Formulation to minimize the cost of meeting the demand of simultaneous requests on Video on Demand centers.

Julio César Hidalgo Sánchez
Electrical Engineering, Universidad de las Fuerzas Armadas - ESPE
Av. del Progreso SN, Sangolguí, Ecuador
jchidalgo7@espe.edu.ec

***Abstract*—This document is a solution for the Video On Demand Case Study in order to determine the number of servers to attend all requests per hour in a VOD center. This paper models requests on a Video on Demand (VOD) Center in order to minimize the cost function of weighted number of servers of each type with some constrains such as the availability of bandwidth, total available servers for the VOD service, the cost of turning on and off a server, and the total number of servers on or off at the beginning of the i_{th} hour. Once the model is developed and Service Level is established in order to add complexity to the model, a simulation is run to verify assumptions and sensitivity of the objective function.**

***Index Terms*—video on demand centre, simulation, requests on data centres, modelling data centres, energy conservation.**

I. Introduction

The rapid growth of video on demand services in market data, causes many problems of efficiency because it should work within certain levels to maintain good quality of service. Currently, virtual platforms have greatly helped resolve the issues of energy efficiency. Furthermore, it has now raised (Gallego et al, 2013) that this can be solved with Mixed Integer Programming large scale using software to resolve this issue. However, to understand how it should set a data center in particular, an empirical model should made be aware to understand the problem and a possible solution for starting and stopping services within a data center. An initial approach to this particular problem is considered in this paper in order to minimize an objective function with some constrains.

II. Mathematical Formulation

A. Description of the Mathematical Formulation

The mathematical model is formulated in order to minimize the cost of meeting the demand of simultaneous requests per hour that that are needed to satisfy a calculated service level. The total cost function will be stated in terms of four decision variables as following:

- A variable for the numbers of servers that are turned ON
- A variable for the number of servers that are turned OFF
- A variable for the number of servers that are kept ON from the previous hour to the next one
- A variable for the number of servers that are kept OFF from the previous hour to the next one.

In addition to the four decision variables mentioned above, the mathematical model will include the following costs per hour:

- The cost of keeping ON a number of servers during any hour
- The cost of turning ON a number of servers during any hour
- The cost of turning OFF a number of servers during any hour

The solution obtained by stating the mathematical formulation only in terms of minimizing the objective function (stated in terms of the previously mentioned decision variables and their associated costs per hour) will be that none of the servers should be turned ON at any hour during a cycle time. In order to prevent this situation, four constraints must be added to our model as following:

- A constraint to guarantee that the VOD center is able to satisfy the required number of requests per hour according to a calculated service level. Furthermore, this constraint is responsible for not letting the VOD center to exceed its fixed and pre-established capacity.
- A constraint to guarantee that the total number of available servers is not exceeded at any hour (the total number of servers will be measured in terms of the number of servers kept ON, servers turned ON, servers turned OFF, and servers kept OFF).
- A constraint to prevent that a fixed number of servers are kept ON during the whole cycle time in which the VOD center is working.
- A constraint to guarantee that the number of turned ON and OFF servers at each hour is consistent with the number of servers that were ON and OFF during the previous hour.
- A constraint to guarantee that all servers are OFF at the first time the VOD is launched
- Note: It is assumed that the VOD center started working

for the first time at the beginning of a certain day. However, every time the system reaches the time zero after the first cycle, it will carry out some of the servers that were ON during the last hour of the previous one to the next cycle.

Furthermore, the VOD center for which the solution is designed is going to be working 24 hours a day, and the number of requests for each hour is characterized by a Pareto Distribution with parameters α_i and β_i for each hour.

The sets, indices, variables and parameters required to develop the mathematical formulation are:

Set

$I \rightarrow$Set of all working hours available during a day; I=0,1,2..,23,24. The consideration of the number zero is because of the initialization of the model with all servers OFF.

Indices

$i \rightarrow$index for hours during a day.

Variables

$x \rightarrow$Random variable for the number of requests per hour. (distributed Pareto with parameters α_i and β_i.

$s \rightarrow$Variable that represents the number of requests that the VOD center is able to serve per hour.

$s^* \rightarrow$Optimal value of number of servers that must be on at the it^h hour in order to maximize the expected profit

$y_i \rightarrow$Number of servers turned ON at the beginning of the it^h hour.

$n_i \rightarrow$Number of servers ON at the beginning of the ith hour that were ON during the $(i-1)^{th}$ hour.

$v_i \rightarrow$Number of servers turned OFF at the beginning of the ith hour that were ON during the $(i-1)^{th}$ hour.

$z_i \rightarrow$Number of servers OFF during the ith hour that were OFF during the $(i-1)^{th}$ hour.

$\xi_i \rightarrow$Binary variable 1,0

Parameters

$r_i \rightarrow$Total users served during ith hour with service level p.

$q_i \rightarrow$Bandwidth required by Low and High customers during ith hour.

$h_i \rightarrow$Bandwidth partition into high quality during ith hour.

$l_i \rightarrow$Bandwidth partition into low quality during ith hour.

$w_i \rightarrow$Number of clients in high quality during i-th hour.

$f_i \rightarrow$Number of clients in low quality during ith hour.

$o \rightarrow$Maximum permitted operating cost per hour.

