# TWO-ECHELON REPAIRABLE ITEM INVENTORY SYSTEM WITH LIMITED REPAIR CAPACITY UNDER NONSTATIONARY DEMANDS

Hoong Chuin LAU        Huawei SONG
The Logistics Institute – Asia Pacific,
National University of Singapore
Singapore 117543, Republic of Singapore
Email: lauhc@comp.nus.edu.sg , tlish@nus.edu.sg
Phone: (+65) 9767-2913

**ABSTRACT**

We study a repairable item inventory system under limited repair capacity and nonstationary Poisson demands, motivated by corrective maintenance of military equipment. Our goal is to minimize the cost of both spare and repair resource allocation. We propose an efficient analytical model that combines optimization modeling and queuing theory.

**Keywords:** Inventory, Queueing, Repairable Item, Service Network

## 1. INTRODUCTION

Military equipments such as aircrafts are expensive complex structures that break down because components are either worn out or damaged during operations. To support operational readiness (or availability), spare components and maintenance resources are required. However, since spares and resources are costly, consume space and become obsolete over time, there is a tradeoff between cost and availability. The purpose of the military planner is to sustain the demands for equipment to support operations with respect to cost and availability.

In this paper, we propose a model for a two-echelon repairable item inventory system under nonstationary Poisson demands (In this paper, a nonstationary Poisson process refers to one which has a time-dependent arrival rate. see, for example [12]) and limited repair capacity. The goal is to determine the optimal allocation of spares *as well as repair resources* given that the demand follows a *time-dependent* Poisson distribution. In most academic and even commercial offering, it is often assumed that the demand is given by a stationary Poisson process with a mean such as the annual demand rate. In reality however, the demand rate varies as time due to seasonality. This situation is acute in the military context, since military systems usage varies from one mission to another, and transits rapidly from peacetime to wartime. This leads to varying utilization rate over time, which is used in military. Therefore, mean values will cause high errors at a given time point. Another concern peculiar the military context - and perhaps increasingly so in today's commercial context as well - is that the planning period is relatively short so that the underlying inventory system may not converge to steady state. Finally, military systems usually exhibit low failure rates, which makes it difficult even to forecast the average rate in the first place.

This paper is organized as follows. A literature survey is provided in section 2. In section 3, we describe the two-echelon reparable item network structure and present a mathematical model. In section 4, we study a *nonstationary multi-class finite-server queueing system*, which is the core computational problem underlying our model. We develop an analytical approach to compute the expected number of each class of customers in the queueing system at any time efficiently. An optimization algorithm is developed in section 5, considering the effect of time. A summary and future work is given in section 6.

## 2. LITERATURE REVIEW

METRIC (Multi-Echelon Technique for Recoverable Item Control) is a pioneer study presented by Sherbrooke in [16]. In METRIC, it assumes that there are infinite repair resources at the depot and the failures at bases are Poisson processes. Sherbrooke provides an optimization procedure for METRIC by employing marginal analysis. In [5], Graves proposes an approach to use negative binomial distribution instead of Poisson by introducing variance because the variance-to-mean ratio should be 1 under Poisson distribution, but it is usually greater than 1 in practice.

The models with limited repair resources have been studied recently because the assumption of infinite repair capacity is unrealistic in industrial applications. In [4], Díaz and Fu develop a model, considering limited repair facilities at the depot where all failed LRUs are repaired. Unfortunately the variance of per-class number in queue and repair is derived only for single-server due to analytical complexity. In [1] and [2], Alfredsson proposes OPRAL, a model for optimum spare allocation as well as repair facility allocation. The queue at the repair facility is modelled as *M/M/s* so that the expected waiting time for an available resource can be calculated.

In [10], Jung presents a methodology for a repairable inventory system with time-dependent demand by implementing discrete event simulation. In [18], Slay et al. propose an aircraft sustainability model that can handle time-dependent demand rates but infinite repair resources.

