Autonomic Resource Management for Multimedia Services Using Inventory Control

Ramy Farha and Alberto Leon-Garcia

University of Toronto, Toronto, Ontario, Canada
ramy.farha@utoronto.ca, alberto.leongarcia@utoronto.ca

Abstract. The proliferation of real-time multimedia services has triggered attempts by service providers to seek ways to reduce the cost of managing those services, while satisfying customer demands for the resources involved in the service delivery. A key issue is to avoid costly service interruptions in such real-time multimedia services. In this paper, we propose the use of the inventory control approach from the Operations Research community to remedy to this problem. We show two possible inventory control models to manage the resources involved in real-time multimedia service delivery. We also perform extensive simulations to compare the advantages and disadvantages of the two inventory control models for resource management of real-time multimedia services.

1 Introduction

Recent years have witnessed a proliferation of real-time multimedia services delivered to customers. In parallel, the number of wireless customers asking for such services has also increased. Service providers offering such services are therefore faced with the challenge of offering real-time multimedia services to customers with desirable Quality of Service (QoS) characteristics [1]. In order to achieve this, real-time multimedia service providers are exploring novel techniques to automate resource management for their services. The goal is to obtain a self-managing network infrastructure, which adapts to changes in customer demands, satisfying the needs for networking, computing, and storage resources.

Inventory control is commonly used in the Operations Research community to analyze inventory systems of industries and businesses, placing and receiving orders when needed to meet demands for a given product [2]. An analogy between inventory control and autonomic resource management for multimedia real-time services, is that of balancing the two extreme cases: Having a small amount of resources leads to costly service interruptions, thus leading to customer dissatisfaction, while having a large amount of resources leads to idle capital expenditures, thus leading to lower profits for service providers. An important factor in the formulation and solution of a resource inventory model is whether the demand (per unit time) for a resource is deterministic (known with certainty) or probabilistic (described by a probability distribution). Since the first assumption is too optimistic in a real-world environment, we will turn our

D. Krishnaswamy, T. Pfeifer, and D. Raz (Eds.): MMNS 2007, LNCS 4787, pp. 161–172, 2007.

attention to probabilistic inventory models. In this paper, we will use inventory control for autonomic resource management of multimedia real-time services.

The rest of this paper is structured as follows. In section 2, we summarize some related work. In section 3, we describe two inventory control models we examine in this paper for autonomic resource management of multimedia real-time services. In section 4, we show how the inventory control models are applied to a situation involving multimedia real-time service providers. In section 5, we show some simulation results. In section 6, we conclude this paper.

2 Related Work

The use of inventory control to solve problems for real-time multimedia services is a relatively new concept. One paper [3] considered applying inventory control to capacity management for utility computing. A framework consisting of theoretical foundations, problem formulations, and quality of service (QoS) forecasting is presented. While the paper considers inventory control to solve a problem in the IT world, it does not tackle the problem explored in this paper.

However, adaptive resource management for wireless networks offering real-time multimedia services is not a new idea. Several papers have been written on the issue. One paper used flow and admission control algorithms for efficient resource utilization in wireless networks [4]. This approach is based on control theory concepts to offer network-aware multimedia applications in wireless networks. Another paper [5] proposed bandwidth adaptation in case of insufficient bandwidth in order to allocate the desired bandwidth to every multimedia connection originating in a cell or being handed off to the cell.

The main differences between previous approaches and the one proposed in this paper are that a) we consider predictive rather than reactive mechanism, b) we do not need to manage at the granularity of individual multimedia connections as the overhead might become prohibitive, and c) we do not interfere with the regular operation of the existing service instances which are active in a given cell (by adapting their bandwidth or dropping some using priority schemes).

3 Inventory Control

The developed models for probabilistic inventory control in Operations Research [2] are broadly categorized under continuous and periodic review situations. We will focus on the continuous review model, where the inventory is being monitored on a continuous basis so that a new order can be placed as soon as the inventory level drops to the reorder point. A continuous review inventory system for a particular resource is based on two critical numbers: reorder point R and order quantity y. The inventory policy is a simple one: whenever the inventory level of the resource drops to R units, place an order for y more units to replenish inventory. All inventory problems seek to answer two questions: *when* to order, and *how much* to order.

