Multiserver Multichannel Real-Time System with Limited Maintenance Facilities under Maximum Load

Edward Ianovsky¹, Joseph Kreimer²

¹Raval ACS Ltd., Beer-Sheva, Israel
²Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beersheba, Israel
Email: e.ianovsky@hotmail.com, kremer@bgu.ac.il

Received 23 June 2015; accepted 21 July 2015; published 24 July 2015

Abstract

We consider a multi-server and multichannel real-time system with identical servers (e.g. unmanned aerial vehicles, machine controllers, etc.) that provide services for requests of real-time jobs arriving via several different channels (e.g. surveillance regions, assembly lines, etc.) working under maximum load regime. Each channel has its own constant numbers of jobs inside at any instant. Each channel has its own specifications, and therefore different kinds of equipment and inventory are needed to serve different channels. There is a limited number of identical maintenance teams (less than the total number of servers in the system). We compute analytically steady-state probabilities of this system, its availability, loss penalty function and other performance characteristics, when both service and maintenance times are exponentially distributed.

Keywords

Availability, Performance, Real-Time System, Steady-State

1. Introduction

Real-time systems (RTS) are defined as those for which correctness depends not only on the logical properties of the computed results, but also on their temporal properties. In RTS, a calculation that uses temporally invalid data may be useless and sometimes harmful—even if such a calculation is functionally correct. Examples include industrial automation, traffic control, aerospace, robotic, intelligence and defense system, telecommunication and distributed process control, just to name a few.

Several researchers ([1]-[3]) have proposed a number of classic priority algorithms for scheduling real-time

tasks on a single processor. The problem of minimizing the number of processors in multiprocessor computer system executing real-time tasks is studied in [4].

Different scientific communities are treating RTS problems. During last three decades several meta-heuristic methods, such as Tabu Search [5], Simulated Annealing [6] and Greedy Randomized Adaptive Search Procedure [7] are developed. Good surveys on Artificial Intelligence and real-time decision problems (RTDP) are given in references [8]-[11].

The use of analytical methods of queueing theory [12] and stochastic processes [13] has significant benefits in developing RTS.

We will focus on RTS with a zero deadline for the beginning of job processing. In these RTS, jobs are processed immediately upon arrival, if there are available servers. That part of the job which is not processed immediately is lost forever, since queueing of jobs or their parts are not allowed. The particular interest in such RTS is aroused by military intelligence problems involving unmanned aerial vehicles (UAV), which demonstrate a very high efficiency in many local conflicts. It is proved that the non-mix policy of never relieving an operative server maximizes the availability of a multiserver single-channel RTS involving preventive maintenance and working in general regime with any arrival pattern under consideration and constant services and maintenance times ([14] [15]). This policy appears to be optimal for any finite time interval, and not only for infinite horizon. Multiserver (identical servers) and multichannel (identical channels) RTS [16] [17] (with limited and unlimited number of maintenance facilities respectively), working under maximum load regime were treated as finite source queues. Two-dimensional birth-and-death processes [18] [19] were applied in analysis of a multiserver RTS (with ample and limited number of maintenance teams respectively) with two different channels operating under a maximum load regime, when both service and maintenance times are exponentially distributed. In [20] the results of [18] for different channels operating under a maximum load regime were extended. Optimal assignment probabilities maximizing availability of RTS (with ample and limited number of maintenance teams respectively) with large number of servers and two channels were obtained [21] [22]. Multiserver and multichannel RTS with identical servers and channels and ample maintenance facilities, working in general regime with exponentially distributed service, maintenance, jobs inter-arrival and duration times were treated as Markov chains [23] in order to obtain various performance characteristics. The researchers [24] [25] obtained moments of RTS with ample and limited maintenance facilities respectively. In [26] RTS with preemptive priorities policy were considered. In [27] the RTS with exactly one job in each channel was studied. The researchers [28] computed analytically (for exponentially distributed service times) and via Cross Entropy [29]-[31] simulation approach (for generally distributed service times) optimal routing probabilities for RTS with ample maintenance facilities.