In recent years, the Queueing Systems community has been actively working on approximations under time-dependent arrival and service rates (see [6] [7] [8] [15] [19]). However these models suffer one or more of the following limitations: single-class customers, assumption that the maximum traffic intensity is less than one, only long-run average and not time-dependent performance measures, single server, and high computational time complexity.

## 3. MATHEMATICAL MODEL

### 3.1 Service Support Network Structure

In this paper, we will address a two-echelon, single-indenture repairable item inventory model with finite repair capacity under nonstationary Poisson demands. In our environment, one depot supports a number of sites called *bases* where military systems are deployed. Each military system is composed of multiple LRUs (Line Replaceable Unit) which are assumed to be connected in series. For simplicity, we assume all systems are identical. Our goal is to optimize the allocation of both LRU spares as well as repair resources over the life cycle of military systems across all sites within the network.

Due to the limited space at the bases, spares can be allocated at both bases and the depot whereas repair resources are only allowed to be allocated at the depot. When an LRU fails, the system will be grounded at the base and the failed LRU will be removed and replaced by a good component if it is available at the base. Otherwise, the base will place an order from the depot and there is a backorder at the base. The "down" system has to wait until a spare is available. The failed and removed LRU will be delivered to the depot where finite repair resources comprising technicians and tools are allocated. If the required repair resources are available, this LRU will be repaired. Otherwise, it has to wait. When repair is completed, the good LRU will be sent to the depot stock for future demands.

In this paper, the following assumptions are made:

1. All LRUs must and can be repaired at the depot.
2. Continuous resupply, i.e. an LRU can be sent between the depot and bases immediately at any time.
3. The structure supplies a serviceable item for a failed one on a *one-for-one* basis.
4. The repair time of an item follows an exponential distribution.
5. Each LRU requires exactly one repair resource. Different types of LRUs may compete for the same repair resource, and FCFS (First Come First Serve) is accepted as scheduling policy.

6. FCFS (First Come First Serve) replenishment from the depot to bases.

## 3.2 Notations

We adopt and extend the notations of those in [1] and [17]. We use 0 to denote the depot and index the bases by $j$, $j$ = 1, …, $J$. As done in [1], the types of LRU are indexed by $k$, $k$ = 1, „„ $K$. We use $g$, $g$ = 1, …, $G$ to index the repair resource types. Other notations include:

### Input Variables

$T$: the length of planning horizon
$MTBF_k$: mean time between failures of LRU $k$
$TAT_k$: mean repair time of LRU $k$
$OST_k$: constant order-and-ship time for LRU $k$ from the depot to the base
$Nsys_j$: number of military systems deployed at base $j$
$QPM_k$: quantity of LRU $k$ that military system has
$UR(t)$: utilization rate across all bases at time $t$
$Cs_k$: unit cost of spare for LRU $k$
$Cr_g$: unit cost of repair resource $g$

### Immediate Variables

$\lambda_{jk}(t)$: demand rate of LRU $k$ at base $j$ at time $t$
$\lambda_{0k}(t)$: demand rate of LRU $k$ at the depot at time $t$
$N_k(t)$: random variable representing number of LRU $k$ in repair facility at time $t$
$EBO_{jk}(t)$: EBO of LRU $k$ at base $j$ at time $t$
$EBO_{0k}(t)$: EBO of LRU $k$ at the depot at time $t$
$BP_{jk}(t)$: random variable representing number of LRU $k$ in the pipeline of base $j$ at time $t$
$OSP_{jk}(t)$: random variable representing number of LRU $k$ in the order-and-ship pipeline to base $j$ at time $t$
$RP_{0k}(t)$: random variable representing number of LRU $k$ in the depot repair pipeline at time $t$
$f_{jk}(t)$: fraction of LRU $k$ at base $j$ contributing to the EBO at the depot

### Decision Variables

$s_{jk}$: number of spare units of LRU $k$ at base $j$
$s_{0k}$: number of spare units of LRU $k$ at the depot
$r_g$: number of repair resource $g$ at the depot