The costs involved in the Inventory Control problem are:

- Purchasing Cost: Based on the price per unit of the resource. It may be constant or variable.
- Setup Cost: Represents the fixed charge incurred when an order is placed. This cost is independent of the size of an order.
- Holding Cost: Represents the cost of maintaining the inventory in stock. It includes the cost of storage, maintenance, and handling.
- Shortage Cost: Represents the penalty incurred when we run out of stock. It includes potential loss of income, as well as the more subjective cost of loss in customer's goodwill. It can also include the lost revenue opportunity and the cost of possible delays resulting from the shortage in a given resource.

Determining the best inventory control approach revolves around making a managerial decision on the desired service level. We will consider two possible continuous review probabilistic inventory models in this paper and compare their advantages and shortcomings. The first model, which we will refer to as Model 1, uses a buffer stock to account for the probabilistic demand. The second model, which we will refer to as Model 2, is a more exact model which includes the probabilistic demand directly in the formulation. In the first model, the goal is to minimize the number of shortages encountered during operation. In the second model, the goal is to minimize the total expected cost during operation.

The assumptions of the continuous review probabilistic inventory models are:

1. Each use of the model involves a single resource.
2. The inventory level is under continuous review.
3. The only decisions to be made are to choose R and y.
4. There is a lead time L between the time when the order is placed and when the order quantity is received. This lead time L can be either fixed or variable.
5. The demand for resources from the inventory during the lead time L is uncertain. However, the probability distribution $f(x)$ of the demand is stationary with time, and the corresponding expected demand per unit time is given by D.
6. If a stockout occurs before the order is received, the excess demand is backlogged, so that the backorders are filled once the order arrives. Therefore, the excess demand is not lost, but is instead held until it can be satisfied once enough resources are delivered to replenish the inventory.
7. The fixed setup cost is denoted by K, and is incurred each time an order is placed. But except for this setup cost, the cost of the order is proportional to the order quantity y of a resource with unit cost c.
8. The holding cost h is incurred for each unit of resource in the inventory per unit time.
9. When a stockout occurs, the shortage cost p is incurred for each unit of resource backordered per unit time until the backorder is filled.

3.1 Model 1

Model 1, shown in Fig. 1, reflects the probabilistic nature of the demand by using an approximation that superimposes a constant buffer stock on the inventory

level throughout the entire planning horizon. The size of the buffer is determined such that the probability of running out of stock during lead time L does not exceed a given value. We define the following additional parameters:

- x_L: random variable representing demand during lead time L.
- μ_L: average demand during lead time L.
- B: buffer stock size.

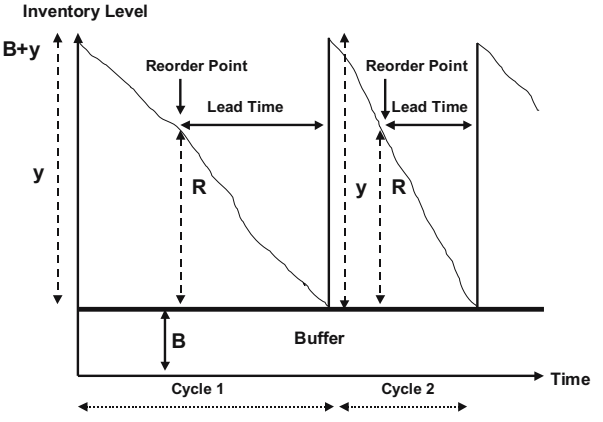

Fig. 1. Continuous Review Probabilistic Inventory Model 1

As mentioned before, the demand x during the lead time L is random with probability distribution $f(x)$. The buffer stock B needs to be found for a given probability of stock shortage. This maximum allowable probability of running out of stock during the lead time L is given by α. The probability statement used to determine B can be written as:

$$P\{x_L \geq B + \mu_L\} \leq \alpha \tag{1}$$

3.2 Model 2

Model 2, shown in Fig. 2, allows for a shortage of demand. The reorder level R is a function of the lead time L between placing and receiving an order. The optimal values of y and R are determined by minimizing the expected cost per unit that includes the sum of all costs incurred in this inventory model.