This work extends the models [27] [28] as follows. The RTS under consideration assumes the constant number of jobs, specific for each channel. We compute analytically steady-state probabilities of this system, its availability, loss penalty function and other performance characteristics.

The paper is organized as follows: in Section 2 describes the model; Section 3 provides steady state probabilities; Section 4 presents various performance measures; Finally, Section 5 summarizes our results.

2. Description of the Model

We consider a multiserver RTS consisting of N identical servers that provide service for requests of real-time jobs arriving via r different channels required to be under nonstop surveillance. There are exactly \( r \) requests of real-time jobs in \( k \)-th channel at any instant (there are no additional job arrivals to the busy channel), and therefore \( r \) servers at most are used to serve the \( k \)-th channel (with others being on stand-by or providing the service to another channel or in maintenance or waiting for maintenance) at any given time. Thus, the total number of operating servers in the system is at most \( \sum_{k=1}^{N} r_k \).

Each channel has its own specifications and conditions, etc., and therefore different kinds of equipment and inventory are needed to serve different channels.

A server providing service for the \( k \)-th channel is operative for a period of time \( S_k \) before requiring \( R_k \) hours of maintenance. \( S_k \) and \( R_k \) are independent exponentially distributed random variables with parameters \( \mu_k \) (\( k = 1, \ldots, r \)) and \( \lambda \) respectively.

It is assumed that there are \( K \) (\( K \leq N \)) maintenance teams available to repair (with repair times \( R_k \) being i.i.d.r.v.) the servers. Thus, a shortage of maintenance facilities is possible. In that case the server waits for
maintenance. This server is assigned to the $k$-th channel with probability $p_k$ ($k = 1, \ldots, r$). It receives the appropriate kind of maintenance (equipment, programming, etc.), and therefore cannot be sent to another channel. Assignment probabilities $p_k$ may depend upon inventory conditions. They also can be used as control parameters. The duration $R_k$ of repair is exponentially distributed with parameter $\lambda$, and does not depend on the channel. After maintenance, the server will either be on stand-by or serving the channel it was assigned to.

The system works under a maximum load (worst case) of nonstop data arrival to each one of $r$ channels. This kind of operation is typical in high performance data acquisition and control systems, such as self-guided missiles, space stations, satellites, etc.

If, during some period of time of length $T$, there is no available server to serve one of the jobs, we will say that the part of the job of length $T$ is lost.

Queues of jobs or their parts cannot exist in RTS, nevertheless they can be fitted into a framework of finite source queues, while using a dual approach of changing the roles between jobs and servers.

3. Steady State Probabilities

Denote: $\lambda_k = \lambda p_k$ and $\rho_k = \lambda_k / \mu_k$, $(n_1, n_2, \ldots, n_r)$ the state of the system, where $n_k$ ($k = 1, \ldots, r$) is a number of fixed servers assigned to the $k$-th channel (obviously $\sum n_k \leq N$), and $p_{n_1, n_2, \ldots, n_r}$ the corresponding steady state probability. There are $\binom{N + r}{r}$ states in total.

The above RTS can be presented as a closed Jackson queuing network (Figure 1) consisting of $N$ customers ($N$ servers of the RTS), $r + 1$ nodes ($r$ channels and maintenance station of the RTS) with $r_1$ servers at $k$-th node ($r_2$ jobs in the $k$-th channel of the RTS) ($k = 1, \ldots, r$) and $K$ servers at $r + 1$-th station ($K$ maintenance teams of the RTS). The network has transition probabilities $p_k$ (assignment probabilities of the RTS) from $r + 1$-th node to $k$-th one ($k = 1, \ldots, r$), and transition probabilities from $k$-th station to $r + 1$-th are equal to 1. Other transition probabilities are equal to 0. Service times at the network stations (operating and maintenance times of the RTS) are exponentially distributed. The customers of the network cannot leave it and cannot come to it from outside. Thus, the network is a closed Jackson network by definition.

