Inventory Theory Applied to Cost Optimization in Cloud Computing

Andrea Nodari  
Imperial College London (UK)  
a.nodari15@imperial.ac.uk

Jukka K. Nurminen  
Aalto University (FI)  
jukka.k.nurminen@aalto.com

Christian Frühwirth  
Nordcloud Oy (FI)  
christian.fruewirth@nordcloud.com

ABSTRACT
Cloud computing providers offer two different pricing schemes when renting virtual machines: reserved instances and on-demand instances. On-demand instances are paid only when utilized and they are useful to satisfy a fluctuating demand. Conversely, reserved instances are paid for a certain time period and are independent of usage. Since reserved instances require more commitment from users, they are cheaper than on-demand instances. However, in order to be cost-effective compared to on-demand instances, they have to be extensively utilized. This work focuses on finding the optimal combination of on-demand and reserved instances, such that the demand is satisfied and the costs minimized. To achieve this goal, this study introduces a stochastic model of the resources, based on Inventory Theory. The idea is to formulate the optimization problem as an inventory-keeping problem and then derive the optimal strategy. The paper evaluates the proposed model using data from an industry case, comparing the performance with a brute-force approach. The conducted experiments show that the Inventory Theory model provides accurate results and potentially allows prior research on Inventory Theory to be applied to optimal cloud provisioning.

CCS Concepts
• Applied computing → Decision analysis;

Keywords
Cloud provisioning; inventory theory; cost optimization; resource allocation

1. INTRODUCTION
In recent years, cloud computing has become increasingly popular. Cloud computing providers, such as Amazon Web Services\footnote{https://aws.amazon.com}, allow customers to rent virtual machines of two categories: on-demand instances and reserved instances. On-demand instances are virtual machines created and paid for only when utilized. A cloud user adds and removes an on-demand instance with maximum flexibility. Conversely, reserved instances are computational resources reserved and paid for a certain period, with an upfront fee. The latter category requires a higher level of commitment for the user; therefore, if extensively utilized, they result to be cheaper during a long-term utilization.

In order to avoid unnecessary expenses, user of cloud resources needs careful planning. On one hand, reserved instances are useful for cost savings. On the other hand, if reserved instances are underutilized, they generate unnecessary costs. Therefore, the cost optimization problem we investigate in this paper is how to satisfy the uncertain future demand with a combination of reserved and on-demand instances with minimum cost.

In recent years, researchers have studied the field of cost optimization in cloud computing. One of the most promising methods is to utilize Integer Programming to model the optimization problem\cite{4,5,3}. Other authors exploit a two-step approach: first, they propose a demand forecaster and then, they aim to find an optimal solution with evolutionary algorithms\cite{8,11}.

Nonetheless, current state-of-the-art approaches have some limitations. In particular, some works propose stochastic models to calculate the optimal number of reserved instances, assuming the cumulative distribution function of the demand as available\cite{3,5,7}. This paper aims to overcome some of the limitations of state-of-the-art approaches, by explaining how to use the model in real industry cases.

In addition, one of the key contributions of this study is a novel application of Inventory Theory to the problem of cost optimization in cloud computing. Inventory Theory is a branch of operations research that aims to scientifically describe the behavior of an inventory system. The idea is to describe the cost optimization problem as an inventory-keeping problem. After the formulation, it is possible to derive the optimal policy that minimizes the costs. This paper shows the steps needed in order to utilize the findings of the model with observed data. To the best of our knowledge, this is the first study that applies Inventory Theory to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC 2016, April 04-08, 2016, Pisa, Italy  
Copyright 2016 ACM 978-1-4503-3739-7/16/04...$15.00  
http://dx.doi.org/10.1145/2851613.2851869
The idea is to model the hourly cost to satisfy a certain demand as a random variable. Afterwards, the analytical solution to minimize the costs is found. The proposed approach follows Hillier et al. [6], which discuss the applications of operations research in greater details.

The problem of minimizing the cost by using reserved instances has remarkable similarities with the goals of Inventory Theory. Inventory Theory models help companies to deal with stock of goods in order to minimize the costs. The goal of the model is to describe the optimal purchase in order to satisfy the demand of a product. Similarly, in the context of cloud computing, the decision-makers manage a reserved instances portfolio and optimize purchases to satisfy the demand with minimum cost.

Our work proposes a stochastic model which formulates the total cost involved to satisfy the demand of computing capacity. Inventory Theory models help companies to deal with stock of goods in order to minimize the costs. The goal of the model is to describe the optimal purchase in order to satisfy the demand of a product. Similarly, in the context of cloud computing, the decision-makers manage a reserved instances portfolio and optimize purchases to satisfy the demand with minimum cost.

