DETERMINING THE SPARES AND STAFFING LEVEL
FOR A REPAIR DEPOT

by

Stephen C. Graves*

WP #1699-85 September, 1985

* A. P. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139

Introduction

In this paper we consider a problem that arose in the design of a repair depot for the field service organization for an office-automation product. The problem was to determine the staffing level and the inventory level necessary at the depot to implement a particular service strategy. We report here on a linear systems model that was developed to address this question. In the first section we introduce the problem by discussing the product and the repair task, and the proposed service strategy. In the second section we present and analyze a model of the operation of the repair depot. Then we illustrate the use of the model with an example. In the final section we give a short conclusion and summary of our work.

Problem Description

The problem was to determine the resource requirements for the repair depot for a new product. This product (unit) was like a terminal and consisted of a 'box' that contained three electronic modules (components). Most field failures of the product were the result of one or more failures of the electronic modules. The failed product would be returned to the repair depot, where its repair usually entailed 'swapping' electronic modules. That is, a technician would replace the product's components, one at a time, with good ones until the failed module(s) was found. Once the failed module(s) is found, it is replaced with a good module(s) from an inventory of spares. The failed modules could then be repaired (also at the depot), and, when repaired, would be used for the diagnosis and repair of future failed units. To ensure that the product had been properly repaired, it was then subject to an overnight burn-in.

The proposed service strategy was to operate the repair depot so that it could guarantee a two-day turnaround on all failed units. A customer with a failed product would ship its unit to the repair depot, presumably via an overnight express delivery service. Within two days of receipt, the repair depot would return it to the customer, also by an overnight service. Thus, at most, the customer would be without the failed unit for four days. For instance, if a unit fails on Monday, it could be shipped that night to arrive at the repair depot on Tuesday morning. The repair depot would repair the unit so that it would be ready to ship by Thursday morning at the latest. The customer would then have the unit by Friday morning. The manufacturer felt that this would be acceptable service for this product, provided that it was extremely reliable.

This service strategy has several features of note. No inventory of spare units is needed. This is because each customer gets its unit back. This was important since there were expected to be several versions of the product (due to anticipated engineering changes), as well as some customized features to the units. However, inventories for the components are needed, since it is not possible to delay the repair of a unit to wait for the repair of its failed component. To ship a unit within two days requires repair on the day of arrival and overnight burn-in, so as to have a second chance if the unit fails during the burn-in. Thus, a unit that arrives on Tuesday morning, will be repaired that day and then put into burn-in. If it is not ready to ship on Wednesday morning (i.e., it fails during burn-in), then it will be repaired again on Wednesday and put into burn-in so as to be ready to ship on Thursday morning. We assume that the likelihood of twice failing during burn-in is remote.

The problem is to determine the staffing level and component inventory level at the repair depot. The work force is needed for two activities: unit repair and component repair. The work force is to be cross-trained so that a particular technician can do both unit and component repair. Also, it is necessary that the work force be flexible in that it can, within reason, work extra hours (overtime) as needed to satisfy the turnaround requirement. The size and flexibility of the work force depend on the size of the component inventory, which acts as a buffer. When the arrival of failed units runs high, most of the work force will be devoted to repairing these units (rather than working on component repair), provided sufficient component spares are available. The component inventory will be drawn down, and the queue of failed components waiting for repair will grow. When the arrival of failed units slows down, the work force can be shifted back to component repair to restore the inventory of spare components. If component spares are not sufficient, then a greater portion of the work force will need to do component repair when the arrival of failed units runs high. As a consequence, the total work load will be more variable, i.e. higher when arrivals are high, lower when arrivals are low. The cost of the work force depends on the variability of the work load; the more variable is the work load, a larger work force and/or more overtime is needed. Thus, the key tradeoff is to balance the inventory investment in spare components with the cost of the work force.

Model Description

We developed a discrete-time, linear-systems model of the operation of the repair depot. We take the time period to be one day, and assume that all days are alike. All failed units arrive at the start of a day and will be repaired during that day. Failed components that are discovered during the repair of these units, are sent immediately to component repair. However, they may or may not be repaired during that day. We assume that at the earliest, these failed components are repaired and available for unit repair by the start of the following day. Hence, at the start of every day there needs to be sufficient spare components available to repair all of the units that have arrived that day. Indeed, we set the spare component level so that this is ensured with high probability. Figure 1 denotes schematically the flows, repair activities, and inventories for this operation.

For ease of presentation, we will assume that the repair times for both units and components are deterministic. For the same reason, we will also ignore the work load from rework generated by failures during burn-in. Nevertheless, the model can permit both stochastic repair times and the inclusion of rework.