From the description of the RTS as a closed Jackson network we obtain ([12] [32]) that the steady-state probabilities $p_{n_1, n_2, \ldots, n_r}$ are given by the following:

Theorem 1: A real-time system with $N$ servers, $K$ ($K < N$) maintenance crews, $r$ ($r \geq 2$) different channels operating under a maximum load regime with $r_1$ jobs in $k$-th channels at any instant, and exponentially

![Figure 1. Closed Jackson network.](image-url)
distributed operating and maintenance times (with parameters $\mu_k$ (for the $k$-th channel ($k = 1, \ldots, r$)) and $\lambda$ respectively) has following values of the steady-state probabilities

$$p_{n_1, n_2, \ldots, n_r} = \frac{K!}{\min\left\{\sum_{i=1}^{r} n_i, N-K\right\}} \prod_{j=1}^{r} \frac{\rho_{j}^{n_j}}{\min\left(n_j, r_j\right)! r_j^{\max\{n_j-r_j, 0\}}} p_{0,\ldots,0},$$

(3.1)

$$p_{0,\ldots,0} = \left[\sum_{n_1, n_2, \ldots, n_r \geq 0 \atop n_1 + \cdots + n_r \leq N} \frac{K!}{\min\left\{\sum_{i=1}^{r} n_i, N-K\right\}} \prod_{j=1}^{r} \frac{\rho_{j}^{n_j}}{\min\left(n_j, r_j\right)! r_j^{\max\{n_j-r_j, 0\}}} \right]^{-1},$$

(3.2)

where $\left(n_1, n_2, \ldots, n_r\right)$ is a state of the RTS, $n_k$ is number of fixed servers of the RTS in the $k$-th channel and $\rho_k = \lambda_k / \mu_k$ ($k = 1, \ldots, r$).

We will use the following notations:

1) $Q_N^{(k)}(\rho_1, \ldots, \rho_r) = \sum_{n_1, n_2, \ldots, n_r \geq 0 \atop n_1 + \cdots + n_r \leq N} \frac{K!}{\min\left\{\sum_{i=1}^{r} n_i, N-K\right\}} \prod_{j=1}^{r} \frac{\rho_{j}^{n_j}}{\min\left(n_j, r_j\right)! r_j^{\max\{n_j-r_j, 0\}}} \left(\prod_{k=1}^{r} \rho_{k}^{n_k}\right)

= \sum_{n_1=0}^{N} \sum_{n_2=0}^{N-n_1} \sum_{n_3=0}^{N-n_1-n_2} \cdots \sum_{n_r=0}^{N-n_1-n_2-\cdots-n_{r-1}} \frac{K!}{\min\left\{\sum_{i=1}^{r} n_i, N-K\right\}} \prod_{j=1}^{r} \frac{\rho_{j}^{n_j}}{\min\left(n_j, r_j\right)! r_j^{\max\{n_j-r_j, 0\}}} \left(\prod_{k=1}^{r} \rho_{k}^{n_k}\right)

and

2) $P_N^{(k)}(n_k, \rho_1, \ldots, \rho_r)$ the probability that $n_k$ fixed servers are assigned to the $k$-th channel ($k = 1, \ldots, r$).

Then, using Theorem 1, we obtain:

$$p_{0,\ldots,0} = \left[Q_N^{(k)}(\rho_1, \ldots, \rho_r)\right]^{-1},$$

$$P_N^{(k)}(n_k, \rho_1, \ldots, \rho_r) = \left(\frac{(K \rho_k)^{n_k}}{\min\left(n_k, r_k\right)! r_k^{\max\{n_k-r_k, 0\}}} Q_N^{(k)}(\rho_1, \ldots, \rho_r)\right)^{-1}

= \frac{(K \rho_k)^{n_k}}{\min\left(n_k, r_k\right)! r_k^{\max\{n_k-r_k, 0\}}} \left(\prod_{k=1}^{r} \rho_{k}^{n_k}\right)

(3.3)$$
where \( n_k = 0, N, \quad k = 1, r \).

4. Performance Characteristics

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your journal for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

In this section we show how to obtain some useful performance characteristics of the RTS under consideration.