The elements of the cost function are:

1. **Setup Cost:** The approximate number of orders per unit time is $\frac{D}{y}$, so that the setup cost per unit time is $\frac{KD}{y}$.
2. **Expected Holding Cost:** The average inventory is given by:

$$I = \frac{y + E(R - x) + E(R - x)}{2} = \frac{y}{2} + R - E(x) \tag{2}$$

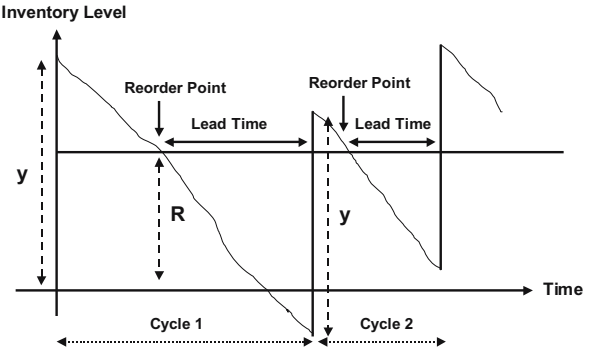

Fig. 2. Continuous Review Probabilistic Inventory Model 2

The expected holding cost per unit time thus equals hI. The formula is based on the average of the beginning and ending expected inventories of a cycle, $y + E(R - x)$ and $E(R - x)$, respectively. As an approximation, the expression ignores the case where $R - E(x)$ may be negative.

3. **Expected Shortage Cost:** Shortage occurs when $x > R$. Thus, the expected shortage quantity per cycle is given by:

$$S = \int_R^\infty (x - R) f(x) dx \tag{3}$$

Because p is assumed to be proportional to the shortage quantity only, the expected shortage cost per cycle is pS, and based on $\frac{D}{y}$ cycles per unit time, the shortage cost per unit time is $\frac{pDS}{y}$.

The resulting total cost function per unit time is given by:

$$TCU(y, R) = \frac{DK}{y} + h\left(\frac{y}{2} + R + E(x)\right) + \frac{pD}{y}\int_R^\infty (x - R) f(x) dx \tag{4}$$

The solutions for optimal y^* and R^* are determined from:

$$\frac{\partial TCU}{\partial y} = -\frac{DK}{y^2} + \frac{h}{2} - \frac{pDS}{y^2} = 0 \tag{5}$$

$$\frac{\partial TCU}{\partial R} = h - \frac{pD}{y}\int_R^\infty f(x) dx = 0 \tag{6}$$

We thus get the following solutions for the optimal values of y and R, denoted by y^* and R^*:

$$y^* = \sqrt{\frac{2D(K + pS)}{h}} \tag{7}$$

$$\int_{R^*}^\infty f(x) dx = \frac{hy^*}{pD} \tag{8}$$

Because y^* and R^* cannot be determined in closed forms from the above two equations, a numeric algorithm is used to find the solution. The algorithm is proved to converge in a finite number of iterations, provided that a feasible solution exists.

For $R = 0$, the last two equations, respectively, yield:

$$\hat{y} = \sqrt{\frac{2D(K + pE(x))}{h}} \tag{9}$$

$$\tilde{y} = \frac{pD}{h} \tag{10}$$

If $\tilde{y} \geq \hat{y}$, unique optimal values of y and R exist. The solution procedure recognizes that the smallest value of y^* is $\sqrt{\frac{2KD}{h}}$, which is achieved when $S = 0$. The steps of the algorithm are:

Step 0. Use the initial solution $y_1 = y^* = \sqrt{\frac{2KD}{h}}$, and let $R_0 = 0$. Set $i = 1$, and go to step i.

