot be both positive.

Suppose demand follows a Poisson process with rate λ, and suppose the lead time is L — the time it takes to process and finish an order. Then, the queueing system in question is an $M/G/\infty$ model; and it is known that N follows a Poission distribution with mean $\rho := \lambda \mathsf{E}(L)$. That is,

$$\mathsf{P}[N = n] = \frac{\rho^n}{n!} e^{-\rho}, \qquad n = 0, 1, 2, ...$$

Combining the above with (1), we can derive the distributions of the on-hand inventory and backorders:

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = 1 - \sum_{n=0}^{R-1} \frac{\rho^n}{n!} e^{-\rho},$$
$$\mathsf{P}[I = n] = \mathsf{P}[N = R - n] = \frac{\rho^{R-n}}{(R-n)!} e^{-\rho}, \quad n = 1, ..., R;$$

and

$$\mathsf{P}[B = 0] = \mathsf{P}[N \leq R] = \sum_{n=0}^{R} \frac{\rho^n}{n!} e^{-\rho},$$

$$\mathsf{P}[B=n]=\mathsf{P}[N=R+n]=\frac{\rho^{R+n}}{(R+n)!}e^{-\rho}, \quad n=1,2,....$$

We can also approximate the Poisson distribution by a normal distribution, and write

$$N=\rho+Z\sqrt{\rho}, \tag{2}$$

where Z denotes the standard normal variate, with density function $\phi(x)=(1/\sqrt{2\pi})e^{-x^2/2}$, and distribution function $\Phi(x)$. Below we shall also use $\bar{\Phi}(x):=1-\Phi(x)$. We can also write

$$R=\rho+k\sqrt{\rho}, \tag{3}$$

where k is known as the "safety factor" in the inventory literature. Then, we have

$$\mathsf{P}[I=0]=\mathsf{P}[N\geq R]=\mathsf{P}[Z\geq k]=\bar{\Phi}(k),, \tag{4}$$

which is the usual stockout probability in the inventory literature. Incorporating (2) and (3) into (1), we can also derive the expectations of I and B as follows:

$$\begin{aligned}\mathsf{E}(I) &= \mathsf{E}[k-Z]^+ \\ &= \int_{-\infty}^{k}(k-z)\phi(z)dz \\ &= k\Phi(k)+\phi(k), \end{aligned} \tag{5}$$

and similarly,

$$\begin{aligned}\mathsf{E}(B) &= \mathsf{E}[Z-k]^+ \\ &= \int_{k}^{\infty}(z-k)\phi(z)dz \\ &= \phi(k)-k\bar{\Phi}(k). \end{aligned} \tag{6}$$

While these are familiar expressions in the inventory literature, from the above we observe that they are rooted in the $M/G/\infty$ queue model.

Next, suppose, due to limited production capacity, replenishment/production orders may have to wait in queue before they can be processed. In other words, in addition to the lead time (i.e., processing time, or, "service time"), there is also queueing time. In this case, a queueing model with a finite number of servers is more appropriate. Suppose we use the M/M/1 model for the corresponding queueing system. Then, N follows a geometric distribution:

$$\mathsf{P}[N = n] = \rho^n (1 - \rho), \qquad n = 0, 1, 2, ...$$

The distributions of I and B in the case are:

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = \rho^R,$$
$$\mathsf{P}[I = n] = \mathsf{P}[N = R - n] = \rho^{R-n}(1 - \rho), \quad n = 1, ..., R;$$

and

$$\mathsf{P}[B = 0] = \mathsf{P}[N \leq R] = 1 - \rho^{R+1},$$
$$\mathsf{P}[B = n] = \mathsf{P}[N = R + n] = \rho^{R+n}(1 - \rho), \quad n = 1, 2, ...$$

Note from (1) that in either case while the on-hand inventory is limited to $I \leq R$, the base-stock level, there is no upper limit on the number of backorders, B.

2 Kanban Control

We now turn to the kanban control. This is a control rule that limits the number of on-hand (or, finished goods) inventory to K units, with K a constant parameter. Specifically, there are K "cards" in the system; each card is attached to a job that is in waiting or in process, or to a job that is completed (waiting for demand). In other words, a card is needed in order to admit an order into the system. When all K cards are exhausted, there are three possibilities:

(i) If all K cards are attached to completed jobs, then production is suspended.

(ii) If all K cards are attached to jobs in waiting or in process (i.e., there is no completed job), then any further arrival (demand) will be blocked, i.e., no more jobs are admitted into the system.