Each server can be in one of the following states:
(i) busy (serving a job of one of the channels);
(ii) in maintenance;
(iii) on stand-by (for one of the channels);
(iv) waiting for maintenance.

Each job of \( k \)-th channel \( (k = 1, r) \) can be in one of two positions:
(i) in service;
(ii) out of service.

Keeping in mind that \( n_k \) is a number of fixed servers assigned to the \( k \)-th channel \( (k = 1, r) \), we can represent the number of channels and servers in different positions in terms of \( n_k \), namely:

Number of fixed servers is \( \sum_{k=1}^{r} n_k \);

Number of fixed servers serving the \( k \)-th channel (also a number of jobs served in the \( k \)-th channel) is \( \min(n_k, r_k), \quad k = 1, r \);

Number of fixed servers on stand-by for the \( k \)-th channel is \( \max(n_k - r_k, 0), \quad k = 1, r \);

Number of broken servers is \( N - \sum_{k=1}^{r} n_k \);

Number of broken servers in maintenance is \( \min\left(N - \sum_{k=1}^{r} n_k, K\right)\);

Number of broken servers waiting for maintenance is \( \max\left(N - \sum_{k=1}^{r} n_k - K, 0\right)\).

Now we can obtain corresponding average values (performance characteristics):

\[
L_k = E\left[n_k\right] = \sum_{n_1=0}^{N} \sum_{n_2=0}^{N-n_1} \cdots \sum_{n_r=0}^{N-r_{k-1}} n_k p_{n_1, n_2, \ldots, n_r}
\]

for average number of fixed servers assigned to the \( k \)-th channel, \( k = 1, r \);

\[
L = \sum_{k=1}^{r} L_k
\]

for average number of fixed servers in the system;

\[
L_{k,b} = E\left[\min(n_k, r_k)\right] = \sum_{n_1=0}^{N-n_k} \sum_{n_2=0}^{N-n_1} \cdots \sum_{n_r=0}^{N-r_{k-1}} \min(n_k, r_k) p_{n_1, n_2, \ldots, n_r}
\]

for average number of busy servers serving the \( k \)-th channel (of jobs served in the \( k \)-th channel), \( k = 1, r \);

\[
L_{k,sv} = E\left[\max(r_k - n_k, 0)\right] = r_k - L_{k,b}, \quad k = 1, r
\]
for average number of nonserved jobs in the \( k \)-th channel;

\[
L_b = \sum_{k=1}^r L_{k,b}
\]

for average number of busy servers in the system (average number of jobs in service);

\( N - L \) for average number of broken servers;

\[
L_m = E \left[ \min \left( N - \sum_{k=1}^r n_k, K \right) \right] = \sum_{n_1=0}^{N-K} \sum_{n_2=0}^{N-K} \cdots \sum_{n_r=0}^{N-K} \min \left( N - \sum_{k=1}^r n_k, K \right) p_{n_1,\ldots,n_r}
\]

\[
= K \sum_{n_1=0}^{N-K} \sum_{n_2=0}^{N-K} \cdots \sum_{n_r=0}^{N-K} p_{n_1,\ldots,n_r} + \sum_{n_1=0}^{N-K} \sum_{n_2=0}^{N-K} \cdots \sum_{n_r=0}^{N-K} n_j = \max \left( 0, N - \sum_{k=1}^r n_k \right) p_{n_1,\ldots,n_r}
\]

for average number of broken servers in maintenance;

\[
p_k L_m
\]

for average number of broken servers in maintenance assigned to the \( k \)-th channel, \( k = 1, \ldots, r \);

\[
L_w = E \left[ \max \left( N - \sum_{k=1}^r n_k - K, 0 \right) \right] = \sum_{n_1=0}^{N-K} \sum_{n_2=0}^{N-K} \cdots \sum_{n_r=0}^{N-K} \max \left( N - \sum_{k=1}^r n_k - K, 0 \right) p_{n_1,\ldots,n_r}
\]

for average number of broken servers waiting for maintenance;