Step 1. Use y_i to determine R_i from the second equation. If $R_i \approx R_{i-1}$, stop; the optimal solution is $y^* = y_i$, and $R^* = R_i$. Otherwise, use R_i in the first equation to compute y_i. Set $i = i + 1$, and repeat step i.

4 Application of Inventory Control

Having introduced some approaches to traditional inventory control, we now turn our attention to their possible applications for management of resources in next generation network infrastructures. With the increasing trend towards virtualization of physical resources to deal with heterogeneity, and with the rise of on-demand computing through the exchange of virtual resources as commodities on a market of service providers, we envision service providers in the future to exchange virtual resources (networking, computing, and storage) in an open market based on demand. Thus, the use of inventory control models to improve the performance of such markets is investigated. More specifically, we consider the use of the inventory control model in the context of service providers dealing with multimedia and real-time services that should not be interrupted because of lack of resources as customers move across cells in next generation networks. Fig. 3 shows the application of the inventory model for management of real-time multimedia services in wireless networks managed by a given service provider which owns a stock of physical resources hosting virtual resources managed by the inventory control system.

In traditional work on resource management, we assume that orders are placed as needed between the service providers and that resources purchased are instantaneously replenished. However, in practice, a lead time elapses between the time at which an order is placed and the time at which the virtual resource amount ordered is delivered. Hence, the policy would be to use the continuous

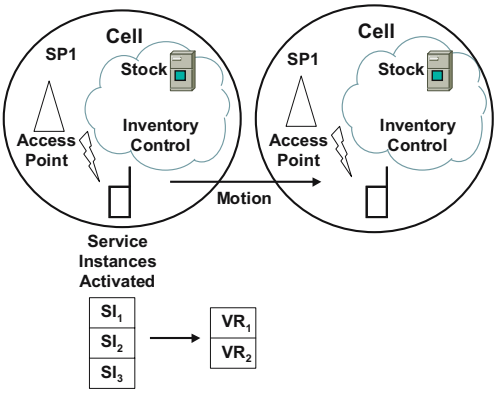

Fig. 3. Example of the Inventory Model for Wireless Real-Time Multimedia Services

review probabilistic inventory model to order virtual resources whenever needed to avoid real-time multimedia service interruptions. Since demand varies in unpredictable ways, we use collected historical data to approximate the distribution of this demand put on service providers for a given virtual resource, in order to find the expected demand during the lead time. The empirical distribution could be found by summarizing the raw data gathered over a training period where the demand for a given virtual resource is studied. The summary is in the form of an appropriate frequency histogram and allows the service providers to determine the associated empirical probability distribution functions of demand.

The inventory control provides autonomic resource management to allow resources to be ordered when needed, as shown in Algorithm 1. The resources are used by service instances activated by the customers for the services they had bought, hence service providers need to keep track of each virtual resource consumption and make appropriate decisions on *when* to order and *how much* to order in order to avoid service interruptions to the customers. In case of mobile customers, the variation in demands for a given resource is driven by customers moving between cells managed by the same or different service providers.

Fig. 4 shows how inventory control models can be incorporated in an autonomic loop for management of virtual resources by service providers subjected to variable demands by fixed and mobile customers. We consider the use of the inventory model in the context of service providers owning several resources needed by the various real-time multimedia services offered by this service provider. Assuming that the service provider has a stock of N virtual resources needed by V services. The virtual resources are referred to as: VR_1, VR_2, ..., VR_N. The stock of virtual resource VR_i at time t is denoted by $A_t(VR_i)$ and needs to be constantly monitored. Initially, the stock of virtual resources is formed by buying from other service providers offering virtual resources, or by using previously owned virtual resources amounts. Let the initial amounts of the N virtual resources be referred to as: $A_0(VR_1)$, $A_0(VR_2)$, ..., $A_0(VR_N)$. Customers,

whether fixed or mobile, place variable demands on those N Virtual Resources. Each service provider is therefore running N inventory models in parallel.