$b \rightarrow$Bandwidth per user in High Quality.

$u \rightarrow$Bandwidth per user in Low Quality.

$h \rightarrow$Cost of having 1 server ON during 1 hour.

$k \rightarrow$Total bandwidth capacity when all servers are ON at any hour.

$t \rightarrow$Total number of servers at the VOD center.

$\beta \rightarrow$Cost of turning ON 1 server at the beginning of the ith hour.

$g \rightarrow$Cost of turning OFF 1 server at the beginning of the ith hour .

$\alpha \rightarrow$Bandwidth capacity per server.

$a \rightarrow$Constant to weight the quantity of preferred users.

$\varphi_L \rightarrow$Price for hour for a user in low quality bandwidth capacity.

$\varphi_H \rightarrow$Price for hour for a user in high quality bandwidth capacity.

$r \rightarrow$Average revenue per request per hour in dollars.

$c \rightarrow$Average cost per request per hour in dollars.

$p \rightarrow$Loss of goodwill cost in terms of unsatisfied requests per hour.

$m \rightarrow$High consumption of resources cost in terms of having idle capacity for 1 extra request.

$C_u \rightarrow$Cost of underage or the cost of having a capacity requests less or equal than the actual demand

$C_o \rightarrow$Cost of overage or the cost of having a capacity request greater than the actual demand

$s_L \rightarrow$Service level or the probability that the demand will be less than the capacity

$\bar{s}\rightarrow$Service level or the probability that the demand will be less than the capacity

B. Mathematical Model of the VOD Center

Minimize

$$\mathbf{G}(\Gamma) = \sum_{i\in I}[(h+\beta)\times y_i + h\times n_i + g\times v_i] \qquad (1)$$

Subject To

$$q_i \leq \quad \alpha\times(n_i+y_i) \quad \leq k \qquad i\in I \qquad (2)$$

$$y_i+n_i+v_i+z_i \quad = t \qquad i\in I \qquad (3)$$

$$\beta\times y_i + g\times v_i \quad \leq 0 \qquad i\in I \qquad (4)$$

$$t - z_i - [y_{i-1}+n_{i-1}] - [y_i+v_i] \quad = 0 \qquad i\in I; i=1..24 \qquad (5)$$

$$z_0 \quad = t \qquad (6)$$

(1) Minimize sum of weighted number of servers of each type.

(2) The availability of bandwidth must be greater than the demand during the ith hour and less than total bandwidth capacity.

(3) The number of servers of each type during the ith hour must be equal to the total number of servers available at the VOD center.

(4) The cost of turning on and off a certain number of servers during the ith hour must be less than an established operational cost

(5) The total number of servers turned ON or OFF at the beginning of the ith hour must equal to the total number of servers at the VOD center subtracted the number of servers OFF during the same hour less the number of servers that were ON during the (i-1)th hour.

(6) At time zero, all the available servers must be OFF

III. Service Level

A. Description of the Service Level

Since the number of requests per hour is a random variable (distributed Pareto with parameters α_i and β_i for each hour i), the questions that now arise are the following:

- How many requests should the VOD serve every hour?
- Can a quantitative approach be derived to measure the capacity of the VOD to meet the number of requests per hour if the latter is understood as the service level? If this is the case, in terms of which parameters could the VOD centers service level be expressed?

One approach that will lead us to answer the previous inquiries is computing the VODs expected profit function for every hour (this is done in order to deal with the random variables that might appear in our initial objective function so that we can make them tractable for subsequent applications). Afterwards, the First Order Necessary Conditions and Second Order Necessary and Sufficient Conditions will be applied to the Expected Profit function in order to find a service level (probability) that is associated to the number of requests per hour that will allow us to maximize this function.

B. Mathematical Model of the Service Level

First, let π be Profit function in terms of the random variable X representing the number of simultaneous requests per hour, and the variable s representing the number of requests that the VOD will be able to handle per hour.

$$\pi(X,s) = \begin{cases} sr-[sc+(X-s)p] & x>s \\ Xr-[sc-(s-X)m] & x\leq s \end{cases} \qquad (7)$$

By definition, the expected value of a function of $g(x)$ is

$$E[g(x)] = \int_{-\infty}^{+\infty} g(x)f(x),dx \qquad (8)$$

Where $f(x)$ is the probability density function that characterizes the random variable x. Applying the above definition to equation (7),then

$$E[\pi(x,s)] = \int_{-\infty}^{+\infty} \pi(x,s)f(x),dx$$

Integrating between $-\infty$ and s, and between s and $+\infty$

$= \int_{-\infty}^{s}\pi(x,s)f(x),dx + \int_{s}^{+\infty}\pi(x,s)f(x)dx$
$= \int_{-\infty}^{s}(xr-sc-(s-x)m)f(x)dx + \int_{s}^{+\infty}(sr-sc-(x-s)p)f(x)dx$
$= \int_{-\infty}^{s}rxf(x)dx - \int_{-\infty}^{s}scf(x)dx - \int_{-\infty}^{s}(s-x)mf(x)dx + \int_{s}^{+\infty}srf(x)dx - \int_{s}^{+\infty}scf(x) - \int_{s}^{+\infty}(x-s)pf(x)dx$
$= r\int_{-\infty}^{s}xf(x)dx - sc\int_{-\infty}^{s}f(x)dx - m\int_{-\infty}^{s}(s-x)f(x)dx + sr\int_{s}^{+\infty}f(x)dx - s - p\int_{s}^{+\infty}(x-s)f(x)dx$
$= r\int_{-\infty}^{s}xf(x)dx - scF(s) - msF(s) + m\int_{-\infty}^{s}xf(x)dx + sr(1-F(s)) - sc(1-F(s)) - p + ps(1-F(s))$