## 3.3 System Performance Function

First we will compute the demand rate. Similar to the assumption in METRIC, we assume $Nsys_j$ is large. The demand rate of LRU $k$ at base $j$ and the depot are computed by definition respectively as follows:

$$\lambda_{jk}(t) = \frac{UR(t)}{MTBF_k / QPM_k} \times Nsys_j \quad \text{and} \quad \lambda_{0k}(t) = \sum_{j=1}^{J} \lambda_{jk}(t) \tag{1}$$

As with existing literature, we will use time-dependent EBO as our objective function since the demand for an LRU is given by a time-dependent Poisson process. Additionally, $EBO(s|\lambda)$ is used to denote EBO given stock level $s$ when the mean pipeline is $\lambda$ and the computation is in [17].

From the assumption that demand for an LRU is given by a time-dependent Poisson process and from the splitting theorem for non-homogeneous Poisson process, we know that a sum of Poisson processes is still a Poisson process [18]. By the linearity of expectation, the expected value of $BP_{jk}(t)$ is given as:

$$E[BP_{jk}(t)] = E[OSP_{jk}(t)] + f_{jk}(t - OST_k) \cdot EBO_{0k}(t - OST_k) \tag{2}$$

Based on the assumption that $OST_k$ is constant, the dynamic form of Palm's theorem tells us that $OSP_{jk}(t)$ is a Poisson random variable with mean equal to

$$E[OSP_{jk}(t)] = \int_{(t-OST_k)^+}^{t} \lambda_{jk}(u)du \tag{3}$$

The number of LRU $k$ in the repair pipeline at time $t$ is a Poisson random variable with mean equal to the expected number of LRU $k$ in the repair facility, consisting of those in queue and in process at time $t$. That is $E[RP_{ok}(t)] = E[N_k(t)]$. Therefore, given the stock of LRU $k$ at the depot $s_{0k}$, the depot EBO at time $t$ is $EBO_{0k}(t) = EBO(s_{0k} \mid E[RP_{0k}(t)])$.

Since we assume FCFS replenishment policy and the utilization rates are the same across all bases, i.e. demand rates vary synchronously over time, the waiting time for an available spare from the depot at the bases are the same. Hence, we distribute EBO according to the proportion of demand rate, i.e. $f_{jk}(t) = \lambda_{jk}(t)/\lambda_{0k}(t)$.

Following that, the expected number of LRU $k$ in the pipeline of base $j$ at time $t$ can be calculated according to equation (2). Therefore, given the stock of LRU $k$ at base $j$ $s_{jk}$, the EBO of LRU $k$ at base $j$ at time $t$ is given by $EBO_{jk}(t) = EBO(s_{jk} \mid E[BP_{jk}(t)])$. Finally, the total time-dependent EBO is the sum of EBOs over all LRUs at all bases at time $t$, i.e. $EBO(t) = \sum_{k=1}^{K}\sum_{j=1}^{J} EBO_{jk}(t)$.

### 3.4 Optimization Model

In this paper we only consider the investment cost incurred by spares and resources. We will use the terminology LSC (standing for *Life Support Cost* in [13] and [14]) as the notation for total cost, where

$$LSC = \sum_{k=1}^{K} Cs_k (\sum_{j=0}^{J} s_{jk}) + \sum_{g=1}^{G} Cr_g r_g \tag{4}$$

Our optimization model is to determine a minimum-cost spare and repair resource allocation such that the EBO *at any time* will be less than or equal to a given target. So given the planning horizon $T$, we define $MaxEBO = max_{t \in T}\ EBO(t)$ as our final objective function. Given a budget, the model should determine an allocation of spares and repair resources that minimize the EBO while not exceeding the budget. The ultimate goal is to generate a curve (called C/E-curve) where each point on the curve is an optimal allocation for a given cost and a target EBO.

