ELSEVIER

European Journal of Operational Research 97 (1997) 480–492

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

Theory and Methodology

Models for multi-echelon repairable item inventory systems with limited repair capacity

Angel Díaz [a], Michael C. Fu [b,*]

[a] *Instituto de Estudios Superiores de Administración, av. IESA, San Bernardino, Caracas 1010, Venezuela*

[b] *College of Business and Management, Institute for Systems Research, University of Maryland, College Park, MD 20742, USA*

Received 1 June 1995; accepted 1 August 1996

Abstract

The dominant models for inventory control of repairable items, both in the literature and in practical applications, are based on the assumption of ample repair capacity. This assumption can introduce a serious underestimation of the spare parts requirements in systems with high repair facility utilization, as is typical in industry. In this paper, we introduce approximations that can deal with limited repair facilities, under the scenarios of single-class exponentially distributed repair distributions, single-class general repair distribution, and multi-class general repair distributions. We provide numerical experiments that demonstrate how these models significantly outperform traditional models in the case of high repair facility utilization. Their ease of implementation is illustrated in a case study of the spare parts requirements at the Caracas subway system. © 1997 Elsevier Science B.V.

Keywords: Maintenance; Queueing; Multi-echelon inventory control; Spare parts management

1. Introduction

Inventories represent about one-third of all assets of a typical company. An important type of multi-echelon inventory systems is found in maintenance, where spare parts that can be economically repaired are kept in locations known, using military terms, as bases. When a failure takes place, the defective part is removed, exchanged for a fresh part taken from the base stock and sent to a repair facility, known as the depot, where it is repaired and held in stock to be eventually sent down to the bases to cover another part used in a repair. Parts are thus subject to cycles and not freshly brought from the outside (assuming that all items can be repaired). Variations of this basic model include repairs at both echelons, more than two echelons, lateral transshipment between bases and total failures (a proportion of the failures are non-repairable). To insure continuity of operations, an ample supply of spare parts must be maintained; however, this must be traded off with the cost of tying up capital in non-revenue-generating spare parts inventories. In the context of standard inventory control, stockouts can lead to unavailability of equipment, translating to a loss of service or production capability, whereas spare parts inventories incur holding costs. The magnitude of these economical

* Corresponding author. Email: mfu@umd5.umd.edu.

PII S0377-2217(96)00279-2

implications are such that in the United States military world alone, repairable items represent about $30 billion (O'Malley, 1994). However, these inventories are also of paramount importance for industries or services with heavy utilization of equipment, such as continuous chemical or petrochemical processes and mass transit systems.

We consider multi-echelon models for repairable (or cyclic) items that follow a one-for-one replenishment policy, in which the parts are always ordered in lots of one, so that a part is ordered every time it is used. This considerably simplifies the models and is based on the fact that for low demand, high cost items, the EOQ tends to a size of one. The dominant model for repairable items, both in the literature and in practical applications, is METRIC (Multi-Echelon Technique for Recoverable Item Control), developed by Sherbrooke (1968) and extensively used in the military world. Over the years, some of the constraints imposed on the original METRIC model have been relaxed and versions that deal with compound-Poisson demands, wear-out processes (non-Poisson failures), cannibalism (taking parts from equipment in temporary disuse) and lateral resupply (sending parts from one base to another without going through the depot) have been developed. In addition, other models have been proposed, e.g., Graves (1985) and Axsäter (1993); see Nahmias (1981) and Díaz and Fu (1995) for reviews.

METRIC is predicated on the assumption of ample repair facilities and a large parts population. While these assumptions may be justified in military applications, they seem less appropriate in a resource-constrained environment, as in most industrial settings. Research for the relaxation of these assumptions has been dominated by exact closed queueing network models of limited practical appeal because of prohibitive computational complexity, mainly by Gross and Ince (1978), Gross et al. (1983, 1987, 1993), Albright and Soni (1988), Albright (1989), and Albright and Gupta (1993).

In this paper, we introduce exact and approximate analytical models that relax the assumption of ample repair capacity and lead to expressions of simple implementation. In Section 2, we discuss the basic principles in the general model, discussing METRIC and introducing the ideas of the double convolution analysis and our specific double negative binomial model. We then propose the following models for handling limited repair facilities, based on results and recent approximations from queueing theory:

- Limited repair facilities I: This model applies exact results for exponential interarrival and repair distributions with limited repair facilities.
- Limited repair facilities II: We relax the service time distribution assumption of the previous model by using a two-moment queueing model approximation to handle general repair distributions (exact for one server).
- Limited repair facilities III: The previous model is relaxed even further to allow for different classes of parts in the system, each with a different repair distribution, again using a two-moment queueing approximation.

In Section 3, we provide a case study of the Caracas Metro subway system for the application of the proposed methods. Section 4 provides conclusions and discussion of further research, including a proposal for the utilization of Mean Value Analysis (MVA) as an approximation for systems with a relatively small parts population.

2. The models

2.1. Basic principles

We will introduce the basic principles in the models for repairable items with a simple example. We consider a system consisting of a depot with a target level of s_0 spares at the depot and a target level of s_i spares at a base. This system can be represented as a network of elements, shown in Fig. 1 for the case of a single base:

- a set of working parts at the base "field" (e.g., in the factory or service operation),
- a base-to-repair facility transportation pipeline of failed parts (the in-pipeline),

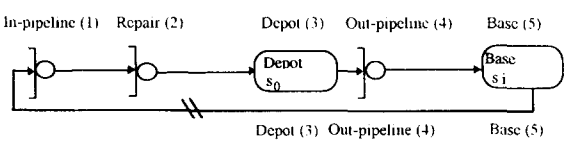

Fig. 1. Schematic representation of the system.

- a repair facility of failed parts to be repaired,
- a depot storage facility stocking spare parts,
- a depot-to-base transportation pipeline of parts (the out-pipeline),
- a base storage facility stocking spare parts.