By doing the previous calculations, the final form of the expected profit function is

$$= (r+m)\int_{-\infty}^{s}xf(x)dx - p\int_{s}^{+\infty}xf(x)dx - msF(s) - psF(s)$$

$-rsF(s) + s(p+r-c)$(9)

By differentiating the expected profit function respect to variable s (recalling the fundamental theorem of Calculus developed by Leibniz)

$\frac{\partial E(\pi(x,s))}{\partial s} = (r+m)sf(s)+psf(s)-mF(s)+msf(s)-pF(s)-psf(s)-rF(s)-rsf(s)+(p+r-c)$
Then is obtained

$$\frac{\partial E(\pi(x,s))}{\partial s} = -F(s)(m+p+r)+(p+r-c) \quad (10)$$

Taking the second derivative with respect to s then,

$$\frac{\partial^2 E(\pi(x,s))}{\partial s^2} = -f(s)(m+p+r)$$

For a probability density function defined $f(s) \geq 0$ and for all values of $(s) \geq 0$ then, $-f(s)(m+p+r) \leq 0$ concave function.
It is a good thing that $\frac{\partial^2 E(\pi(x,s))}{\partial s^2}$ is a concave function because the solution is the global maximum, then

$$\frac{\partial^2 E(\pi(x,s))}{\partial s^2} = 0$$

$\frac{c-p-r}{m+p+r} = -F(s)$
$\frac{(r-c)+p}{[(r-c)+p]+(m+c)} = F(s)$
$C_u = (r-c)+p$
$C_o = [(r-c)+p]+(m+c)$

$$s_L = \frac{C_u}{C_u+C_o} \quad (11)$$

s_L is an implicit expression of s^* or the value by which equation (9) is maximized. In fact, equation s_L is known as the service level. Service level is understood as the probability that the number of demanded requests is equal or less than the number of requests that the VOD center can serve at a particular hour. The service level can be also quantified as following

$$\frac{c+m}{[(r-c)+p]+(m+c)} = F(\bar{s})$$

(12)

$$\bar{s} = \frac{C_o}{Cu_C o}$$

Service level $\bar{s}$ can be understood as the probability that the number of demanded requests will exceed the number of requests that the VOD center is able to handle at a certain hour. Recalling the pareto cumulative distribution function

$$F(x) = 1-(\frac{x_m}{x})^k$$

Using the result obtained in (11) and the pareto distribution, then

$$\frac{(r-c)+p}{[(r-c)+p]+(m+c)} = 1-(\frac{x_m}{x})^k$$

$(\frac{x_m}{x})^k = 1-[\frac{(r-c)+p}{(r-c)+p+m+c}]$
$ln(\frac{x_m}{x})^k = ln(1-[\frac{(r-c)+p}{(r-c)+p+m+c}])$
$kln(\frac{x_m}{x}) = ln(1-[\frac{(r-c)+p}{(r-c)+p+m+c}])$
$k(lnx_m - lnx) = ln(1-[\frac{(r-c)+p}{(r-c)+p+m+c}])$
$klnx_m - ln(1-[\frac{(r-c)+p}{(r-c)+p+m+c}]) = klnx$
$lnx_m - ln(\frac{(1-[\frac{(r-c)+p}{(r-c)+p+m+c}])}{k}) = lnx$

Now an expression to calculate the actual number of requests the VOD center has to be able to handle in order to the satisfy the required service level can be obtained as

$$e^{\left(lnx_m - ln\frac{\left(1-\left[\frac{(r-c)+p}{(r-c)+p+m+c}\right]\right)}{k}\right)} = x \approx s(13)$$

C. Validation of the Service Level Model

A simulation will be run in order to validate the mathematical model that was explained before. First, with computational tools a large number should be chosen in order to run the simulation, 10.000 random numbers, with Pareto distribution and parameters alpha and beta, will be generated for each of the time slots in which a working day for the VOD center is divided

$$e^{\left(ln\alpha - ln\frac{RandomNumber \sim U_{(0,1)}}{\beta}\right)} = x(14)$$

Second, if an event in which a generated random number is less than the fixed number of requests during the i_{th} hour (calculated based on a particular service level), then the event will be recorded as a success and the variable ξ_i will have a 1 value; otherwise, the variable ξ_i will get a 0 value. After all 10.000 random numbers have been generated, the service level s_L for the i_{th} hour will be calculated as following:

$$\textbf{Service Level} = \frac{\sum_{i \in I} \xi_i}{10.000}$$

The Visual Basic code to run the simulation is shown as following

A service level of 83.75% has been calculated using equation (11) and the following arbitrary input values
$r = \$8.00$
$c = \$1.00$
$p = \$6.40$
$m = \$1.60$
The results of the simulation are shown in table I.