(iii) The situation in between is that some of the K cards are attached to completed jobs while others to jobs in waiting or in process. Then, any arriving demand will be supplied by one of the completed jobs, with its card detached and given to another new job (representing the outstanding order).

This clearly relates to the base-stock control mechanism. In particular, $K = R$ — any unit of demand will trigger production (or, replenishment), whereas when the finished goods inventory reaches K, production will be suspended. Kanban, however, has the additional feature of blocking arrivals (of demand) when the on-hand inventory drops down to zero, i.e., when all K cards are associated with outstanding orders, which is the situation in (ii) above. Hence, kanban corresponds to a *finite* queueing system with K being the buffer capacity — the upper limit on the total number of jobs allowed in the system.

With this in mind, in particular, with $R = K$, the models in the last section should be modified to $M/G/\infty/K$ and $M/M/1/K$ queues. In terms of the distributions of N, I and B, this amounts to a renormalization: dividing those probabilities in the last section by $\mathsf{P}[N \geq K]$. The inventory and backorder expressions in (1) remain valid, with $R = K$. For instance, in the $M/G/\infty/K$ case, we have

$$\mathsf{P}[I=0] = \mathsf{P}[N=K] = \frac{\rho^K}{K!}e^{-\rho}\left[\sum_{n=0}^{K}\frac{\rho^n}{n!}e^{-\rho}\right]^{-1},$$

$$\mathsf{P}[I=n] = \frac{\rho^{R-n}}{(R-n)!}e^{-\rho}\left[\sum_{n=0}^{K}\frac{\rho^n}{n!}e^{-\rho}\right]^{-1}, \quad n = 1, ..., R;$$

and

$$\mathsf{P}[B=0] = 1,$$

since $N \leq K$.

3 Generalized Kanban Control

Note that in the above kanban control mechanism, the cards control both admission (of demand arrivals) into the system and the upper limit of finished goods inventory. A more general kanban control mechanism is one that uses two types of cards. In addition to the usual type, which now (only) controls admission, there is a second type of cards controlling finished goods inventory. Let K and R, with $K \geq R$, denote, respectively, the number of cards for these two types, referred to, respectively, as kanbans and "production cards". Specifically, every job within the system has a kanban attached to it. In addition, to be processed by the server, a job also requires a production card. Both the kanban and the production card stay with the completed job until it supplies a demand. At that point, the released kanban admits a new order into the system, while the released production card authorizes the service (production) of another waiting order.

(Note that the above kanban control with two types of cards can be shown to be equivalent to the kanban control with three types of cards, the so-called (a, b, k) model in [3].)

This way, the total number of jobs in the system is limited to K, while the total number of *completed* jobs (that are waiting for demand) is limited to R. Since now $R \leq K$, backorders are allowed, up to a limit of $K - R$.

Therefore, what is needed for this more general kanban control is the $M/G/\infty/K$ and $M/M/1/K$ models, in connection with the relations in (1) — with $R \leq K$. The earlier distributions for I and B can be modified accordingly. For instance, in the $M/M/1/K$ case, we have

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = \frac{\rho^R - \rho^{K+1}}{1 - \rho^{K+1}},$$

$$\mathsf{P}[I = n] = \mathsf{P}[N = R - n] = \frac{\rho^{R-n}(1 - \rho)}{1 - \rho^{K+1}}, \quad n = 1, ..., R;$$

and

$$\mathsf{P}[B = 0] = \mathsf{P}[N \leq R] = \frac{1 - \rho^{R+1}}{1 - \rho^{K+1}},$$

$$\mathsf{P}[B = n] = \mathsf{P}[N = R + n] = \frac{\rho^{R+n}(1 - \rho)}{1 - \rho^{K+1}}, \quad n = 1, ..., K - R.$$

The associated expectations can be derived as follows:

$$\mathsf{E}[I] = \frac{R - (R + 1)\rho + \rho^{R+1}}{(1 - \rho)(1 - \rho^{K+1})},$$

$$\mathsf{E}[B]=\frac{\rho^{R+1}-(K-R+1)\rho^{K+1}+(K-R)\rho^{K+2}}{(1-\rho)(1-\rho^{K+1})}.$$

4 Network of Inventory Queues

Now consider an inventory/distribution network, what's known as a supply chain. Each node in the network represents a stocking location. Suppose a base-stock control policy is followed at each node. With the discussion above, we can adapt the standard decomposition approach in analyzing queueing networks to study this inventory network.