Given a budget **B**, it can be expressed as the following mathematical model:

$$\min\ MaxEBO$$
$$\text{s.t.}\ LSC \le \mathbf{B}$$

## 4. NONSTATIONARY MULTI-CLASS FINITE-SERVER QUEUEING SYSTEM

The computation of EBO involves the calculation of the expected number of LRUs in the repair facility, which has a queueing system behavior. While Palm's theorem is based on infinite channel queueing, our problem is subject to finite repair resources. Furthermore, compared with traditional queueing models such as *M/M/s*, we have multiple classes of customers (LRUs), each of which has its own arrival and service rates, where the arrival process is a non-stationary Poisson process.

In queueing theory terminology, a repair resource is called server and LRU is called customer. Based on our assumption that each LRU requires exactly one repair resource, LRUs can be partitioned into resource groups, each of which consists of all LRUs competing for the certain type of resource. The scheduling policy in each resource group is FCFS, i.e. all demands will wait in a single queue regardless of type until one of the identical parallel repair resources is free. Then the first customer (LRU) in the queue enters service and so on. We call this queueing system *nonstationary multi-class finite-server queueing system*. (See Figure 1)

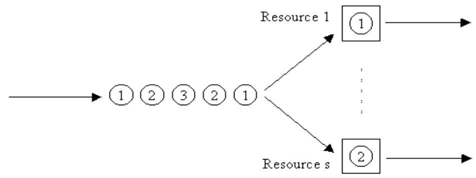


**Figure 1: Multi-Class Queueing System with Finite Servers**

Our aim is to compute the expected number of customers in the queueing system of each class at any time. To our knowledge, there is no literature on how to compute the required measure.

### 4.1 Approximation of Queueing System

In [15], a method was proposed that gives good approximations for nonstationary single-class queueing systems. In [1] and [4], the authors view multiple classes of customers as a single class by using cumulative arrival rate and mean service rate. Our method is essentially to first merge multiple classes into a single class as [1] and [4] and employ the method in [15] to estimate the expected number of customers in the system of all classes. We then calculate the expected number of customers of each class via disaggregation.

Before presenting our method, we introduce some notations.
$\lambda_c(t)$ = arrival rate of class $c$ at time $t$
$\mu_c(t)$ = service rate of class $c$ at time $t$
$\lambda(t)$ = cumulative arrival rate at time $t$
$\mu(t)$ = mean service rate at time $t$
$N_c(t)$ = expected number of class $c$ customers in the system at time $t$
$N(t)$ = expected number of customers of all classes in the system at time $t$
$Q_c(t)$ = expected number of class $c$ customers in queue at time $t$
$Q(t)$ = expected number of customers of all classes in queue at time $t$
$R_c(t)$ = expected number of class $c$ customers in repair (service) at time $t$
$U(t)$ = utilization of servers (i.e. expected number of customers of all classes in repair) at time $t$

The cumulative arrival rate at time $t$ which is the summation of arrival rates of all classes at time $t$ and the mean service rate at time $t$ are given by

$$\lambda(t) = \sum_c \lambda_c(t) \quad \mu(t) = \frac{\lambda(t)}{\sum_c \frac{\lambda_c(t)}{\mu_c(t)}} \tag{5}$$

From [11] and [15], we know how to compute the expected number of customers of all classes in the queueing system at time $t$. The algorithm is given in [15], starting from $t = 0$ when the system is empty. After computing $N(t)$, we calculate the time-dependent expected number of each class customers in the system. We have $N(t) = \sum_c N_c(t)$ for all $t$.

We know from [1] and [4], the expected waiting time for a server is equal for all customers under steady state. Let $W$ be the expected waiting time for a server under steady state. We have $N_c = \lambda_c W + \lambda_c / \mu_c$ for stationary models. If we divide the expected number of customers in the system into two parts: one in queue $Q_c$, and the other in repair (service) $R_c$,, we can see that $Q_c = \lambda_c W$, which is proportional to the arrival rate of class $c$, and $R_c = \lambda_c/\mu_c$, which is proportional to the utilization of class $c$.