The model has two parts. The first part sets the repair levels for both unit and component repair by a production smoothing rule. We assume here that ample component spares are available, and then can determine the work force cost. The second part determines the level of component spares necessary to support the first part. By modifying the parameters of the production smoothing rule, we can obtain a range of solutions that will show the tradeoff between work force cost and inventory cost.

PRODUCTION SMOOTHING

To present the production smoothing model we need first define the arrival process for failed units. We will define this in terms of the implied work loads, rather than the number of failed units, as follows:

$A(1,t)$: The work load, measured in man-hours, in unit repair from the arrival of failed units at the start of day t.

$A(2,t)$: The work load, measured in man-hours, in component repair generated by the arrival of failed units at the start of day t.

We assume that both $A(1,t)$ and $A(2,t)$ are i.i.d. processes, over time, and that their sum, $A(t) = A(1,t) + A(2,t)$ is an i.i.d. random variable. However, $A(1,t)$ and $A(2,t)$ are not independent random variables, but will be highly correlated since the failed components are generated by the failed units. Our model permits these processes to be correlated. We will assume that $A(t)$, $A(1,t)$, and $A(2,t)$ are normally-distributed random variables.

The decision variables for the model are the production levels for unit repair and component repair, defined as follows:

$P(1,t)$: The production level, measured in man-hours, in unit repair during day t.

$P(2,t)$: The production level, measured in man-hours, in component repair during day t.

These production levels are set as a function of the total work load (backlog) in unit repair and in component repair, where we define the total work load as follows:

$B(1,t)$: The backlog of work, measured in man-hours, in unit repair just after the arrival of failed units at the start of day t.

$B(2,t)$: The backlog of work, measured in man-hours, in component repair just after the arrival of failed units at the start of day t.

Now we can relate the backlog to the arrival process and the production decision variables by the following balance equation:

(1) $$B(i,t) = B(i,t-1) + A(i,t) - P(i,t-1) \qquad \text{for } i = 1,2.$$

To specify the model we need to set the control rule that determines the production levels in each period. To satisfy the two-day service requirement, we need set the unit repair level equal to the work backlog:

(2) $$P(1,t) = B(1,t) .$$

From (1) and (2), we see that $B(1,t)=A(1,t)$ and thus, $P(1,t) = A(1,t)$. As required, we set the production level to repair all units on their day of arrival.

We have assumed here that ample components are available to perform the unit repairs. In fact, we will use the model to set the level of component spares so that sufficient components are available most of the time (e.g. with probability .95). We assume that on the days when the component spares are inadequate, that the unit repairs will be completed either by expediting the repair of stocked-out components or by 'borrowing' the needed components from elsewhere (e.g. the primary production activity). Thus, the service level that is planned for the component inventory should reflect how frequently the depot manager is willing to resort to an expediting and/or 'borrowing' mode.

We set the component repair level by smoothing the total production level $P(t)$, where $P(t) = P(1,t) + P(2,t)$. Since the work force is cross-trained, $P(t)$ reflects the work load on day t. We set $P(t)$ by

(3) $$P(t) = \min[\, B(t), K + (B(t)/n)\,],$$

where $B(t) = B(1,t) + B(2,t)$ is the total backlog of work, and K and n are nonnegative control parameters with $n \geq 1$. First, the total production level can not exceed the backlog. Second, in an attempt to smooth the work load, we set the production level equal to the sum of a constant term (K) and a term that is proportional to the backlog $(B(t)/n)$. We chose this control rule due to its simplicity, due to the fact that it permits some analysis of the model, and due to its correspondence to actual practice. As the backlog increases, so will the production level, where the parameters K and n determine the exact response. From (2) and (3) we obtain the component production level $P(2,t)$.

To get some analytic insight into the performance of this production smoothing rule, we consider the following simpler rule:

(4) $$P(t) = K + (B(t)/n) .$$

That is, we ignore the restriction that $P(t)$ cannot exceed $B(t)$. We will then set K and n so that the probability that $P(t)$ exceeds $B(t)$ is small (less than .01). Hereafter, we use (4) as an approximation to (3).