D. Sensitivity Analysis of the Service Level

The sensitivity analysis for the service level will be done by varying one parameter at a time while leaving the other

```
Sub simulation()

Dim beta(24), alpha(24), servers(24)
iterations = InputBox ("please type the quantity of the required random numbers according a Pareto Distribution with alpha y beta")
For i = 1 To 24
  beta(i) = Cells(2 + i, 6).Value
  alpha(i) = Cells(2 + i, 7).Value
  servers(i) = Cells(2 + i, 8).Value
  Sum = 0
  s = 0
    For j = 1 To iterations
      n = Exp(Log(alpha(i)) - (Log(Rnd()) / beta(i)))
      Cells(2 + i, 9).Value = n
        If n <= servers(i) Then
          Sum = 1 + Sum
        End If
    Next j
  s = Sum / iterations
  Cells(2 + i, 10).Value = s
Next i
End Sub
```

Fig. 1. Sub simulation()

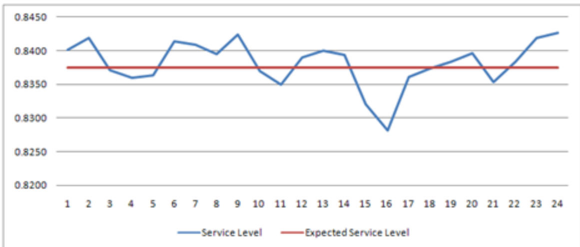

Fig. 2. Comparison between the service level of the simulation and the expected service level

parameters fixed. The base value for the different parameters is shown in table II.

Varying the p parameter

It is found that the service level is very sensitive to the values of p. If the parameter p is varied between the values of $\$0.00$ and $\$8.00$, a variation between 43.75% and 93.75% is obtained in the service level respectively (with a total variation of 50%). The results are shown in table III.

The previous results are expected because it is very logical that if the value of the p parameter increases the service level value should increase as well. In other words, the model compensates the high values of p with a higher service level that eventually will be translated into a greater fixed capacity of requests per hour.

Varying the m parameter

It is found that the service level is very sensitive to the values of r. If the parameter m is varied between the values of $\$0.00$ and $\$8.00$, a variation between 93.75% and 43.75% is obtained in the service level as well (with a total variation of 50%). The results are shown in table IV.

It is clear to see that the variation in the service level due to the changes in the m parameter is proportionally inverse

TABLE I
RESULTS OF THE SIMULATION

I	Hour	Mean of simult. Requests	Std. dev of simult. Requests	Mean + Standard Deviation	Beta 1 K
1	12-1am	374	1021	1395	1.93
2	1-2am	241	501	742	1.88
3	2-3am	178	255	433	1.72
4	3-4am	89	103	192	1.50
5	4-5am	93	151	244	1.79
6	5-6am	103	409	512	1.97
7	6-7am	156	666	822	1.97
8	7-8am	201	256	457	1.62
9	8-9am	319	684	1003	1.88
10	9-10am	527	927	1454	1.82
11	10-11am	699	772	1471	1.42
12	11am-12pm	743	902	1645	1.57
13	12-1pm	1458	1932	3390	1.66
14	1-2pm	1021	2193	3214	1.89
15	2-3pm	856	1228	2084	1.72
16	3-4pm	1672	2055	3727	1.58
17	4-5pm	923	1327	2250	1.72
18	5-6pm	467	1291	1758	1.93
19	6-7pm	584	841	1425	1.72
20	7-8pm	992	2231	3223	1.90
21	8-9pm	642	836	1478	1.64
22	9-10pm	592	901	1493	1.75
23	10-11pm	855	1127	1982	1.65
24	11pm-12am	604	1307	1911	1.89

I	Alpha 1 Xm	Total users served during ith hour with service level p.ri	Service level	Expected Service Level	Difference
1	180.27	463	0.840	0.8375	0.003
2	112.58	297	0.842	0.8375	0.004
3	74.27	215	0.837	0.8375	-0.001
4	29.80	100	0.836	0.8375	-0.002
5	40.98	114	0.836	0.8375	-0.001
6	50.66	128	0.841	0.8375	0.004
7	76.90	194	0.841	0.8375	0.003
8	76.87	237	0.839	0.8375	0.002
9	149.73	393	0.842	0.8375	0.005
10	237.87	645	0.837	0.8375	-0.001
11	208.29	746	0.834	0.8375	-0.003
12	268.84	858	0.839	0.8375	0.001
13	577.63	1731	0.840	0.8375	0.002
14	479.36	1257	0.839	0.8375	0.001
15	357.46	1030	0.832	0.8375	-0.006
16	614.70	1940	0.828	0.8375	-0.009
17	385.90	1111	0.836	0.8375	-0.001
18	225.32	578	0.837	0.8375	-0.000
19	244.38	704	0.838	0.8375	0.001
20	468.71	1223	0.840	0.8375	0.002
21	250.66	759	0.835	0.8375	-0.002
22	254.46	718	0.838	0.8375	0.001
23	337.29	1014	0.842	0.8375	0.004
24	283.88	744	0.843	0.8375	0.006
		AVG.	0.838		