The idea is to model the hourly cost to satisfy a certain demand using reserved instances. If the reserved instances are not sufficient to satisfy the demand, additional on-demand instances are added. The on-demand instances are more expensive than reserved instances, using the Inventory Theory terminology this could be associated to shortage costs.

The optimization is performed in a fixed amount of time of past data [2] and assumes observations to be independent identically distributed.

### Table 1: The parameters of the inventory theory model

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_{od}$</td>
<td>Hourly cost for an on-demand instance</td>
</tr>
<tr>
<td>$c_{ri}$</td>
<td>Effective hourly cost for a reserved instance, which is the hourly cost amortized in a year of usage with the upfront fee</td>
</tr>
<tr>
<td>$y$</td>
<td>Number of reserved instances</td>
</tr>
<tr>
<td>$D$</td>
<td>Random variable representing the demand of instances</td>
</tr>
<tr>
<td>$d_i$</td>
<td>$i^{th}$ observation of the demand</td>
</tr>
<tr>
<td>$C(D, y)$</td>
<td>Total cost to satisfy demand $D$ with $y$ reserved instances</td>
</tr>
<tr>
<td>$E[C(D, y)]$</td>
<td>Expected value of the total cost</td>
</tr>
<tr>
<td>$f_D$</td>
<td>Probability distribution function of the demand</td>
</tr>
<tr>
<td>$F_D$</td>
<td>Cumulative distribution function of the demand</td>
</tr>
<tr>
<td>$F_D^{-1}$</td>
<td>Quantile function of the demand</td>
</tr>
</tbody>
</table>

Starting from a empty inventory (i.e. no initial reserved instances) the total cost of a period of $N$ hours can be expressed as:

$$ C(D, y) = \sum_{i=1}^{N} (c_{ri} y + c_{od} \max\{0, d_i - y\}) $$ (1)

Let us now assume $D$ a continuous random variable with probability density $f_D(x)$, and cumulative distribution function $F_D(x) = \int_0^x f_D(x) dx$. In order to minimize the total cost, the goal of the next step is to find the optimal value for $y$, the number of reserved instance to purchase. The cost for an hour with demand $D$ and $y$ reserved instances is expressed as follows.

$$ C(D, y) = c_{ri} y + c_{od} \max\{0, D - y\} $$ (2)

Let us now calculate the expected value:

$$ C(y) = E[C(D, y)] $$ (3)

$$ = \int_0^\infty C(x, y) f_D(x) dx $$ (4)

$$ = \int_0^\infty (c_{ri} y + c_{od} \max\{0, x - y\}) f_D(x) dx $$ (5)

$$ = c_{ri} y + \int_y^\infty c_{od} (x - y) f_D(x) dx $$ (6)

At this point, it is necessary to minimize the expected total cost by taking the derivative and set it to zero. Assuming that the cost function has one point of minimum, the expression is:

$$ C(y) = c_{ri} y + c_{od} \int_0^\infty (x - y) f_D(x) dx $$

$$ - c_{od} \int_0^y (x - y) f_D(x) dx $$ (7)

$$ \frac{dC(y)}{dy} = c_{ri} - c_{od} + c_{od} \int_0^y f_D(x) dx $$ (8)

$$ = c_{ri} - c_{od} \left[ - \int_0^y f_D(x) dx \right] $$ (9)

$$ = c_{ri} - c_{od} \left[ - F_D(y) \right] = 0 $$ (10)

Solving this expression results in:

$$ F_D(y) = \frac{c_{od} - c_{ri}}{c_{od}} $$ (11)

Therefore, the value $y$ such that the condition above holds, minimize the cost. To demonstrate that the solution minimizes $C(y)$, let us calculate the second order derivative and verify that it is $\geq 0$ for every value of $y$.

$$ \frac{d^2C(y)}{dy} = c_{od} f_D(y) \geq 0 $$ (12)

### 3. Evaluation

The goal of the experiment is to validate the correctness of the proposed model (i.e. Equation 11). The result from the model is compared with the result from Algorithm 1 which is utilized to compute the ground-truth value for the optimal number of reserved instances. This simple brute-force algorithm computes the cost of satisfying the demand by increasing the number of reserved instances, stopping when
the costs increase. Figure 2 illustrates the result when running the algorithm. The input data is a real-world trace of demand, representing the number of utilized instance each hour (Figure 1). The trace has been obfuscated for confidentiality.