We can now use (4) and the balance equation for $B(t)$ that corresponds to (1), to find the simple smoothing equation:

$$P(t) = A(t)/n + (n-1)P(t-1)/n .$$

For large t, we can reexpress this smoothing equation by recursive substitution as

$$P(t) = \sum_{i=0}^{\infty} (1/n)\,[(n-1)/n]^i\, A(t-i) , \tag{5}$$

where we have assumed an infinite history. From (5) we obtain

$$E[P(t)] = E[A(t)] , \tag{6}$$

$$Var[P(t)] = Var[A(t)]/(2n-1), \tag{7}$$

where E[ ] denotes the expectation, and Var[ ] denotes the variance. Furthermore, since we assume that $A(t)$ is a normal random variable, then $P(t)$ is also a normal random variable with mean and variance given by (6)-(7). We note that only the control parameter n (which we term the smoothing parameter) affects the moments of $P(t)$. The other control parameter K will only influence the spares' level, as will be seen.

The above derivation is virtually identical to that for a comparable rule in Cruickshanks et al. (1984), to which we refer the interested reader.

From (6)-(7) we can determine the size of the work force and its cost, including overtime. For instance, if the work force is k people that work a eight-hour day, then the daily work force cost is the regular time cost for k people plus the expected overtime cost, given by

$$- \int_{8k}^{\infty} (x-8k)\, f(x)\, dx$$

where f(x) is the probability density function for P(t), the daily production level in man-hours. Thus, for A(t) a normal random variable, f(x) is the normal density with mean and variance from (6)-(7). For a given smoothing parameter n, one can easily find the size of the work force k that minimizes the total work force cost.

DETERMINATION OF SPARES' LEVEL

To determine the level of spare components, we will characterize the backlog at the start of the day in component repair, B(2,t). We then convert this backlog (measured in man-hours) into the number and type of components in repair or waiting for repair. From this we can determine the level of component spares that is necessary to provide acceptable service.

From (4) and (5) we can write the total backlog as

$$B(t) = \sum_{i=0}^{\infty} [(n-1)/n]^i A(t-i) \quad - \quad nK .$$

Since B(1,t) = A(1,t), we can use the above expression to find the component backlog to be

$$B(2,t) = A(2,t) + \sum_{i=1}^{\infty} [(n-1)/n]^i A(t-i) \quad - \quad nK . \tag{8}$$

From (8) we obtain

$$E[B(2,t)] = E[A(2,t)] + (n-1)E[A(t)] \ - nK , \tag{9}$$

$$Var[B(2,t)] = Var[A(2,t)] + [(n-1)^2/(2n-1)]Var[A(t)] . \tag{10}$$

If A(2,t) and A(t) are normal random variables, then so is B(2,t) with its moments given above. From (8)-(10) we see the influence of the parameter K; the larger is the parameter K, the smaller is the expected component backlog. However, we recall that the derivation of (8) is based on the control rule (4), which is an approximation to (3). We conjecture that in order for (4) to be a good approximation to (3), we need choose K so that the likelihood that B(2,t) is negative is small. The assumption of the normality of B(2,t) and (9)-(10) provide sufficient information to do this.

We also see from (9)-(10) the influence of the smoothing parameter n. The variance of the component backlog increases linearly with n. As a consequence, provided we choose K as indicated above, the expected backlog will also increase with n, albeit at a slower rate.

We use (9)-(10) to characterize the number and type of components in repair. For instance, suppose that the repair time for each type of component is the same and equal to r hours. Then for N(t) equal to the total number of components in repair at time t, we have that N(t) = B(2,t)/r, and thus

(11) $E[N(t)] = E[B(2,t)]/r$,

(12) $Var[N(t)] = Var[B(2,t)]/(r^2)$.

Now, for a given value for N(t), we assume that the number of failed components of type j at time t is binomial with parameters N(t) and p_j, where p_j is the probability that a failed component is of type j. Then, for N(j,t) being the number of components of type j in repair at t, we find that

(13) $E[N(j,t)] = p_j\, E[N(t)]$,

(14) $Var[N(j,t)] = p_j(1-p_j)\, E[N(t)] + p_j^2\, Var[N(t)]$.

Although the derivation of (13)-(14) assumes that the component repair time is deterministic and the same over all types of components, we can get comparable results for stochastic and varying repair times.

To use (13)-(14) we must relate N(j,t) to the stockage level for component j, call it s_j. We observe that at any point in time for type j the number of components in repair plus the number of available components for unit repair must equal s_j. This follows since each failed component from unit repair is repairable and is swapped for a good component from the spares inventory. Since N(j,t) is the number of components in repair after accounting for the failed components from the unit arrivals on day t, then $s_j - N(j,t)$ is the number of available components of type j leftover at the start of day t. Thus, a nonnegative value for $s_j - N(j,t)$ signifies that sufficient components were available for that day's unit repairs. We define the service level from s_j as the probability that $s_j - N(j,t)$ is nonnegative: that is, the probability that sufficient spares are available to meet that day's repair requirements. To assess this probability we need know the discrete probability distribution for N(j,t). We will approximate this distribution by a normal distribution with parameters given by (13) and (14). Alternatively, we could use a two-moment discrete distribution such as a binomial or negative binomial. In this instance, the choice of a normal distribution is due to its easy use: for a desired service level, we set s_j by

$$s_j = E[N(j,t)] + z^*\{Var[N(j,t)]\}^{.5},$$

where z is the number of standard deviations of safety stock necessary to achieve the desired service level.