TABLE II
BASE VALUES OF THE "NEWSVENDOR" MODEL

Parameter	Values
r	\$8.00
c	\$1.00
p	\$6.40
m	\$1.60

TABLE III
SENSITIVITY VARYING THE p PARAMETER

$\$p$	Service level
0	0.4375
0.8	0.4875
1.6	0.5375
2.4	0.5875
3.2	0.6375
4	0.6875
4.8	0.7375
5.6	0.7875
6.4	0.8375
7.2	0.8875
8	0.9375

TABLE IV
SENSITIVITY VARYING THE m PARAMETER

$\$m$	Service Level
0	0.9375
0.1	0.8875
0.2	0.8375
0.3	0.7875
0.4	0.7375
0.5	0.6875
0.6	0.6375
0.7	0.5875
0.8	0.5375
0.9	0.4875
1	0.4375

the variation that is obtained in the service level when the parameter p is changed. This is an expected result because the model penalizes low values of m with a lower service level. In fact, this situation will be translated in a small amount of capacity requests per hour.

Varying the c parameter

It is found that the service level is very sensitive to the values of c. If the parameter c is varied between the values of \$0.00 and \$8.00, a variation between 90.00% and 40.00% is obtained in the service level correspondly (with a total variation of 50%). The results are shown in table V.

The previous results are expected since a low value of c will increase the value of the service level. In other words, a lower value of c will represent a higher profit in which situation the model tells that a higher risk should be taken by fixing a greater capacity of requests per hour.

Varying the r parameter

It is found that the service level is not sensitive to the values of r. If the parameter r is varied between the values of \$8.00 and \$24.00, a variation between 83.75% and 85.63% is obtained in the service level is obtained (with a total variation of 1.88%).

TABLE V
SENSITIVITY VARYING THE c PARAMETER

$\$c$	Service Level
0	0.9
0.8	0.85
1.6	0.8
2.4	0.75
3.2	0.7
4	0.65
4.8	0.6
5.6	0.55
6.4	0.5
7.2	0.45
8	0.4

The results are shown in table VI.

TABLE VI
SENSITIVITY VARYING THE r PARAMETER

$\$r$	Service Level
8	0.8375
13.33	0.8469
17.14	0.8511
20	0.8536
22.22	0.8551
24	0.8663

IV. PRICING MODEL

A. Description of Pricing Model

Once the number of available requests has been calculated using the service level quantified for the system, the number of customers in high and low quality can be found (it is clear that the number of requests per hour is equivalent to the number of users per hour). Consequently, the key point is to identify how many customers belong to high and low quality. First, let us assume that the price for both services is the same. Then, it is obvious that more customers would prefer the high quality because their willingness to have a better service. Thus, the number of high quality customers should be higher (see equations (15) and (16)). Furthermore, given the information about the number of requests per hour, a spreadsheet has been done to calculate the number of clients that the system may have during any hour with a known probability of blocking Pb (Pb is the factor that gives the quality of the system), and alpha (α) and beta (β) are calculated for each hour. Afterwards, the same is made for the number of customers needed using the information of the pareto distribution. The mathematical formulation below gives the information about how many customers are needed for each type of service (high or low quality).

B. Mathematical Formulation of the Pricing Model

Given the total number of clients, the spreadsheet calculates the number of customers in high (CHQ) and low quality (CLQ) with the following expression:

$$CHQ = \frac{a \times \varphi_L}{\varphi_H + a \times \varphi_L} \times r_i \tag{15}$$

$$CLQ = \frac{\varphi_H}{\varphi_H + a \times \varphi_L} \times r_i \tag{16}$$

Where:

$$\begin{cases} a = 2 & \text{if } \varphi_H = \varphi_L \\ a = 1 & \text{if otherwise} \end{cases}$$

V. SPREADSHEET EXPLANATION

A. Description of the Spreadsheet

In order to calculate the values of the spreadsheet, the model has to have the following inputs are shown in table VII.

Some of those values are listed in the mathematical model explained before. However, they need a further explanation to understand the model in the spreadsheet.

The yellow cells with red text are inputs that have to be written down directly to the model. For this example, the use of some management tools such as benchmarking have been used to get an idea of the real values for video on demand service.

Also, some assumptions have been made to these inputs. First of all, 1 Kilobit per second [Kbps] equals 1000 bits per second [bps], and not 1024 [bps]. Also, the assumption of 1 Megabit per second [Mbps] equals 1000 [Kbps], and 1 Gigabit per second [Gbps] equals 1000 [Mbps] have been made.

On the other hand, the bandwidth per user in low and high quality has been chosen with what we have considered is a good service in high and low quality. It is supposed that this service is offered with some means of communication directly from the VOD center to the users. Indeed, the bandwidth of 1000 [kbps] for high quality customers is good enough with the correct use of some compression algorithms for video. On the other hand, the bandwidth 200 [kbps] for low quality service is assumed for this model.

The bandwidth capacity of each server is assumed to be 10 [Mbps]. Even though this value can be greater today with the advancement of technology, some other consideration also limits the bandwidth of each server such as the microprocessor speed, the RAM memory, the storage capacity, and so on.

The average revenue per request per hour in dollars, r, is calculated and weighted with the pricing model explained above. In fact, the revenue per user depends on the customers connected on high quality times their price (φ_H) plus the low quality users times their price (φ_L). That expression divided by the total number of possible users results in the weighted revenue per user.