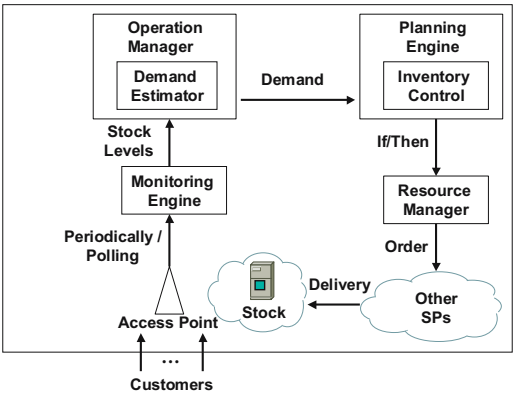

Fig. 4. Inventory Control Autonomic Loop

Algorithm for Autonomic Resource Management using Inventory Control
Training Period
For each Virtual Resource VR_i:
Run network for a training period
Generate empirical distributions $f(x_i)$ based on statistical data on demand collected in training period
Deduce mean demand per unit time D_i
Deduce initial order policy [If/Then]: If $A_0(VR_i) < R_i$, order y_i
Inventory Control Period
Turn Inventory Control policy ON:
Monitor Virtual Resource VR_i stock
Continuously update empirical distributions $f(x_i)$ based on statistical data on demand collected on the fly
Update estimate of mean demand per unit time D_i
Update order policy [If/Then] at time t: If $A_t(VR_i) < R_i$, order y_i
When inventory level of VR_i drops below R_i, order y_i

Algorithm 1. Pseudo Code for Inventory Control

5 Simulation Results

To study the performance of the two inventory control models for autonomic resource management of multimedia real-time services, we built a custom simulator using the Java programming language. We divide the network infrastructure into several cells over a city, modelled as a rectangular grid, which has main and secondary streets, with one or more access points in each cell, allowing both fixed and mobile access over several access network technologies based on the

customer's mobile device capability. Access independence is made possible through fixed mobile convergence, proposed in the IP Multimedia Subsystem (IMS) [6].

We also emulate the movement of mobile customers at a given speed which depends on whether the cell is mainly spanning a main street in the city, or a secondary speed, with an exponentially distributed sojourn time in each cell, and transfer the activated service instances and the amounts of virtual resources they need to new access points as the customers move between cells. Note that, as the mobile customers move to new cells, if they cannot find a bootstrap access point supporting their mobile device's access network technologies, they have to terminate their existing service instances which were previously activated and were still running, and continue their motion, waiting to move to a new cell which might have access points supporting any of the mobile device's access network technologies to reactivate those service instances.

We create customer entities connecting to access points through physical resources, to activate the service instances corresponding to real-time multimedia services bought from service providers. The virtual resources in the network infrastructure are distributed on physical resources which are geographically spread in the cells. Each physical resource hosting a given service has a random capacity for each virtual resource used in the network infrastructure. The customers buy several services from the real-time multimedia service providers and activate several instances of such services. The service instances are activated according to start times which follow an exponential distribution, and are kept active for service times which follow another exponential distribution.

The holding, shortage, and unit of resource costs are kept constant for the duration of a given simulation. We monitor the inventory level of the stock of virtual resources at each time step, and update the empirical distributions of the virtual resource demands. The order policies for different virtual resources are updated on a periodical basis using the most updated demand distributions. The total cost calculated for each simulation is given by the sum of the purchasing costs, the setup costs, the holding costs, and the shortage costs, which are all calculated on a unit time basis, and then added to the total cost.

The parameters used in the simulations are as follows. The unit time step is set to 1 second. The training time is set to 1000 seconds, where data is collected, and empirical distributions of the demand on a service provider are deduced. The policy is reviewed every 20 seconds. The lead time is chosen to be 10 seconds. In problems where the costs are varied, the holding costs, shortage costs, and setup costs are proportional. The shortage cost is chosen to be 100 times the holding cost, and the setup cost is chosen to be 100 times the holding cost. The unit cost is chosen to be equal to the shortage cost. The reason for this choice is that while Model 1 always works, Model 2 only produces a unique and optimal solution when $\tilde{y} \geq \hat{y}$, thus for a fair comparison we need to set the costs appropriately for both inventory control models to operate normally. We run each simulation for 20000 seconds.