Example

In this section we present an illustrative example of the use of the model. The parameters for this example are based on on the projected parameters for the repair depot but have been disquised. The moments for the arrival process are as follows:

$$E[A(t)] = 105 \qquad E[A(2,t)] = 72$$

$$Var[A(t)] = 225 \qquad Var[A(2,t)] = 144$$

where the dimensions of the arrival process are man-hours. We assume that there are three distinct components per unit with $p_1 = .3$, $p_2 = .2$, and $p_3 = .5$. The repair time for each component is two hours ($r = 2$). Thus, we expect 36 failed components to enter repair each day, with a standard deviation of 6. We set the control parameter K so that the probability that B(2,t) is negative is less than .01; thus, we expect that the control rule (4), which we analyze, will be a good approximation to the actual control rule (3). For each component we set its spare level s_j to provide a service level of at least .95; that is, each spare will stock out at most once every twenty days.

In Table 1 we give the illustrative results from the model for a range of values for n for this parameter set. We note that the expected work load, E[P(t)], is 15.9 man-days in all cases, where we assume 6.6 hours per man-day. We show in the table how the standard deviation of the work load varies with the smoothing parameter n: as we increase n, the work load becomes smoother. We also report P*, which we define as the 95th percentile of P(t). Thus, we might consider P* as an indicator of the staffing level necessary for each value of n. For instance, we would need 17 people for n=10, 18 people for n=2.5, and 19 people for n=1.25. A more rigorous analysis would set the staffing level to minimize the total work force cost for each value of n, as discussed in the production smoothing section.

Table 1 also gives the level of spares necessary for each choice of the smoothing parameter. We again see the affect of the smoothing parameter. With no smoothing (n=1) we require the minimal number of spares, 53. [To put this in perspective, recall that we assume that failed components cannot be used for unit repair until the day after their arrival at the earliest. Hence, we need enough spares to cover at least one day's requirements, which are 36 components on average.] This requirement increases with n, and more than doubles when we set n=10.

To make the most economic choice for n, we need to balance the work force cost versus the inventory holding cost for the spare components. Without going into a detailed cost analysis, we will try to illustrate the required comparison. For instance, consider the parameter choices n=10, n=2.5, and n=1.25. At n=10, we need a work force of size 17 and an inventory investment in 117 spares (without loss of generality we assume that the three types of components have the same value). We can reduce the inventory to 74 spares by setting n=2.5, but need increase the work force by one since we are doing less smoothing. Hence, this is desirable only if the labor cost of one person is less than the holding cost for 43 spares. Similarly, further reduction in the smoothing (to n=1.25) is warranted only if the savings from reducing the inventory level from 74 to 58 exceeds the labor cost of an additional person. In this manner, one can find the best choice for the smoothing parameter for a given service level.

Conclusion

In this paper we have considered the problem of setting the work force level and the inventory level in a flexible repair facility with a rigid service requirement. We have proposed a linear systems model that highlights the tradeoff between the work load variability and the inventory level. The model shows that as we smooth the work load, which should reduce the work force cost, we require more inventory to ensure satisfactory service. We have illustrated the use of the model with a small example. We note that this model should also be useful for guiding the day-to-day decisions of setting the production levels both in unit repair and in component repair.

As we have noted throughout, the model requires a series of assumptions and approximations. Possibly the most severe are the assumption of the control rule (3) and its approximation (4). It would be useful to contrast this control rule with alternatives. Furthermore, it would be of value to explore under what conditions does (4) provide an adequate approximation to the behavior of (3). Such a comparison has been done for a comparable setting by Cruickshanks et al. (1984). We have assumed that by setting the control parameter K properly, then (4) will be indicative of (3). This is not obvious and should be examined in more detail.

References

Cruickshanks, A. B., R. D. Drescher, and S. C. Graves, "A Study of Production Smoothing in a Job Shop Environment," Management Science, **30**, 3, (March 1984), 368-380.