The repair, failure and transportation processes can be represented as queues – as parts may experience delays in requests for repairs or transportation – and the depot and bases by physical inventories. Let $N_j(t)$ denote the number of parts in location j at time t, where $j=0$ represents the field operations, $j=1$ represents the in-pipeline, $j=2$ represents the repair facility, $j=3$ represents the depot, $j=4$ represents the out-pipeline, and $j=5$ represents the base. We allow $N_j(t)<0$ for $j=0$, $j=3$ and $j=5$, indicating a backordered condition in the field, at the depot, or at the base, respectively, i.e., $N_0(t)$, $N_3(t)$, and $N_5(t)$ can be thought of as the inventory levels in the usual inventory control parlance, which as we shall see are derived from the other three processes. In addition, let $B_0(t)$ and $B_1(t)$ denote the backorder levels at the depot and base, respectively, i.e., $B_0(t)=N_3^-(t)$ and $B_1(t)=N_5^-(t)$, where $x^-=\max(0,-x)$ is the negative part of x. Then, independent of any assumptions on the failure process, the repair process, or the transportation processes, we have the following key relationships:

$$N_1(t)+N_2(t)+N_3(t)=s_0, \tag{1}$$

$$B_0(t)=N_3^-(t)=\left[N_1(t)+N_2(t)-s_0\right]^+, \tag{2}$$

$$N_4(t)+N_5(t)+B_0(t)=s_1, \tag{3}$$

$$B_1(t)=N_5^-(t)=\left[N_4(t)+B_0(t)-s_1\right]^+, \tag{4}$$

$$N_0(t)+B_1(t)=m, \tag{5}$$

where $x^+=\max(0,x)$ is the positive part of x, and m denotes the total nominal working population of parts. METRIC assumes that m and hence $N_0(t)$ is essentially infinite. Under this infinite working parts population assumption the failure rate is a constant value (i.e., independent of the number of working parts), and the system acts essentially like an open network, i.e., we can "cut" the network in Fig. 1 at the indicated place, and ignore Eq. (5).

2.2. The double convolution general model

Eqs. (1)–(5) can be explained as follows (see also Fig. 2), where we drop the time argument on the quantities henceforth, for notational convenience here and later to indicate the corresponding steady-state quantities. Whenever a part is pushed into N_1, another is pulled from the stock at the depot. Thus, by obtaining the distribution for the sum of N_1 and N_2, we can obtain the distribution of outstanding parts at stock at the depot via Eq. (1), from which we can obtain the backorder distribution at the depot $B_0(s_0)$ via Eq. (2). Outstanding parts at the base, N_5, can be obtained from the sum of $B_0(s_0)$ and N_4 via Eq. (3), as whenever a part is pulled from the stock at the base, another is pulled from the depot stock into the out-pipeline, or is backordered if the stock at the depot is zero. From N_5, backorders at the base are obtained as at the depot, via Eq. (4).

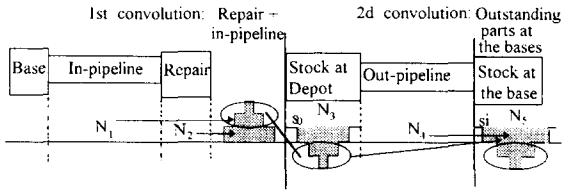

Fig. 2. Double convolution general model.

If mutual independence between all the pipelines and the repair facility processes is assumed, the analysis in essence reduces to solving two convolutions: the first to obtain the distribution of inventory at the depot (N_3) from the distribution of parts in the in-pipeline (N_1) and the distribution of parts being repaired (N_2); and the second to obtain the distribution of inventory at the base (N_5) from the depot backorder distribution derived from the first convolution and the distribution of parts at the out-pipeline (N_4). In the case of multiple bases, the first convolution is obtained from the parts in all in-pipelines, while the second requires decomposing the distribution of backorders at the depot into the bases and then convolving with each base out-pipeline; or alternatively convolving the distribution of backorders with parts in all out-pipelines and then decomposing this into the bases, as in Graves (1985). The key relationships (Eqs. (1)–(5)) are modified to the following:

$$\sum_i N_1^i(t)+N_2(t)+N_3(t)=s_0, \tag{1'}$$

$$B_0(t)=N_3^-(t)=\left[\sum_i N_1^i(t)+N_2(t)-s_0\right]^+, \tag{2'}$$

$$N_4^i(t) + N_5^i(t) + \alpha_i B_0(t) = s_i, \quad (3')$$

$$B_1^i(t) = N_5^{i-}(t) = \left[N_4^i(t) + \alpha_i B_0(t) - s_i\right]^+, \quad (4')$$

$$N_0(t) + \sum_i B_1^i(t) = m, \quad (5')$$

where the superscript i indicates the corresponding base quantity, and α_i is the apportioning of depot backorders to base i. If backorders are dispatched to the bases according to first-come, first-served (i.e., backorders are satisfied in the order in which they were requested), then the apportioning is simply proportional to the demand.

The idea behind METRIC is to calculate both convolutions using Poisson distributions, analytically convenient since the convolution of Poisson distributions is again Poisson. This requires the assumption of a Poisson failure process, an infinite parts population (so that the arrival rate at the depot is constant and independent of the actual number of working parts), and ample repair capacity (so that the distribution of parts in the repair facility is Poisson, independent of the service time distribution if the arrival process is Poisson; see, e.g., Palm, 1938). Under these conditions, $N_1 + N_2$ is Poisson distributed, but B_0 is not. So, furthermore, METRIC assumes that both B_0 and N_4^i are Poisson distributed, so that the second convolution for N_5, needed in calculating backorders at the bases via Eq. (4), is also Poisson. Graves (1985) improves upon the latter assumptions by proposing the fitting of a negative binomial distribution to the second convolution instead, which he showed gave better results of METRIC. A natural question is whether or not the distribution of backorders at the depot obtained under the hypothesis of ample repair capacity is realistic, as the assumption of ample repair facilities is not always the case in industrial environments, and if the distribution of parts waiting to be repaired is important, relative to the distribution of demand in the in-pipelines (bases to depot), the error introduced by the ample server assumption could be considerable. This is intuitively the case when the depot facilities have only one, or few, servers, the intensity of use of the facility is high and the bases are close to the depot, a common industrial scenario. Thus, the models we propose improve upon the first convolution under the case of limited repair facilities.