Under time-dependent case, since the in-rate is dependent on the arrival rate $\lambda(t)$ and out-rate is dependent on both the arrival rate $\lambda(t)$ and service rate $\mu(t)$, we will divide the expected number of customers of all classes in the system at time $t$ $N(t)$ into two parts: one is that in queue $Q(t)$ and the other is that in service i.e. $U(t)$. Hence, the expected number of each class customers in queue is proportional to the arrival rate and that in service is proportional to the utilization. We have:

$$N(t) = Q(t) + U(t) \text{ and } N_c(t) = Q_c(t) + R_c(t) \tag{6}$$

Since $Q_c(t)$ is proportional to $\lambda_c(t)$ and $R_c(t)$ is proportional to $\lambda_c(t) / \mu_c(t)$, we have

$$Q_c(t) = \frac{\lambda_c(t)}{\lambda(t)} Q(t) = \frac{\lambda_c(t)}{\lambda(t)} (N(t) - U(t)) \text{ and } R_c(t) = \frac{\lambda_c(t)/\mu_c(t)}{\lambda(t)/\mu(t)} U(t) \tag{7}$$

### 4.2 Results and Discussion

We verify the accuracy of this method by experimental comparisons. We run a number of experiments using a combination of different arrival and service rates and compare the results with those generated by simulation. We call our method "analytical" method in contrast with "simulation".

In all of our experiments, we set the number of replications to be 1000, the discrete time step $\delta$ to be 0.1 and start with empty system. We first conduct a stationary system with 3 servers and 2 classes of customers whose arrival and service rates are $\lambda_1 = 1/3$, $\mu_1 = 0.2$ and $\lambda_2 = 0.25$, $\mu_2 = 0.25$. The results are shown in Figure 2. From that, we can find that our analytical method agrees well with simulation.

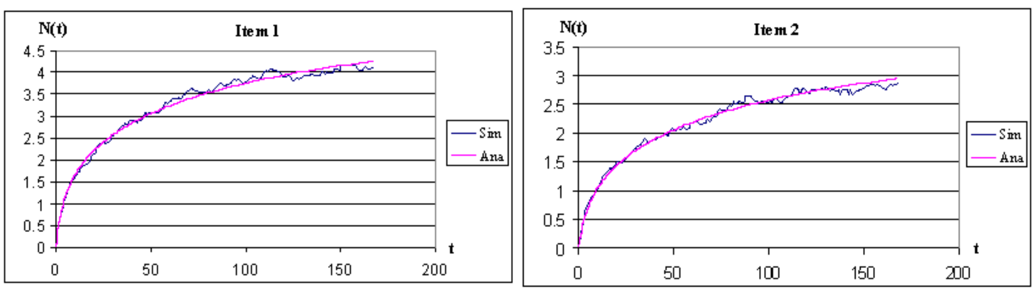


**Figure 2: Expected number of customers of each class in the system**

Next we consider a nonstationary system similar to the first case. There are 3 servers and 2 classes of customers whose service rates are $\mu_1 = 0.2$, $\mu_2 = 0.25$. But the demand rates vary as time, which are given as follows: $\lambda_1 = 1/3$(0-48hr), 1/12(48-120hr), 1/6(120-168hr); $\lambda_2 = 1/4$(0-48hr), 1/16(48-120hr), 1/8(120-168hr). The results are shown in Figure 3. From that we can see that our analytical model agrees well with simulation under nonstationary scenario.