Other calculations are also indicated in the table. However, some parameters have not been mentioned before such as factor and the weights of c, p, and m. First, factor is a parameter that helps to determine the values of p and m in the table. There is an inverse relationship between the loss of goodwill cost in terms of unsatisfied requests per hour (p), and high consumption of resources cost in terms of having idle capacity for 1 extra request (m). In fact, the value of 1 in factor gives the high importance of satisfied customers with little consideration in the cost of the resources. Second, the weight of c helps to determine the percentage of revenue per user that is considered as a tolerable cost per user. In addition, the weight of p helps to determine the value in the cost of unsatisfied request as a function of the revenue per user and factor. Consequently, the weight of m is calculated as 1 weight of p because of the inverse relationship.

B. Spreadsheet Model

The information given in the problem such as mean and standard deviation was useful to calculate the parameters of the pareto distribution. Those values have to be considered for each hour with the use of the following formulation.

Recalling the probability distribution function of the pareto distribution;

$$f(x) = \frac{\beta \times (\alpha)^\beta}{x^{\beta+1}}$$

Let μ be the mean and σ^2 the variance of the pareto distribution.

$$\mu = \frac{\beta\alpha}{\beta - 1}, \beta > 1$$

$$\sigma^2 = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \beta > 2$$

Solving the equations for α and β;

$$\beta = 1 \pm \sqrt{1 - \frac{\mu}{\sigma}^2} \tag{17}$$

$$\alpha = \frac{\beta - 1}{\beta}\mu \tag{18}$$

In the spreadsheet, the positive value of the radical of β is considered because it will generate a positive α. The explanation of the spreadsheet model are shown in table VIII

The column I is the index of the data for each hour of the VOD system. Also, the mean and standard deviations were given in the information of the problem. The values of Alpha and Beta are calculated according to the pareto distribution and the explanation given before.

The number of users served during the ith hour is calculated with alpha, beta, and the probability calculated with the newsvendor model. Once the information of clients that the VOD system has to serve, the values have to be divided in low and high quality users considering the pricing model. In fact, this information is calculated in r_i, w_i and f_i column according are shown in table IX.

The bandwidth can be calculated with the number of customers classified in high and low quality and per hour. For high quality users, the number of clients is multiplied by the bandwidth of each user. The same calculation is done for the low quality users. The required bandwidth is divided by the capacity in kbps of each server. That information rounded up gives the solution of the number of servers needed during each hour according are shown in table X.

The servers are allocated to serve by bandwidth to clients in high and low quality users. Besides, the demanded bandwidth is divided by the capacity of each server to find the total number of servers needed. In fact, there is no distinction between high and low quality clients for each server. Also, the model does not consider the type of videos that are demanded. The model assumes an unlimited service in videos and bandwidth. For instance, if all the clients request a movie like Star wars, the model assumes that all servers can provide the service in low and high quality and they only serve the number of customers according to their capacity in bandwidth.

The column with name available servers to the system compares if the solution of needed servers is below the total number of servers. If not, the column takes the total number of servers as a solution for that particular hour. Once this information has been validated, the new bandwidth (q_i) is calculated.

The solution to this model as an example is shown in the next table. However, at the beginning of the first hour, the model assumes that the system has all servers OFF to be in accordance with the mathematical model are shown in table XI.

In order to verify that our solution meets the criteria established by the mathematical model, the previous solution is considered as an inputare shown in table XII.

The objective function has a cost of $\$11,123$ for this particular problem. This value is calculated with the summation of costs at each hour. In addition, the constraints are put in the model to verify their bounds.

C. Sensitivity Analysis for the Spreadsheet Model using Top Rank

The goal of the sensitivity analysis is to identify the percentage in which certain variables (input) affect the objective function (output). The objective function used for this analysis is the one defined in the mathematical model, and the input variables are varied using a uniform distribution because the probability to have any value on the range is equally likely.
The input variables that were chosen with their corresponding intervals are shown in the table XIII.

The variable "Factor" can only take the values of 1 or 2.
To simulate this particular case of the model, the use of "Palisade" tools and "top rank" program is necessary. This software is available in the lab and runs under Microsoft Excel. In fact, this simulation helps to identify which of those variables affects the objective function.

From the graph, it is shown that the variable that influences the most the objective function is the cost of having one server on, followed by the bandwidth per high quality user, the bandwidth per server, and the bandwidth per low quality user. The other variables are easily identified in the graph and can be modified to see the effect in the objective function.

VI. Conclusion

Expected values of objective functions that involve optimization (maximization or minimization) is a very useful approach that allows the modeler to efficiently deal with random variables while taking into consideration both a long run tendency value and its inherent characteristic of uncertainty. Nevertheless, its applications should be confined to those

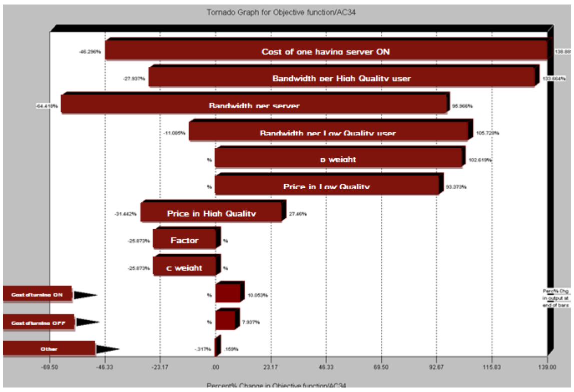

Fig. 3. Tornado graph to visualize the variables that affect the objective function

applications in which the long term scope of the problem is a considered assumption. For example, in our case model of the VOD Center, the application of a service level in order to maximize the expected profit is a very relevant situation in which this model can be efficiently used since this type of business is very profitable with very long cycle lives. In other applications, such as the Newsvendor problem, although the cycle time is not that long (1 or 2 months at last), the application of this model happens to be very useful because of the seasonality of the business whose frequency is being repeated infinitely through the past of time.