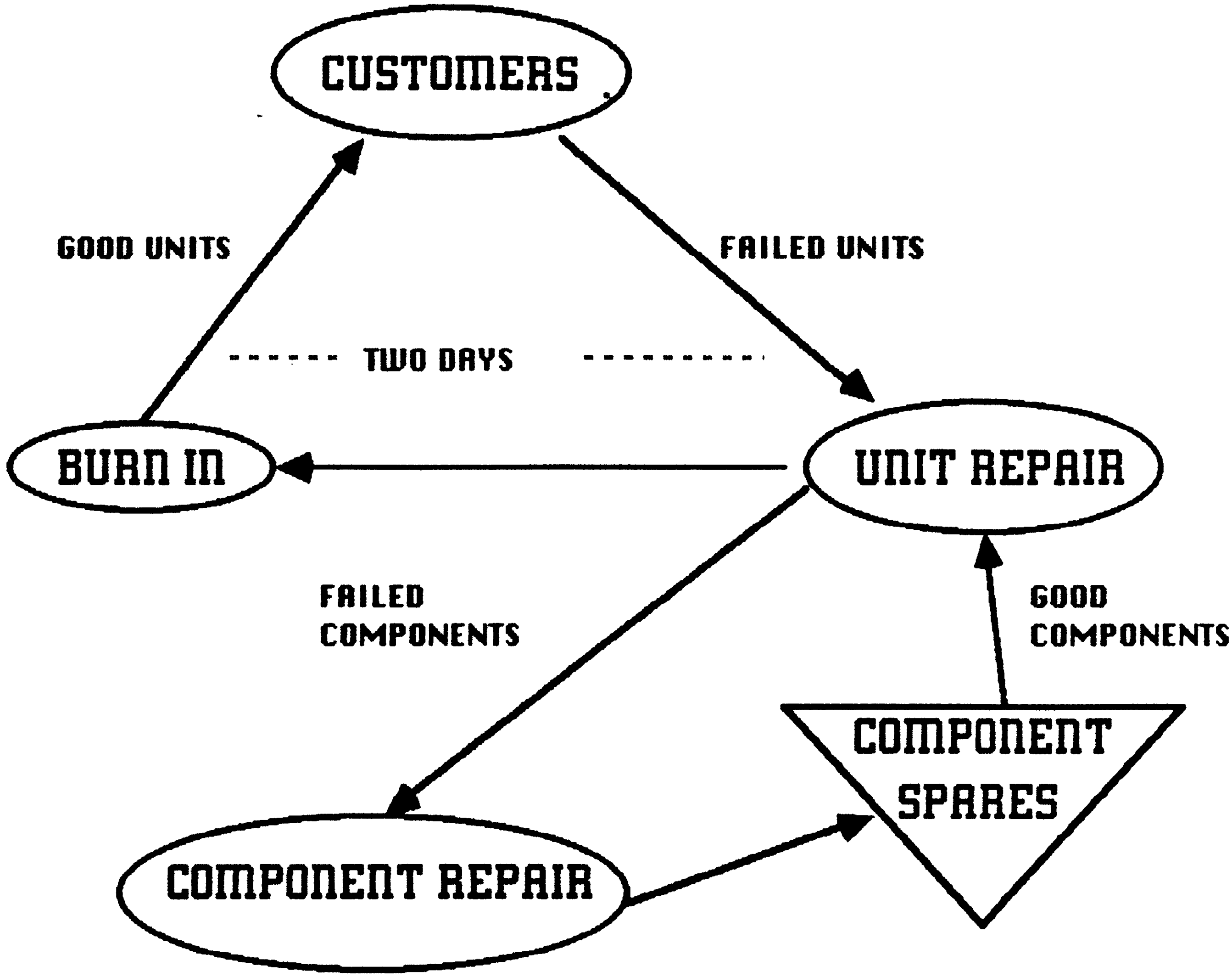

FIGURE 1 : MATERIAL FLOW, REPAIR ACTIVITIES AND INVENTORIES AT REPAIR DEPOT

n	Std. Dev. of P	P*	S(1)	S(2)	S(3)	Total Spares
1	2.3	19.7	16	12	25	53
1.11	2.1	19.3	17	12	26	55
1.25	1.9	19.0	18	13	28	58
1.43	1.7	18.7	19	13	29	61
1.66	1.5	18.4	20	14	31	65
2	1.3	18.1	21	15	33	69
2.5	1.1	17.8	23	16	35	74
3.33	.85	17.5	25	18	38	81
5	.76	17.2	28	20	45	93
10	.52	16.8	35	25	57	117

TABLE 1 : RESULTS FROM EXAMPLE
(P* IS 95-TH % OF P)