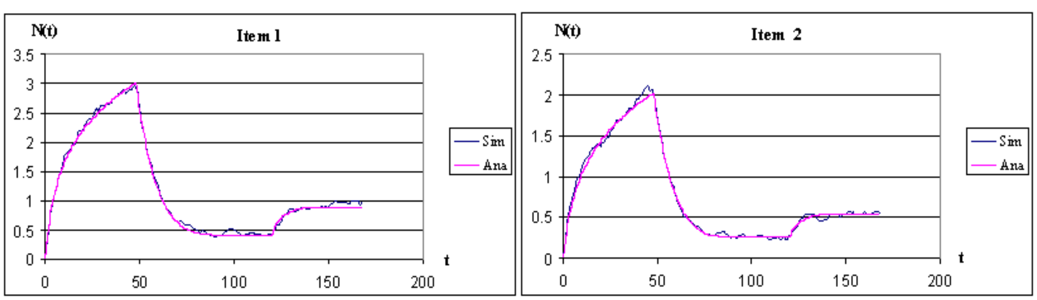


**Figure 3: Expected number of customers of each class in the nonstationary system**

It can be seen that our approach can be efficiently implemented to estimate the expected number of each class customers in the system at a certain time for a nonstationary multi-class finite-server queueing system. Furthermore, our approach gives quite good results within short computational time by means of an analytical model when compared with simulation results.

## 5. OPTIMIZATION

### 5.1 Significance Levels and Resource Groups

In [13] and [14], the concept of significance level is introduced. A stock position is defined by both the item to be stored and the site where it is stored. The significance level describes the importance of different stock positions. The more important the stock position, the higher the significance level is. In our problem, the stock positions of LRUs at bases are the most important because they have direct effects on EBO whereas the spares at the depot are of next importance because they determine the resupply delay for LRUs. The repair resources at the depot are the least important stock position, only influencing the repair pipeline of LRUs at the depot. Consequently, the significance level of repair resources at the depot is set to be 0, that of spares at the depot is 1, and that of spares at the bases is 2. Since the computation of EBO at a certain level only requires the information on stock positions whose significance levels are less or equal, optimization can be carried out level-by-level from the lowest-significance level 0-until the highest level.

Our algorithm follows and extends the approach proposed in [1]. Under the assumption that each failed LRU requires only one repair resource, different LRUs can be partitioned into independent *resource groups*, where each group contains LRUs that share a common repair resource.

### 5.2 Optimization Algorithm

Marginal analysis is a greedy heuristic where one unit of the item with maximum marginal utility is increased at each step. However here we should consider EBO values at different time points together, in order to prevent marginal analysis from being misled. This lead us to develop a generalization of marginal analysis, called *cross-time* optimization, presented as follows. First we divide the time horizon into several periods, each of which is short. And then we apply cross-time optimization. The algorithm is given as follows. It proceeds in 3 phases. The first phase is to apply marginal analysis to find the optimal allocation of repair resources at significance level 0 and generate a C/E curve:

1. Initialize *MaxEBO* and *LSC* based on the initial allocation (such as zero units for all spares and repair resources).
2. Until (*LSC* > **B** or *MaxEBO* is not improved) do the following. Apply marginal analysis to make incremental allocations. Here, an additional unit from each resource group constitutes a candidate, and the marginal utility is computed as the increase in *MaxEBO* per unit cost incurred. Each incremental allocation will result in a new point on the C/E curve.

The next phase is to apply marginal analysis to generate a C/E curve for significance level 1:

3. For each point on the C/E curve of significance level 0 do the following: Generate a C/E curve using that point, and initialize *MaxEBO* and *LSC* based on the existing repair resource allocation at that point until (*LSC* > **B** or *MaxEBO* is not improved).
4. Apply marginal analysis to make incremental allocations. Here, an additional spare unit of each LRU constitutes a candidate, and the marginal utility is again computed as the increase in *MaxEBO* per unit cost incurred.
5. Output a convex curve that envelops all curves obtained from the previous step.

The third phase is to output a C/E curve for significance level 2. In the same way, for each point on the C/E curve of significance level 1, generate a C/E curve for optimal spares allocations at the *bases*, and output a convexification of these curves. Note that our algorithm can be extended to handle multi-echelon problems in a straight-forward fashion.