The newsvendor model developed for this problem was successfully implemented in the spreadsheet to determine the probability of blocking customers or service level of the system. In fact, the use of parameters according to their importance such as the cost of losing one client or the cost of resources made this newsvendor model a good approach to treat uncertainty on demand. Also, it helps to estimate the probability that was used in the pareto distribution. Therefore, the newsvendor model can be adapted very well to a particular problem when it is necessary to estimate the maximum expected profit.

The calculated service level is very sensitive to changes in the value of the parameters of c (average cost per request per hour in dollars), p (loss of goodwill cost in terms of unsatisfied requests per hour) and m (high consumption of resources cost in terms of having idle capacity for 1 extra). For example, if the value of the c parameter increases then the profit will decrease and, since there is going to be a less chance of winning money, then the service level model will yield a less service level probability. Indeed, this situation, at the same time, will be translated into a smaller number of fixed capacity requests at the beginning of the ith hour.

The value of the service model is not sensitive to changes in the values of the r parameter (average revenue per request per hour in dollars). This situation happens because a unit increment in the r value, while leaving the other parameters fixed, makes a unit increment in both the numerator and denominator of the service level formula. Thus, this situation makes that the service level ratio varies in a very small range.

For example, for a unit increment between the values of r of 10 to 15 (for some arbitrary values of c, p, and m) we obtain results are shown in table XIV.

References

[1] J. Gallego, Y. Myoung, R. Polansky, E. Perez, L. Ntaimo, N. Gautam, Integrating virtualization, speed scaling, and powering on/off servers in data centers for energy efficiency, in IEE Transactions, vol. 45, pp. 1114-1136, 2013.

TABLE VII
LIST OF PARAMETERS USED IN THE SPREADSHEET.

Variable or parameter	Value	Observation
x	Given	Random variable for the number of requests per hour.
s	Calculated	Demand in number of requests per hour value enter as an input in the formulation
o	150	Maximum permitted operating cost per hour.
b	1000	Bandwidth per user in High Quality. [Kbps]
u	200	Bandwidth per user in Low Quality. [Kbps]
h	20	Cost of having 1 server ON during 1 hour.
k	5000000	Total bandwidth capacity when all servers are ON at any hour [Kbps]. FORMULA ($t \times \alpha$).
t	500	Total number of servers at the VOD center.
β	5	Cost of turning ON 1 server at the beginning of the ith hour.
g	3	Cost of turning OFF 1 server at the beginning of the ith hour.
α	10000	Bandwidth capacity per server. [Kbps]
a	1	Constant to weight the quantity of preferred users. FORMULA IF(φ_L=φ_H,2,1)
φ_L	5	Price for hour for a user in the low quality bandwidth capacity.
φ_H	20	Price for hour for a user in high quality bandwidth capacity.
r	8	Average revenue per request per hour in dollars. FORMULA SUMPRODUCT(% clients LQ:% clients HQ,φ_L:φ_H)
c	3.2	Average cost per request per hour in dollars. FORMULA $r \times c$ weight
p	13.6	Lost of good cost in terms of unsatisfied requests per hour. FORMULA IF(Factor=1,$r \times (1 + pweight), rweight)$
m	2.4	High consumption of resources cost in terms of having idle capacity for 1 extra request. FORMULA IF(Factor=2,$r \times (1 + mweight), rweight)$
$Factor$	1	Factor of the importance. 1 High, 2 Low
$c\ weight$	0.4	weight of c
$p\ weight$	0.7	weight of p
$m\ weight$	0.3	weight of m. FORMULA 1-p weight
$F()$	0.766666667	Probability that the number of requests during the ith hour is less than the capacity
$Clients\ in\ low$	80%	Percentage of clients in low quality
$Clients\ in\ high$	20%	Percentage of clients in high quality

TABLE VIII
LIST OF COLUMNS USED IN THE SPREADSHEET MODEL.

Title	Observation
I	Index for the information of each hour
Hour	Hour of the system
Mean of simult.	Requests Expected value of the number of requests per hour
Std. dev of simult.	Requests Standard deviation of the number of requests per hour
Beta 1 K	The parameter or K of the pareto distribution
Alpha 1 Xm	The parameter or Xm of the pareto distribution
Total users served during ith hour with service level p. (ri)	Users served with some probability or service quality
Number of clients in high quality during i-th hour. (wi)	Users in high quality calculated with the pricing model
Number of clients in low quality during ith hour. (fi)	Users in low quality calculated with the pricing model
Bandwidth partition into high quality during ith hour. (hi) [Kbps]	wi * bandwidth per user needed in high quality
Bandwidth partition into low quality during ith hour.. (li) [Kbps]	fi * bandwidth per user needed in low quality
Bandwidth required by Low and High customers during ith hour. qi [Kbps]	hi + li
Servers needed	qi / bandwidth capacity of each server
Available servers to the system	FORMULA IF($qi < t, qi, t$). Checking that needed servers are not greater than available servers
New qi with the available servers for the system	New qi calculated if there are over capacity.
y off to on	Servers to be turned ON
v on to off	Servers to be turned OFF
n on to on	Servers kept in ON
z off to off	Servers kept in OFF

TABLE IX
FIRST CALCULATIONS IN THE SPREADSHEET MODEL.