As in METRIC, we will assume throughout that the failure processes are Poisson. Before we consider the various models, we introduce some common notation to be used for various quantities of interest at the limited repair facility:

- k = number of servers at the repair facilities,
- μ = service rate of each server,
- λ_i = failure rate at base i,
- λ = arrival rate at repair facility (sum of λ_i above),
- ρ = $\lambda/(k\mu)$ = utilization of the repair facility,
- S = service time at the repair facility, $E[S] = 1/\mu$,
- N = number of parts at the repair facility (sometimes also the original notation N_2),
- Q = number of parts in queue at the repair facility,
- W = queue time spend by a part at the repair facility,
- C_X = coefficient of variation for random variable X (e.g., C_S, C_N, C_Q, C_W),
- p_n = probability that there are n parts at the repair facility = $P(N = n)$,
- L_{i0} = in-pipeline time from base i to the depot.

Our general model can be described as a double negative binomial approximation, as we fit a negative binomial distribution to both the first convolution – with moments obtained from modeling the repair facility as a finite server queue and the in-pipelines as M/G/∞ queues – and also to the second convolution, as in Graves (1985). Since we model the in-pipeline as an M/G/∞ queue, the number in repair is independent of the number in the in-pipeline (Mirasol, 1963). For the in-pipelines, we have

$$E[N_1] = V[N_1] = \sum_i \lambda_i L_{i0}.$$

Under first-come, first-served (FCFS) apportioning of backorders, we have $\alpha_i = \lambda_i/\lambda$.

The corresponding general algorithm is as follows:

1. Calculate the expected value and variance of $N_1 + N_2$ by adding the corresponding quantities for the number in repair N_2 and for the in-pipelines N_1^i.

2. Estimate the expected value and variance of backorders B_0 at the depot (dependent on the value of s_0) via Eq. (2′) with the distribution of $N_1 + N_2$ fitted to a negative binomial distribution using the central moments obtained in Step 1.
3. Calculate the expected value and variance of $N_4^i + \alpha_i B_0$ by adding the corresponding quantities for the out-pipelines N_4^i and the depot backorder apportioning $\alpha_i B_0$.
4. Fit the distribution of $N_4^i + \alpha_i B_0$ to a negative binomial distribution using the central moments obtained in Step 3, from which the expected value and variance of backorders B_i at each base (dependent on the value of s_i) can be calculated via Eq. (4′), or alternatively, the fill rates, via $P(N_4^i + \alpha_i B_0 > s_i)$.

2.3. Limited repair facilities I: M/M/k single-class model

For the first model, we assume limited repair facilities, in which service times at the repair facility are exponentially distributed. Standard queueing literature (e.g., Gross and Harris, 1985) gives the following results for an M/M/k queue:

$$E[N_2] = k\rho + p_0\,\rho(k\rho)^k / \left[(1-\rho)^2 k!\right], \tag{6}$$

$$V[N_2] = k\rho\left(1 + \frac{(k\rho)^k}{k!(1-\rho)}p_0\right) + \frac{\frac{\rho(k\rho)^k}{k!(1-\rho)}p_0\left[1+\rho\left(1-\frac{(k\rho)^k}{k!(1-\rho)}p_0\right)\right]}{(1-\rho)^2}, \tag{7}$$

where

$$p_0 = \left[\sum_{n=0}^{m-1}\frac{(k\rho)^n}{n!} + \frac{(k\rho)^k}{k!(1-\rho)}\right]^{-1}. \tag{8}$$

Thus, in the general algorithm, Step 1 is obtained using Eq. (6) and Eq. (7).

For comparison, a set of experiments, adapted from Graves (1985), was designed. There were four bases, taking 10%, 20%, 30% and 40% of total demand, and in- and out-pipelines of length one unit of time each. Six fill rates levels were used to compare results at the bases: 0.84, 0.87, 0.90, 0.93, 0.96 and 0.99 (we used fill rate instead of expected backorders, as did Graves, as results are easier to grasp intuitively). There were four levels of total demand, 0.5, 1, 2 and 4 per unit of time. There was a single server, under four levels of ρ – 0.2, 0.4, 0.6, and 0.8 – to simulate the effect of different degrees of intensity of use of the repair facility, giving equivalent mean repair times varying from 0.05 to 1.6 units of time. For the METRIC case, which doesn't distinguish between repair and in-pipeline times, the repair cycle times were taken as 1 (length of the in-pipeline) plus the repair time. The s_0 were chosen as in Graves, using integer values ranging approximately from the average number in the repair cycle (defined as in METRIC) minus one standard deviation to the average plus two standard deviations (up to a maximum of six values).

Table 1
% cases where the number of parts allocated was incorrect

ρ	0.2	0.4	0.6	0.8
METRIC	5.32	12.70	45.04	91.10
Graves	2.08	10.52	42.26	90.91
Proposed approximation	0.23	0.20	2.18	8.14

The objective of the experiment was to obtain, for each method, the allocation of spares at the bases to attain a given fill rate. The exact solution in this case can be analytically obtained by solving the double convolution. The complete results of the experiment, measured in terms of the proportion of cases where an incorrect allocation of spare parts was made, yielded the results given numerically in Table 1. It is clear from the results that traditional methods make serious errors when the intensity of use of the repair facilities is high ($\rho \geq 0.6$) due to the underestimation of the M/G/∞ approach of the effect of additional waiting time (queueing) at the repair place produced by a fairly large saturation of the server. This can also be seen in Table 2, where the first two moments of the number in the in-pipeline plus the repair place for different values of ρ ($L_{i0} = 1$ for all cases) are calculated. Note in particular the large variance that the waiting in queue produces for conditions of heavy utilization of the repair facilities.