## 6. CONCLUSION AND FUTURE WORKS

In this paper, we presented the technology for automating military corrective maintenance. The work essentially involves an innovative application of analytical modeling, optimization, queuing theory, and simulation to solve a two-echelon repairable item inventory problem under non-stationary Poisson demands and finite repair resources. We proposed a cross-time optimization algorithm, based on marginal analysis. We investigate EBO values at different time points together, to prevent marginal analysis from being misled. In considering time-dependency, we studied a nonstationary multi-class finite-server queueing system at the depot. We developed an analytical approach to compute the expected number of each class customers in the queueing system at any time.

In our work, the assumption that each LRU requires exactly one repair resource is crucial in reducing the complexity of the problem. What is challenging for future work is to develop analytical models that relax this assumption. We also assume the utilization rates are the same across all bases so that the demand rates vary synchronously. Another challenging aspect of future work is the development of a model where utilization rates vary with different systems at different units at different time.

## References


[1] Alfredsson P., *Optimization of multi-echelon repairable item inventory systems with simultaneous location of repair facilities*, European Journal of Operational Research, v99, 584-595, 1997

[2] Alfredsson P., *OPRAL – A Model for Optimum Resource Allocation*, Systecon AB, Box 5205, SE-102 45 Stockholm, Sweden, 1999

[3] Carrillo M.J., *Extensions of Palm's Theorem: A Review*, Management Science, v37, 739-744, 1991

[4] Díaz A., Fu M.C., *Models for multi-echelon repairable item inventory systems with limited repair capacity*, European Journal of Operational Research, v97, 480-492, 1997

[5] Graves S., *A multi-echelon inventory model for a repairable item with one-for-one replenishment*, Management Science, v31 n10, 1247-1256, 1985

[6] Green L., Kolesar P., *The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals*, Management Science, January 1991

[7] Green L, Kolesar P., Svoronos A., *Some Effects of Nonstationary on Multiserver Markovian Queueing Systems*, Operations Research, May-June 1991

[8] Green L., Kolesar P., *The Lagged PSA for Estimation Peak Congestion in Multiserver Markovian Queues with Periodic Arrival Rates*, Management Science, January, 1997

[9] Isaacson K.E., Boren P., Tsai C.L., Pyles R., Dyna-METRIC Version 4: Modeling Worldwide Logistics Support of Aircraft Components, The RAND Corporation, R-3389-AF, 95-96, Santa Monica, CA, May 1988

[10] Jung W., *Recoverable Inventory Systems with Time-Varying Demand*, Production and Inventory Management Journal v34 n1, 77-81, 1993

[11] Matta I., Shankar A.U., Z-Iteration: *A Simple Method for Throughput Estimation in Time-Dependent Multi-Class Systems*, Proc. ACM SIGMETRICS / PERFORMANCE'95, Ottawa, Canada, 1995

[12] Nelson Barry L., Stochastic Modeling Analysis and Simulation, McGraw-Hill International Edition, 1995

[13] OPUS9 Version 1.6 Users Guide, Systecon AB, January 1992

[14] OPUS10 User's Reference – Logistics Support and Spares Optimization version 3, Systecon AB, May 1998

[15] Rothkopf M.H., Oren S.S., *A closure approximation for the nonstationary M/M/s queue*, Management Science, v25, 522-534, 1979

[16] Sherbrooke C.C., *METRIC: Amulti-echelon technique for recoverable item control*, Operations Research, v16 n2, 122-141, 1968

[17] Sherbrooke C.C., *Optimal Inventory Modeling of System: Multi-Echelon Techniques*, Reading, John Wiley & Sons, New York, 1992

[18] Slay F.M., Bachman T.C., Kline R.C., O'Malley T.J., Eichorn F.L., King R.M., *Optimizing Spares Support: The Aircraft Sustainability Model*, Logistics Management Institute, 2000 Corporate Ridge, McLean, Virginia 22101-7805, October 1996

[19] Whitt W., *The Pointwise Stationary Approximation for* $M_t/M_t/s$ *Queues in Asymptotically Correct as the Rates Increase*, Management Science, v37 n3, March 1991