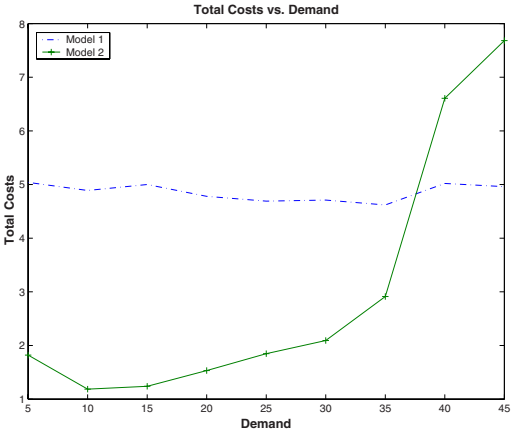

Fig. 5. Total Costs vs. Demand for both Inventory Models

Fig. 5 shows the variation in the total costs incurred by the service provider for both models as the demand increases. The holding cost is set to 10, so the unit cost is set to 10, the shortage cost is set to 1000, and the setup cost is set to 100000. The observation here is interesting. For higher demands, Model 1 seems to perform better than Model 2 because of the lower costs incurred on the service provider, even though Model 2 was designed for reduction of costs. The problem is that for high demand, the cycle time in Model 2 reduces to lower than the lead time, and in extreme cases, can become lower than a unit time. This leads to consistent shortage in Model 2, and therefore while Model 1 orders a large amount of resources, Model 2 is paying the price for shortage and this contributes to its total cost increasing.

Fig. 6 shows the variation in the total costs for low and high demands as the costs per unit time are increased in both inventory models. Initially, the holding cost is set to 1000, so the unit cost is set to 1000, the shortage cost is set to 100000, and the setup cost is set to 10000000. The total costs for both models increase as all the costs are increased by a factor of 1 to 9. For Model 1, the probability of "no shortage" is set to 0.9. For low demand (Fig. 6a), the demand is set to a maximum of 5 resource units per unit time from a given customer. Model 1 outperforms Model 2 in this setup. For high demand (Fig. 6b), the demand is set to a maximum of 50 resource units per unit time from a given customer. Model 2 outperforms Model 1 in this setup. Thus, we can conclude that the performance of one model versus the other depends not only on the demand for a given resource, but also on the costs chosen in the network, as these costs affect the design of model 2 and the values chosen for y^* and R^*.

Fig. 7 shows the inventory level variation as a function of time for both inventory models. For Model 1 (Fig. 7a), the "no shortage" probability was chosen to be 0.9. As expected, the inventory level drops from the initial value of 50000 resource units at the start time as demand for that resource starts in the network.

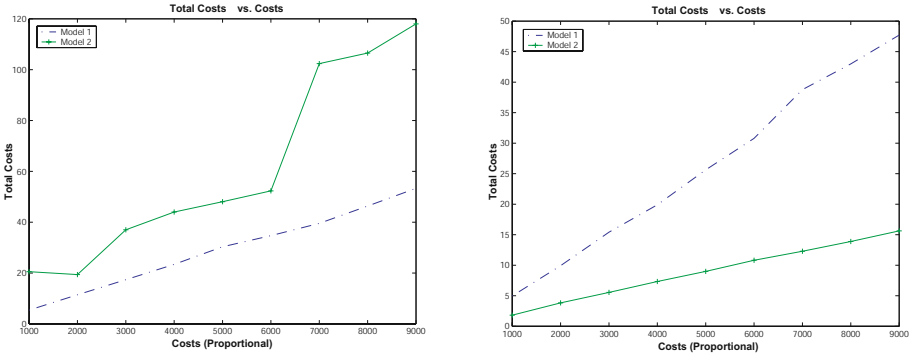

Fig. 6. Total Costs vs. Costs for Low and High Demand for both Inventory Models

When the threshold value is reached, an order is placed for an additional 50000 resource units. As seen in the figure, few orders are needed, and the inventory level is always positive, i.e. no shortages have occurred. For Model 2 (Fig. 7b), the inventory level drops from the initial value of 50000 resource units at the start time as demand for that resource starts in the network. When the threshold value R^* is reached, an order is placed for an amount y^* of resource units. As seen in the figure, several orders are needed, and the inventory level is always fluctuating around zero, so several shortages have occurred.