Table 2
First two moments of the first convolution

	Considering in-pipeline separately from repair (1 server), as in the proposed method		Considering in-pipeline and repair (∞ servers) Poisson distributions, as in METRIC	
	E	V	E	V
$\rho = 0.2$	1.25	1.31	1.2	1.2
$\rho = 0.4$	1.66	2.11	1.4	1.4
$\rho = 0.6$	2.5	4.75	1.6	1.6
$\rho = 0.8$	5	21	1.8	1.8

For example, at $\rho = 0.8$, the actual variance was 21, whereas the METRIC model would estimate 1.8.

2.4. Limited repair facilities II: M / G / k single-class model

The use of an exponential distribution to characterize service processes has been criticized by different authors (Suri et al., 1993), on the grounds that coefficients of variation of less than one are usually observed in reality. The model we developed previously can be extended to general distributions by obtaining an approximation for the distribution of parts in repair at the depot in systems with more than one server, or by using exact methods for systems with one server. We do this by extending our idea of fitting a negative binomial distribution to the first convolution. Henceforth the subscript on N_2 will be dropped.

The following approximations for GI/G/k models are based on (Whitt, 1983, 1993). The expected number in the system, $E[N]$, is approximated by

$$E[N] = \lambda\left[\left(\frac{C_a^2 + C_S^2}{2}\right)\left(\frac{p_0}{k\mu}\frac{(k\rho)^k}{k!(1-\rho)^2}\right) + \frac{1}{\mu}\right], \tag{9}$$

where C_a and C_S are the coefficients of variation for the interarrival times and service times, respectively. This approximation is known to work well for the M/G/k case, where $C_a = 1$. The variance of the number in the system, $V[N]$, is obtained from the first two moments via

$$V[N] = E[N^2] - (E[N])^2. \tag{10}$$

We present two approximations. The first approximation simply assumes that the squared coefficient of variation for number in the system is equal to the value for the M/M/k case,

$$E[N^2] \approx E\left[N_{M/M/k}^2\right] E[N]^2 / E\left[N_{M/M/k}\right]^2. \tag{11}$$

For the second approximation, which is considerably more complex, we define another random variable $Q+ = (Q \mid Q > 0)$, the conditional queue length given that the queue is not empty. Its first moment is given by

$$E(Q+) = E[Q]/P(Q>0) = (E[N] - \lambda/\mu)/P(Q>0),$$

where $P(Q>0) \approx \rho P(W>0) = \rho \min\{\pi, 1\}$, $\pi = \rho^2\pi_a + (1-\rho^2)\pi_b$, with π_a and π_b given by

$$\pi_a = \min\left\{1, \frac{1 - \Phi\left(\frac{(1+C_S^2)(1-\rho)\sqrt{k}}{C_a^2 + C_S^2}\right)}{1 - \Phi\left((1-\rho)\sqrt{k}\right)} \times P\left(W_{M/M/k} > 0\right)\right\},$$

$$\pi_b = \min\left\{1, \frac{1 - \Phi\left(\frac{2(1-\rho)\sqrt{k}}{1+C_S^2}\right)}{1 - \Phi\left((1-\rho)\sqrt{k}\right)} \times P\left(W_{M/M/k} > 0\right)\right\},$$

where $\Phi(\cdot)$ represents the c.d.f. for the standard normal distribution. For an M/G/k queue, $\pi_a = P(W_{M/M/k} > 0)$ since $C_a = 1$; for an M/M/1 queue, the result is exact since $P(W_{M/M/1} > 0) = \rho$.

The second moment can be obtained from its coefficient of variation C_{Q+}, given by

$$C_{Q+}^2 = 1/E(Q+) - 1 + (P(Q>0)/P(W>0))(C_D^2 + 1),$$

where C_D, the coefficient of variance of D, the conditional delay given that a server is busy, is given by

$$C_D^2 = 2\rho - 1 + 4(1-\rho)d_s^3/\left[3(C_S^2+1)^2\right]$$

with

$$d_s^3 = \begin{cases} 3C_S^2(1+C_S^2) & \text{if } C_S^2 > 1 \\ (2C_S^2+1)(C_S^2+1) & \text{if } C_S^2 < 1. \end{cases}$$

The second approximation is then given by

$$E[N^2] \approx E\left[N_{M/M/k}^2\right]^2 S_{G/G/k}/S_{M/M/k}, \quad (12)$$

$$E\left[N_{M/M/k}^2\right] = \rho\delta k^2 + 2\rho\delta k(1-\rho)^{-1} + (1+\rho)\rho\delta(1-\rho)^{-2} + (1-\rho\delta)\min\left\{k^2, (\rho\delta)^2 + \rho\delta\right\}, \quad (13)$$

$$S_{G/G/k} = P(Q>0)\left(k^2 + 2kE[Q+] + E[Q+^2]\right) + P(Q=0)\min\left\{k^2, (\rho k)^2 + \rho k z\right\},$$

$$S_{M/M/k} = P(Q_{M/M/k}>0)\left(k^2 + 2kE[Q+] + E[Q+^2]\right) + P(Q_{M/M/k}=0)\min\left\{k^2, (\rho k)^2 + \rho k\right\},$$

where $z = (C_a^2 + C_S^2)/(1 + C_S^2)$ and $\delta = p_0(k\rho)^k \times (1-\rho)^{-2}/k!$. The second approximation (Eq. (12)) yielded slightly better results than the first approximation (Eq. (11)) in our experiments.