\[
L_{k,wb} = E \left[ \max \left( n_k - r_k, 0 \right) \right] = \sum_{n_1=0}^{N-K} \sum_{n_2=0}^{N-K} \cdots \sum_{n_r=0}^{N-K} \max \left( n_k - r_k, 0 \right) p_{n_1,\ldots,n_r}
\]

for average number of broken servers waiting for maintenance;

\[
L_{wb} = \sum_{k=1}^r L_{k,wb}
\]

for average number of fixed servers on stand-by in the system;

\[
\lambda_{av} = \lambda L_m
\]

for average rate of servers arriving from maintenance;

\[
\lambda_{av,k} = \lambda p_k L_m
\]

for average rate of servers arriving from maintenance assigned to the \( k \)-th channel, \( k = 1, \ldots, r \);

\[
W_{k,wb} = L_{k,wb} / \lambda_{av,k}
\]

for average time spent on stand-by by server assigned to the \( k \)-th channel, \( k = 1, \ldots, r \) (Little Theorem);

\[
W_k = W_{k,wb} + 1 / \mu_k
\]

for average time of server assigned to the \( k \)-th channel being fixed, \( k = 1, \ldots, r \) (Little Theorem).

\[
R_N^{(k)} (\rho_1, \ldots, \rho_r) = L_{k,b} = \sum_{n=0}^N \min \left( n, r_k \right) P_N^{(k)} (n, \rho_1, \ldots, \rho_r)
\]

is an average number of jobs served in the \( k \)-th channel for any given moment.

It follows from (3.3) that

\[
R_N^{(k)} (\rho_1, \ldots, \rho_r) = \sum_{n=0}^N \min \left( n, r_k \right) \left( K \rho_k \right)^n Q_N^{-1} (\rho, \rho_{k+1}, \ldots, \rho_r) / Q_N (\rho_1, \ldots, \rho_r)
\]
\[ L_{k,ns} = E\left[ \max \left( r_k - n_k, 0 \right) \right] = r_k - L_{k,b} = r_k - R_N^{(i)} (\rho_1, \cdots, \rho_r) \]

for average number of nonserved jobs in the \( k \)-th channel, \( k = 1, \cdots, r \).

The use of RTS relies on the principle of availability. We will introduce therefore several definitions.

**Definition 4.1:** For a channel (with \( r_k \) jobs inside) operating under maximum load regime, the average availability is given by the following formula

\[ Av_k = E\left[ \text{number of served jobs in } k\text{-th channel} \right] / r_k = R_N^{(i)} (\rho_1, \cdots, \rho_r) / r_k. \]

**Definition 4.2:** For a multichannel system (number of channels \( r \geq 2 \)) operating under maximum load regime, the average system availability is given by the formula

\[ Av_N (\rho_1, \cdots, \rho_r) = \sum_{i=1}^{r} R_N^{(i)} (\rho_1, \cdots, \rho_r) / \sum_{i=1}^{r} r_k. \]

Another important characteristic is an average loss penalty (or operation cost) of system operation in equilibrium during time unit.

**Definition 4.3:** Let \( C_k \ (k = 1, r) \) be the cost of time unit during which one job in \( k\)-th channel is out of service (penalty). Then formula

\[ TC_N (\rho_1, \cdots, \rho_r) = \sum_{k=1}^{r} C_k \left( r_k - R_N^{(i)} (\rho_1, \cdots, \rho_r) \right) \]

represents average loss penalty function.

We also have the following important relationship between system availability and its loss penalty function, namely for \( C_1 = C_2 = \cdots = C_r = 1 \)

\[ Av_N (\rho_1, \cdots, \rho_r) = 1 - TC_N (\rho_1, \cdots, \rho_r) / \sum_{k=1}^{r} r_k \] (4.3).

## 5. Conclusion

We presented a multiserver multichannel real-time system working in maximum load regime with the shortage of maintenance teams. The system consists of \( r \) different channels with exactly \( r_k \) jobs in \( k\)-th channel. We obtained analytically the steady state probabilities and various performance measures, providing complete description of this system. These analytical results could be used immediately without long simulations by researchers and practitioners. Our next goal is to find the optimal routing probabilities, maximizing the system availability.

## References