We performed several other experiments on the two inventory control models. However, due to space limitations, we will not present them here. The main conclusions from the experiments are that inventory control Model 1 is more predictable, while inventory control Model 2 is less controllable. While inventory

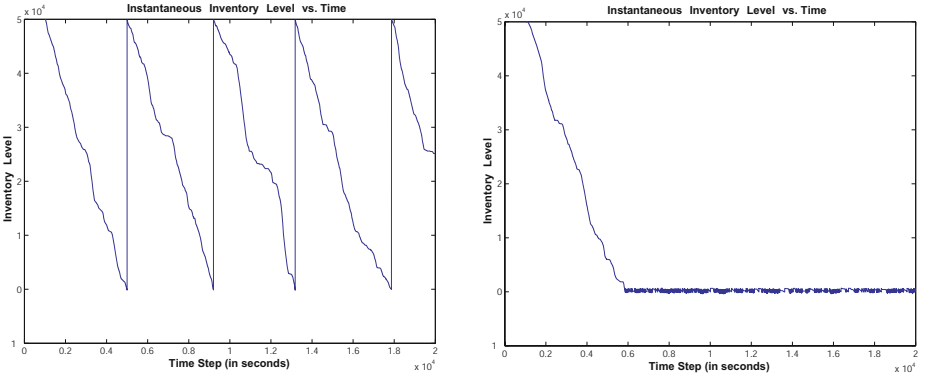

Fig. 7. Inventory Level Variation with time for Both Inventory Models

control Model 1 seems sensitive to changes in the holding costs for a given shortage cost, inventory control Model 2 reduces the total costs incurred, but increases the number of shortages expected and the number of orders that need to be placed. Therefore, inventory control Model 1 is more adapted to real-time multimedia services, but at a higher cost for a service provider. Inventory control Model 2 is better in terms of costs incurred by the service provider, but leads to high customer dissatisfaction as frequent shortages occur.

6 Conclusion

In this paper we proposed to apply inventory control models for autonomic resource management of real-time multimedia services offered by a service provider. Two such models are studied: the first aims at minimizing the probability of resource shortages occurring, while the second aims at minimizing the costs incurred by service providers. Results have shown that while the first model is more appropriate for customers, it increases costs for service providers. On the other hand, the second model is more appropriate for service providers, but leads to customer dissatisfaction. A more detailed study of how demand, policy update period, and cost variation affect the performance of the two models is still needed. However, this paper shows that inventory control is a promising approach for autonomic resource management of real-time multimedia services.

References

1. Ye, J., et al.: A comprehensive resource management framework for next generation wireless networks. IEEE Transactions on Mobile Computing, 249–264 (2002)
2. Hillier, F.: Introduction to Operations Research. McGraw-Hill, New York (2001)
3. Hellerstein, J., et al.: A framework for applying inventory control to capacity management for utility computing. In: 9th IFIP/IEEE International Symposium on Integrated Network Management, pp. 237–250. IEEE Computer Society Press, Los Alamitos (2005)
4. Banerjee, N.L., et al.: Adaptive resource management for multimedia applications in wireless networks. In: Sixth IEEE International Symposium on the World of Wireless Mobile and Multimedia Networks, pp. 250–257. IEEE Computer Society Press, Los Alamitos (2005)
5. Seth, M., et al.: Adaptive resource management for multimedia wireless networks. In: IEEE 58th Vehicular Technology Conference, pp. 1668–1672 (2003)
6. 3rd Generation Partnership Project, Technical Specification Group Services and Systems Aspects : IP Multimedia Subsystem (IMS): Stage 2. 3GPP TS 23.228. (2003)