For the special case of an M/G/1 queue exact results can be obtained through Pollaczek–Khintchine (PK) formulas (for details on the derivation of the

Table 3
Errors in the estimation of $V[N]$

			PK	M/M/1	Δ (%)	Appr. 1	Δ (%)	Appr. 2	Δ (%)
Erlang 3	$\rho = 0.2$	$E[N]$	0.2333	0.2500	7.08	0.2333	0	0.2333	0
		$V[N]$	0.2551	0.3125	22.5	0.2722	6.7	0.2655	4.08
	$\rho = 0.4$	$E[N]$	0.5778	0.6667	15.30	0.5778	0	0.5778	0
		$V[N]$	0.7417	1.1111	49.8	0.8346	12.5	0.8036	8.35
	$\rho = 0.6$	$E[N]$	1.2000	1.5000	25.29	1.2000	0	1.200	0
		$V[N]$	2.08	3.7500	80.29	2.4000	15.3	2.2257	7
	$\rho = 0.8$	$E[N]$	2.9333	4.000	36.65	2.9333	0	2.9333	0
		$V[N]$	9.594	20.000	108.5	10.755	12.1	9.7695	1.83
Erlang 4	$\rho = 0.2$	$E[N]$	0.2313	0.2500	8.08	0.2313	0	0.2313	0
		$V[N]$	0.2485	0.3125	25.75	0.2674	7.6	0.2600	4.62
	$\rho = 0.4$	$E[N]$	0.5667	0.6667	17.64	0.5667	0	0.5667	0
		$V[N]$	0.7011	1.1111	58.47	0.8028	14.5	0.769	9.62
	$\rho = 0.6$	$E[N]$	1.1625	1.5000	29.03	1.1625	0	1.1625	0
		$V[N]$	1.9064	3.7500	96.70	2.2523	18.1	2.0675	8.45
	$\rho = 0.8$	$E[N]$	2.8	4.000	42.85	2.8	0	2.8	0
		$V[N]$	8.56	20.000	133.6	9.8	14.4	8.7532	2.25
Gamma ($C_S^2 = 2$)	$\rho = 0.2$	$E[N]$	0.2750	0.2500	−9.09	0.2750	0	0.2750	0
		$V[N]$	0.4106	0.3125	−23.8	0.3781	−7.9	0.393	−4.29
	$\rho = 0.4$	$E[N]$	0.8000	0.6667	−16.7	0.8000	0	0.8000	0
		$V[N]$	1.8133	1.1111	−38.7	1.6000	−11.7	1.6975	−6.39
	$\rho = 0.6$	$E[N]$	1.9500	1.5000	−23.1	1.9500	0	1.9500	0
		$V[N]$	7.1925	3.7500	−47.9	6.3375	−11.8	6.8991	−4.08
	$\rho = 0.8$	$E[N]$	5.6000	4.000	−28.6	5.6000	0	5.6000	0
		$V[N]$	42.720	20.000	−58.2	39.200	−8.2	42.344	−0.88

moment generating function, see Buzacott and Shanthikumar, 1993), and thus, for the M/G/1 queue, the approximation in Eq. (9) for the mean is exact. The second moment can also be obtained, and from this, an expression for the variance of the number in the system:

$$V[N] = \frac{\lambda^3 E[S^3]}{3(1-\rho)} + \frac{\lambda^4 E[S^2]^2}{2(1-\rho)^2} + \frac{\lambda^3 E[S]E[S^2]}{(1-\rho)} + \frac{\lambda^2 E[S^2](3-2\rho)}{3(1-\rho)} + \rho - E[N]^2, \tag{14}$$

where $E[S^k]$ is the k-th moment of the service time distribution.

Thus, in the general algorithm, in Step 1, the first two moments for parts in repair are obtained using Eq. (9) and Eq. (11) or Eq. (12). For the M/G/1 queue, Eq. (9) is exact and the second moment exact expression is given by Eq. (14). Table 3 shows the results of experiments performed on a single-server system with Erlang service time distributions (which will have squared coefficients of variation less than 1) and with gamma service time distributions chosen such that the squared coefficients of variation were greater than 1. The values show the percentage difference in the estimation of the variance (and the mean) of the number in the repair facility with respect to exact results obtained using the PK formula. It is clear that the two-moment approximations outperform the exponential methods in general. Similar results were obtained in experiments with two servers, benchmarked against simulation. The details can be found in Díaz (1995).

2.5. Limited repair facilities III: M / G / k multi-class model

For the multi-class model, we define the following class quantities:

λ_c = arrival rate for class c,
S_c = service time (random variable) for class c customers,
$C_{S,c}$ = coefficient of variation for service time of class c customers,
ρ_c = $\lambda_c E[S_c]$ = offered load for class c customers,
N_c = number of class c customers at the repair facility (in queue or in repair),
Q_c = number of class c customers in queue,
R_c = number of class c customers in repair.

Note that to lessen the notational burden, we have dropped the base subscripts, replacing them with class subscripts.

First, the aggregated arrival rate and service time moments are found by taking the class-weighted average:

$$\lambda = \sum_c \lambda_c, \tag{15}$$

$$E[S] = \sum_c (\lambda_c/\lambda) E[S_c], \tag{16}$$

$$E[S^2] = \sum_c (\lambda_c/\lambda) E[S_c^2]. \tag{17}$$

The per-class number in the system is made of the number in repair, which has expectation $k\rho_c$, and the per-class number in queue. To obtain the expectation of the latter quantity, we first estimate the expected waiting time in queue, which is the same for all classes under Poisson arrivals and a FCFS discipline. Since a multi-class system has a general repair distribution, we use the formulas from the last section:

$$E[W] = \left[\frac{C_a^2 + C_S^2}{2}\right]\left[\frac{E[S](k\rho)^2}{kk!(1-\rho)^2}\right] \times \left[\frac{(k\rho)^2}{k!(1-\rho)} + \sum_{n=1}^{k-1}\frac{(k\rho)^n}{n!}\right], \tag{18}$$

with the expected per-class number in queue given by Little's Law: $E[Q_c] = \lambda_c E(W)$. Thus, the total expected per-class number at the repair facility is given by

$$E[N_c] = \lambda_c E(W) + k\rho_c. \tag{19}$$

To calculate the variance of the per-class number in the system, we calculate the squared coefficient of variation for Q_c, which is exact for M/G/k queues:

$$C_{Q_c}^2 = 1/E[Q_c] + C_W^2. \tag{20}$$

The second term is the squared coefficient for W, which again we assume is the same for all classes, and is given by:

$$C_W^2 = \big(C_D^2 + 1 - P(W>0)\big)/P(W>0), \tag{21}$$

where C_D^2 and $P(W > 0)$ are obtained as in the last section. From this and Eq. (20), the variance of the per-class number in queue is obtained via $V[Q_c] = (E[Q_c])^2 C_{Qc}^2$.