To be specific, focus on a particular node, j, in the network. Suppose node j has a single upstream node i, which supplies any replenishment orders from j. Since base-stock control is followed throughout the network, each node is driven directly by the external demand process, which we continue to assume to be a Poisson process with rate λ.

Suppose j is modeled as an $M/G/\infty$ queue. Then, the only difference from the analysis in §1 is that the service time (leadtime), L_j, needs to be prolonged, if there is a stockout at node i. Hence, the modified service time

$$\tilde{L}_j = L_j + \tau_i, \quad \text{w.p.} \quad \bar{\Phi}(k_i), \tag{7}$$

while $\tilde{L}_j = L_j$ w.p. $\Phi(k_i)$. Here, τ_i is the extra delay at node i, until the next job (outstanding order) is completed. Following the analysis in [2], we can approximate τ_i as follows, provided node i is also modeled as an $M/G/\infty$ queue:

$$\tau_i = \frac{L_i \mathsf{E}(B_i)}{R_i \bar{\Phi}(k_i)}. \tag{8}$$

Next, suppose both i and j are modeled as single-server queues, say $M/M/1$. Then, the only occasion in which its service time needs to be prolonged is when j has a stockout and server i is forced to become idle. Obviously, the probability is dominated by the stockout probability at node j, which is equal to $\rho_j^{R_j}$. In the single-server queue model, stability requires $\rho_j < 1$, hence, this probability is quite negligible when R_j is reasonably large and ρ_j is not too close to 1. In this case, we can simply forgo the adjustment. Otherwise, an adjustment similar to the one in (7) is to add $\tau_i = L_i$ to L_j

w.p. $\rho_j^{R_j}$. (However, this might result in overcompensation, in particular, the associated probability, $\rho_j^{R_j}$, could be too high.)

Suppose node j interacts directly with external customer demand, and follows a generalized kanban control as outlined in §3. Then, a proportion of the demand will be blocked at lost when all K_j cards are occupied. This is equal to the probability $\mathsf{P}[N_j = K_j]$, where N_j follows the $M/G/\infty/K_j$ or the $M/G/1/K_j$ model, with the adjusted service time $\tilde{L}_j$ as discussed above. Note that in this case, we have a combination of backorder and lost-sales for external demand.

In this type of decomposition, each decomposed node is driven by exactly the same external demand process, due to the base-stock control mechanism. This is quite different from the usual queueing network decomposition, where it is standard to assume that the decomposed queues are independent of each other, which is indeed the case in the special case or "product-form" networks, and hence, by extrapolation, the dependence is quite weak in the non-product-form case. In the network of inventory queues, the dependence among the queues should be "maximal", in the sense that they are all driven by the same arrival process. This, of course, has no bearing upon analyzing each queue in isolation (i.e., the marginal distribution).

To the extent that all nodes are driven by the same demand arrival process, this type of networks behave more like the assembly-to-order system ([5]), where each unit of demands consists of a set of components; and hence, a demand arrival will triger simultaneously the production at all component queues. The difference, of course, is that in the assembly-to-order system, the response time to any order is the *max* of all the component response times; whereas in a tandem of inventory queues discussed above, the order response time is the *summation* of the response times at all queues.

The simple, Poisson demand arrival process assumed here can be readily extended to more involved processes, e.g., those that involve batches, like the model in [2], which makes use of the queueing results in [4]. The same applies to more general leadtime distributions and multiple servers; approximate queueing models can be adapted; refer to [1].

Finally, optimization models can be formulated based on these approximate inventory-queue models, [2] being one such example. As the basic relations in (1) serve as building blocks of the network, one should be able to establish structural properties of the objective function based on properties of the functions in (1), which are convex in R and submodular in (R, N). These are useful properties in identifying efficient solution algorithms.

References

[1] Buzacott, J.A. and Shanthikumar, J.G., *Stochastic Models of Manufacturing Systems,* Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[2] Ettl, M., Feigin, G., Lin, G.Y., and Yao, D.D., A Supply Network Model with Base-Stock Control and Service Requirements. *Operations Research*, to appear.

[3] Glasserman, P. and Yao, D.D., A GSMP Framework for the Analysis of Production Lines, in: *Stochastic Modeling and Analysis of Manufacturing Systems* (Chapter 4), D.D. Yao (ed.), Springer-Verlag, New York, 1994.

[4] Liu, L., Kashyap, B.R.K., and Templeton, J.G.C., On the $GI^X/G/\infty$ System. *Journal of Applied Probability*, **27** (1990), 671 – 683.

[5] Song, J., On the Order Fill Rates in a Multi-Item, Base-Stock Inventory System. *Operations Research*, to appear.