Algorithm 1: Brute-force Approach

**Data:** hourlyDemand: number of instances per hour in a certain time frame

**Result:** optimal number of reserved instances

RIcounter ← min(hourlyDemand)

optimalRI ← 0

currentCost ← +∞

while true do

/* Calculate the cost using RIcounter reserved instances */
tmpCost ← calculateCost(hourlyDemand, RIcounter)

if tmpCost ≤ currentCost then

optimalRI ← RIcounter

currentCost ← tmpCost

RIcounter ← RIcounter + 1

else

break

end

In order to apply the finding of the model to the data, it is necessary to perform some preliminary steps.

Starting from Equation 11 assuming that \( F_D \) is invertible we obtain the following:

\[
y = F_D^{-1}\left(\frac{c_{od} - c_{ri}}{c_{od}}\right)
\]  

(13)

The function \( F_D^{-1} \) is a function such that \( F_D^{-1}(p) \) is the \( p^{th} \) quantile of the distribution. Therefore, we need to find the \( p^{th} \) quantile of the observations of the demand, where \( p = \frac{c_{od} - c_{ri}}{c_{od}} \). The algorithm to find the \( p^{th} \) quantile is simple. Given \( N \) observations of the hourly demand, the steps are the following.

1. Sort \( d_1...d_N \) in \( d_{(1)}...d_{(N)} \)
2. Calculate \( p = \frac{c_{od} - c_{ri}}{c_{od}} \)
3. Output \( D_{\lfloor (Np) \rfloor} \), where \( D_{\lfloor (Np) \rfloor} \) is the smallest \( \lfloor Np \rfloor^{th} \) value of demand observations

The assumption made in these steps is that the observations are independent identically distributed. Extensive empirical tests have been performed using different instance types and demand traces. The estimator of quantile outputs accurate results, always matching with the brute-force algorithm. For more detailed results and explanations see [9].

4. DISCUSSION

The experiment proves that the proposed model provide accurate results. The evaluation includes the assumption that the hourly values of demand are independent identically distributed. Although this assumption might lead to inaccuracy when applied to time series, it does not significantly decrease the performance in the experiments.

A possible alternative to the proposed approach is to use the brute-force algorithm. Figure 3 illustrates the performance of the brute-force approach and the model. The experiment is performed analyzing 30 days of data. The two approaches return the same results, however the brute-force algorithm is slower than the model. In addition, the brute-force approach has unpredictable performance. While sorting \( n \) observations has complexity \( O(n \log(n)) \), a brute-force approach might have varying performance depending on the initialization and the optimal number of reserved instances.

This study shows that Inventory Theory accurately models the dynamics of cost optimization in cloud computing. This paper proposes a first application of Inventory Theory in this context, we believe that further work is possible, exploiting the literature of Inventory Theory and applying some results in the context of cloud computing [6].

5. RELATED WORK

In the last years, researchers have analyzed the problem of finding the optimal number of reserved instances, given a certain demand.

Chaisiri et al. [4] propose models based on Integer programming. The work formulates the optimization problem and solves it using stochastic integer programming. The objective is to minimize the overall cost of the infrastructure. The authors improve their work focusing of Amazon Web Service offering [3]. They develop an algorithm for cost optimization by provisioning of on-demand instances, reserved instances, and spot instances. The model is based
on long-term and short-term provisioning algorithms. Additional improvements are also proposed to efficiently solve the optimization problem, which might be hard when the size of the problem increases [5].

Mark et al. [8] propose a two-step approach. First, the authors design a demand forecaster. The experiments show accurate results of three algorithms: simple Kalman filter, Double Exponential Smoothing, and Markov Chains. The next step utilizes a hybridized algorithm consisting of different evolutionary algorithms for optimization: standard Genetic Algorithm, Particle Swarm, and Ant Colony. The researchers compare the results with Stochastic Integer Programming solution, showing similar results.

Another work thoroughly describes the problem of unnecessary costs in cloud computing [2]. The authors argue that a crucial point for an effective planning is to understand future demand. The study shows the result of the experiments conducted using the load curve of four popular websites. The evaluation demonstrates that Genetic Algorithms generate accurate results.

On the context of Cloud Computing, a potential improvement is to include discount rate to take into account the time value of money in the model. Moreover, a research direction would be to model an entire purchase plan utilizing a multi-product inventory system [1], not only a single instance type. Finally, a further improvement would be to formulate take risks into account.

6. CONCLUSION AND FUTURE WORK
This paper proposes a first application Inventory Theory in the context of Cloud Computing. A potential improvement is to include discount rate to take into account the time value of money in the model. Moreover, a research direction would be to model an entire purchase plan utilizing a multi-product inventory system [1], not only a single instance type. Finally, a further improvement would be to formulate take risks into account.

Acknowledgements
This work has been developed at the R&D team at Nordcloud Oy [9]. Part of the work was supported by the Green Big Data project of Helsinki Institute of Physics.

7. REFERENCES


http://nordcloud.com