The variance of the per-class number in repair is given by $V[R_c] = E[R_c^2] - E[R_c]^2$. For a single-server system, we have $E[R_c^2] = \rho_c$, giving

$$V[R_c] = \rho_c - (\rho_c)^2. \tag{22}$$

The total variance per class is not just the sum of the variances per class in queue and in repair, as the number in repair and in queue are correlated in general. For a single-server system, we have:

$$V[N_c] = V[Q_c] + V[R_c] + 2E[Q_c]\rho_c/\rho - 2E[Q_c]E[R_c]. \tag{23}$$

For the special case of an M/G/1 queue, exact results for the first two moments of the aggregated waiting distribution can again be obtained through the PK formulas. The expected value is $E[W] = E[S^2]/[2(1-\rho)]$, and the second moment is given by:

$$V[W] = \frac{\lambda E[S^3]}{3(1-\rho)} + \frac{\lambda^2 E[S^2]^2}{2(1-\rho)^2} + \frac{\lambda E[S]E[S^2]}{(1-\rho)} + E[S^2] - E[S+W]^2 - C_S^2 E[S]^2, \tag{24}$$

from which an expression for the aggregated square coefficient of the waiting time is given by $C_W^2 = V[W]E[W]^2$.

The general algorithm is slightly modified then for the multi-class case:

1. Calculate the aggregate arrival rate and service time mean and coefficient of variation using Eqs. (15)–(17).
2. Estimate the aggregate expected waiting time via Eq. (18), and with this the expected per-class number in queue per class via Eq. (19).
3. For a single server, use Eq. (24) to obtain the exact coefficient of variation of the waiting time (this requires obtaining the third moment of the service distribution); otherwise, approximate it using Eq. (21). Use this to obtain per-class variance for number in queue using Eq. (20).
4. Obtain variances of the per-class number in repair using Eq. (22).
5. Use Eq. (23) to obtain per-class variances for total number in the system.
6. Fit a negative binomial distribution to each class and proceed as in the previous section.

A total of nine experiments were conducted, using a combination of mixes of Erlang distributions with different arrival and service rates. There were two classes of customers and a single server in all experiments. Listed in Table 4 are the parameters for the experiments, including the class distributions listed down the left side and the class arrival and service rates listed across the top. The results, benchmarked against the exact PK solution, are similar to those previously obtained, as shown in Fig. 3 and Fig. 4.

Table 4
Experiments for the multi-class case: two classes (1 and 2), single server

Parameters:	$\lambda_1 = 1, \mu_1 = 10, \rho_1 = 0.1$ $\lambda_2 = 2, \mu_2 = 4, \rho_2 = 0.5$	$\lambda_1 = 1, \mu_1 = 5, \rho_1 = 0.2$ $\lambda_2 = 2, \mu_2 = 10, \rho_2 = 0.2$	$\lambda_1 = 1, \mu_1 = 10, \rho_1 = 0.1$ $\lambda_2 = 2, \mu_2 = 20/7, \rho_2 = 0.7$
Class distributions:			
1: Exponential 2: Erlang 2	Experiment 1	Experiment 2	Experiment 3
1: Erlang 2 2: Erlang 3	Experiment 4	Experiment 5	Experiment 6
1: Erlang 3 2: Erlang 4	Experiment 7	Experiment 8	Experiment 9

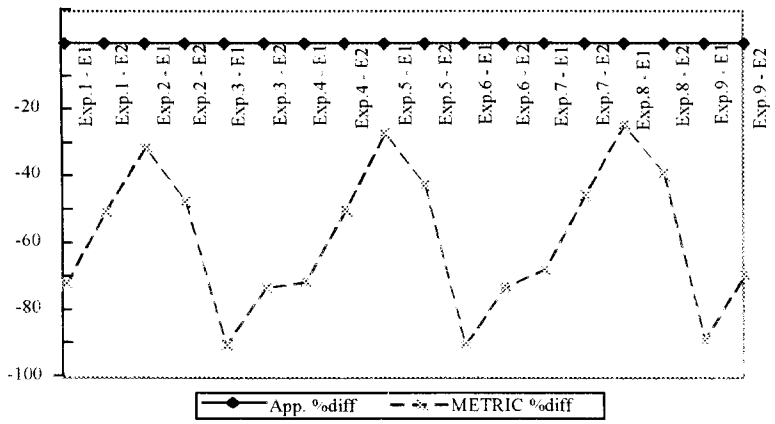

Fig. 3. % differences in the estimation of $E[N_c]$, indicated by Ec, $c = 1,2$.

In the figures, Exp.N is an abbreviation for Experiment N ($N = 1,\ldots,9$), whereas Ec represents the error in class c expectation and Vc represents the error in class c variance ($c = 1,2$). The approximation is almost indistinguishable from the exact solution, whereas METRIC experiences large errors.

3. Case study: Caracas Metro subway system

Caracas is a city of over four million people, situated in a long and narrow valley. This geographic characteristic, together with a high rate (by Latin-American standards) of vehicle ownership, produces spectacular traffic congestion. It was thus with high expectations that a subway system, known as Metro, was inaugurated in 1983. This system consists today of two main lines (a third line is under construction), with almost 40 stations and moving an average of over a million passengers a day (more than twice Washington, DC's Metro, with over twice the number of lines and stations). In a country where there is widespread dissatisfaction with many government services, the government-supported Metro has won a reputation of being clean, safe, fast and cheap. This creates strong political and public opinion pressures to maintain high standards of operation. Unfortunately, the low fare rate combined with economical

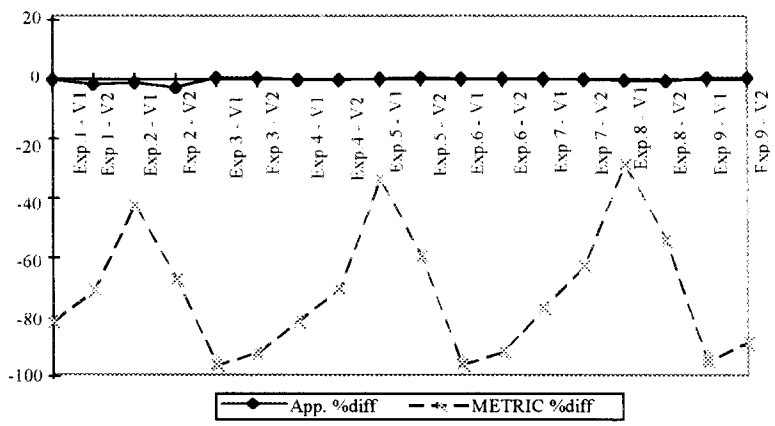

Fig. 4. % differences in the estimation of $V[N_c]$, indicated by Vc, $c = 1,2$.

difficulties in the country have translated into low operational budgets and thus in the loss of technically qualified personnel, and in constraints on the acquisition of support materials. Repairable spare parts, however, are considered to be investments and not common expenditures. At the time of the system acquisition, investments were more easily justifiable in public administration, so the current levels of spare parts of critical items tend to be high. No scientific method is used for their determination, but rather recommendations of the suppliers and experience, so that "the level must be right, as we don't have many serious problems with it" is a common justification for a given level of spare parts. The computerized systems for tracking reliability and inventory control, installed at the start of operations, have been steadily lost, due to the departure of many qualified personnel, so almost any quantitative analysis requires a great deal of manual data gathering. Given this environment, we tried to analyze the system and to apply both METRIC and our proposed models.

The analysis was centered on the type of equipment that the organization regards as more critical, electronic cards. These cards are analog (not digital) and must be repaired in a workshop furnished with specially designed diagnostic equipment. There are some 28,000 cards, of 68 different types (for an average 400 cards of each type). The system can be described as a two-echelon system: at the bases (two different locations at the maintenance yards, one at the end of each line) a few of each type of card are stocked. When a failure occurs, the affected card is exchanged and sent to the central repair facility (the depot). Little stock is kept here, as a repaired card is returned to the corresponding base warehouse. Although the two yards are not very distant (less than one hour by surface), transportation time is almost two days, due to administrative delays. The manager thinks that sometimes the cards stay longer in queue waiting to be transported than in the actual repair process. Almost all repairs are internally done at Metro and the workshop is truly generic: through the use of specially designed electronic testing equipment, the technicians (two, each specialized in certain types of cards) can repair all of the 68 types of cards. These cards can be considered non-indentured, as final repair is done on the card itself and not on its components. We note that although we consider here only the appropriate number of cards to stock for a given number of repairmen, the manager would also like to determine the optimal number of repair technicians (Paulín and Contreras, 1994), and our model could be used to address this issue.

The company does not possess reliable electronic data systems, so the collection of data was a difficult and sometimes incomplete undertaking. Even so, we were able to apply the METRIC and Graves models as well as our proposed models. As we shall see from the analysis, the unnecessary waste inherent in the empirical methods used to set the actual level of spare parts and the error committed by the overestimation of repair facility capacity in the traditional models can be considerable.

The data used in applying the models are summarized as follows:

- The critical part type under consideration had a total population (in operation plus spares) of 378 parts and an average daily failure rate of 0.325.
- Base one corresponded to the yard at line one, where the repair place (depot) is also physically located. The estimated average pipeline time for this base was 1.16 days, and the demand at this base was 70% of total demand (0.228 failures per day). Pipeline length was estimated from the data as the time of arrival at the depot minus the time of occurrence of the failure (which includes administrative delays plus weekends and holidays, during which the depot does not work, although the first level of maintenance does).
- Base two corresponded to the yard at line two. The estimated average pipeline time was 3.88 days, and demand was the remaining 30% of total demand (0.098 failures per day).
- Cards at the depot had a fairly constant repair time of 1.037 hours, but had an average wait of 5.077 days, for an average total stay at the depot of under 5.2 days. There were two reasons for the long stay at the depot: high utilization of the repair facility, and the inclusion of weekends and holidays in the wait times.

Two scenarios were used for our analysis. Under the first, the system was in operation and the time in the repair facility (in queue and in actual repair) and in the pipelines were estimated from sample data. Under the second scenario, only the item repair times

Table 5
Spare parts allocations with corresponding sample fill rates

Model	Empirical (status quo)	M/G/k multi-class 2nd scenario	Graves 1st scenario	METRIC 1st scenario	METRIC 2nd scenario
Total spares	20	10	9	8	6
Sample fill rate	100%	100%	92.6%	88.88%	77.7%

were used (as would be the case for the design of a new system). The application of the multi-class M/G/k model was surprisingly easy. The repair facility acted as two separate single-server FCFS queues, as the two repairmen were exclusive, each being specialized. The focal point in this model was calculating the squared coefficient of variation of the waiting times in queue, C_W^2, from the class parameters, using either PK formulas or approximations. Under the first scenario, this quantity could be estimated directly from the sample data, as the time in the depot minus the repair time gave the waiting time. We further removed from this time all "suspended time", or weekends and holidays (1.87 days, added to the model as an additional pipeline), resulting in $E[W] = 3.226$ days. From this, the expected number of cards waiting in the depot was 1.048 ($E[W]$) and the expected number in repair was 0.014 (ρ_c). C_W^2 was estimated from the sample to be approximately 0.764. From this we followed the previously proposed method to obtain $V[N] = 1.87$ (we assumed the term $2E[Q]\rho_c/\rho$ in Eq. (18) to be zero, as we know that ρ_c is very small and a sensitivity analysis confirmed this for ratios of ρ_c/ρ as high as 0.25). Adding to these quantities the expected values and variances of the quantities in the pipelines, we obtained the expected number in repair plus in-pipeline, 2.308 and its variance, 3.11. With a target fill rate of 99%, our model gave a spare parts requirement of 10.

The corresponding analyses were also carried out under the assumptions of METRIC and the Graves extension, again using a 99% target fill rate. We then used the data to generate a sample fill rate by doing a simulation trace for all the various allocations, including the presently used number of spare parts. The resulting sample fill rate for each of the allocations is given in Table 5, which indicates that the Caracas Metro is overstocking critical parts and that one of the models proposed here can be easily applied to reduce inventories, without deterioration in service quality.

4. Conclusions and directions for further research

Traditional models such as METRIC perform well when the utilization of the repair facility is relatively low. For higher utilizations, the model underestimates the expected value and variance of the number in repair at the depot, as it ignores queueing effects, and the use of the Poisson distribution forces the variance to be equal to the mean. Although the Graves (1985) variation performs better than the basic METRIC model in all cases, due to a better estimation of the variance in the second convolution in our general model, it still suffers from the same limitations of basic METRIC when modeling the first convolution. Given the inherent advantages of METRIC – it doesn't require the complexities of modeling general service time distributions and multiple types of repairs, it is simple to understand and to apply, and it is extensively used in the military world – an interesting topic for further research is to find simple approximations that can be used on existing METRIC implementations to correct the underestimation of the queueing effects in systems with higher utilization of the repair facility.

The double negative binomial approximation proposed in this paper is simple to implement and yields more accurate results. The advantages of this model over traditional models becomes clear at higher utilizations at the repair facility. General service time distributions can be incorporated into the model using approximations, or exact results in the case of single servers.

Systems with multiple types of parts can be analyzed using the aggregation-disaggregation approach. The Caracas Metro system case study was used to illustrate its ease of implementation, its improvement

over METRIC, and its resulting potential savings over the present policy. The models developed here allow for the calculation of the first two moments of the per-class number in the queue for any number of servers. However, the calculation of the variance of the number in repair and the variance of the total number per class was derived for a single server. A natural extension of the model is to derive these quantities for multiple servers.

All of the models assume large parts populations. For relatively small parts populations, an error is introduced by overestimating the arrival rate of failed parts at the depot. A simple approach such as Mean Value Analysis (MVA) may provide a means for evaluating the magnitude of the error introduced by the infinite population hypothesis and allow for a correction factor to be used in conjunction with the previous methods. Preliminary research (Díaz, 1995) shows that even if this method provides only an approximation, its implementation is notably easier than other closed queueing network approaches.

Acknowledgements

This research is supported in part by National Science Foundation Grant D CDR-8803012 and a 1993 University of Maryland General Research Board Summer Grant.

References

Albright, S.C. (1989), "An approximation to the stationary distribution of a multiechelon repairable-item inventory system with finite sources and repair channels", *Naval Research Logistics* 36, 179–195.

Albright, S.C., and Gupta, A. (1993), "Steady-state approximation of a multiechelon multi-indentured repairable-item inventory system with a single repair facility", *Naval Research Logistics* 40, 479–493.

Albright, S.C., and Soni, A. (1988), "Markovian multiechelon repairable inventory system", *Naval Research Logistics* 35, 49–61.

Axsäter, S. (1993), "Continuous review policies for multi-level inventory systems with stochastic demand", in: S. Graves, A. Rinnooy Kan and P. Zipkin (eds.), *Logistics of Production and Inventory*, North Holland, Amsterdam, 175–197.

Buzacott, J.A., and Shanthikumar, J.G. (1993), *Stochastic Models of Manufacturing Systems*, Prentice Hall, Englewood Cliffs, NJ.

Díaz, A. (1995), "Multi-echelon models for repairable items", Ph.D. dissertation, University of Maryland, College Park, College of Business and Management.

Díaz, A., and Fu, M.C. (1995), "Multi-echelon models for repairable items: A review", Working Paper, University of Maryland.

Graves, S.C. (1985), "A multi-echelon inventory model for a repairable item with one-for-one replenishment", *Management Science* 31, 1247–1256.

Gross, D., Gu, B., and Soland, R.M. (1993), "Iterative solution methods for obtaining steady state probability distributions of Markovian multi-echelon repairable items inventory systems", *Computer and Operations Research* 20, 817–828.

Gross, D., and Harris, C.M. (1985), *Fundamentals of Queueing Analysis*, 2nd edition, John Wiley and Sons, New York, NY.

Gross, D., and Ince, J.F. (1978), "A closed queueing network model for multi-echelon repairable items provisioning", *AIIE Transactions* 10, 307–314.

Gross, D., Kioussin L.C., and Miller, D.R. (1987), "A network decomposition approach for approximate steady state behavior of Markovian multi-echelon repairable item inventory systems", *Management Science* 33, 1453–1468.

Gross, D., Miller, D.R., and Soland, R.M. (1983), "A closed queueing network model for multi-echelon repairable item provisioning", *IIE Transactions* 15, 344–352.

Mirasol, N.M. (1963), "A queueing approach to logistics systems", *Operations Research* 12, 707–724.

Nahmias, S. (1981), "Managing reparable item inventory systems: A review", in: L.B. Schwarz (ed.), *Multi-level Production / Inventory Control Systems: Theory and Practice*, Studies in the Management Science 16, North Holland, Amsterdam, 253–277.

O'Malley, T.J. (1994), Private communication.

Palm, C. (1938), "Analysis of the Erlang traffic formula for busy-signal assignment", *Ericsson Techniques* 6, 39–58.

Paulín, P., and Contreras, C. (1994), Private communication.

Sherbrooke, C.C. (1968), "METRIC: A multi-echelon technique for recoverable item control", *Operations Research* 16, 122–141.

Suri, R., Sanders, J.L., and Kamath, M. (1993), "Performance evaluation of production networks", in: S. Graves, A. Rinnooy Kan and P. Zipkin (eds.), *Logistics of Production and Inventory*, North Holland, Amsterdam, 199–286.

Whitt, W. (1983), "The queueing network analyzer", *The Bell Systems Technical Journal* 66, 2779–2815.

Whitt, W. (1993), "Approximations for the GI/G/m queue", *Production and Operations Management* 2, 114–161.