I	Hour	Mean of simult. Requests	Std. dev of simult. Requests	Beta 1 K
1	12-1am	374	1021	1,93
2	1-2am	241	501	1,88
3	2-3am	178	255	1,72
4	3-4am	89	103	1,50
5	4-5am	93	151	1,79
6	5-6am	103	409	1,97
7	6-7am	156	666	1,97
8	7-8am	201	256	1,62
9	8-9am	319	684	1,88
10	9-10am	527	927	1,82
11	10-11am	699	772	1,42
12	11am-12pm	743	902	1,57
13	12-1pm	1458	1932	1,66
14	1-2pm	1021	2193	1,89
15	2-3pm	856	1228	1,72
16	3-4pm	1672	2055	1,58
17	4-5pm	923	1327	1,72
18	5-6pm	467	1291	1,93
19	6-7pm	584	841	1,72
20	7-8pm	992	2231	1,90
21	8-9pm	642	836	1,64
22	9-10pm	592	901	1,75
23	10-11pm	855	1127	1,65
24	11pm-12am	604	1307	1,89

I	Alpha 1 Xm	Total users served during ith hour with service level p. ri	Number of clients in high quality during i-th hour. wi	Number of clients in low quality during ith hour. fi
1	180,27	384	77	307
2	112,58	245	49	196
3	74,27	174	35	139
4	29,80	79	16	63
5	40,98	93	19	74
6	50,66	107	22	85
7	76,90	161	33	128
8	76,87	189	38	151
9	149,73	325	65	260
10	237,87	529	106	423
11	208,29	579	116	463
12	268,84	681	137	544
13	577,63	1391	279	1112
14	479,36	1038	208	830
15	357,46	835	167	668
16	614,70	1543	309	1234
17	385,90	901	181	720
18	225,32	479	96	383
19	244,38	570	114	456
20	468,71	1010	202	808
21	250,66	609	122	487
22	254,46	584	117	467
23	337,29	815	163	652
24	283,88	614	123	491

TABLE X
SOLUTION OF THE SPREADSHEET MODEL WITH THE NUMBER OF NEEDED SERVERS.

I	Hour	Mean of simult. Requests	Bandwidth partition into high quality during ith hour. hi [Kbps]	Bandwidth partition into low quality during ith hour.. li [Kbps]
1	12-1am	374	77000	61400
2	1-2am	241	49000	39200
3	2-3am	178	35000	27800
4	3-4am	89	16000	12600
5	4-5am	93	19000	14800
6	5-6am	103	22000	17000
7	6-7am	156	33000	25600
8	7-8am	201	38000	30200
9	8-9am	319	65000	52000
10	9-10am	527	106000	84600
11	10-11am	699	116000	92600
12	11am-12pm	743	137000	108800
13	12-1pm	1458	279000	222400
14	1-2pm	1021	208000	166000
15	2-3pm	856	167000	133600
16	3-4pm	1672	309000	246800
17	4-5pm	923	181000	144000
18	5-6pm	467	96000	76600
19	6-7pm	584	114000	91200
20	7-8pm	992	202000	161600
21	8-9pm	642	122000	97400
22	9-10pm	592	117000	93400
23	10-11pm	855	163000	130400
24	11pm-12am	604	123000	98200

I	Bandwidth required by Low and High customers during ith hour. qi [Kbps]	Servers needed	Available servers to the system	New qi with the available servers for the system
1	138400	14	14	140000
2	88200	9	9	90000
3	62800	7	7	70000
4	28600	3	3	30000
5	33800	4	4	40000
6	39000	4	4	40000
7	58600	6	6	60000
8	68200	7	7	70000
9	117000	12	12	120000
10	190600	20	20	200000
11	208600	21	21	210000
12	245800	25	25	250000
13	501400	51	51	510000
14	374000	38	38	380000
15	300600	31	31	310000
16	555800	56	56	560000
17	325000	33	33	330000
18	172600	18	18	180000
19	205200	21	21	210000
20	363600	37	37	370000
21	219400	22	22	220000
22	210400	22	22	220000
23	293400	30	30	300000
24	221200	23	23	230000

TABLE XI
SOLUTION TO THIS PARTICULAR PROBLEM.

I	Hour	Mean of simult. Requests	off to on	on to off	on to on	off to off
1	12-1am	374	14			486
2	1-2am	241	0	5	9	486
3	2-3am	178	0	2	7	491
4	3-4am	89	0	4	3	493
5	4-5am	93	1	0	3	496
6	5-6am	103	0	0	4	496
7	6-7am	156	2	0	4	494
8	7-8am	201	1	0	6	493
9	8-9am	319	5	0	7	488
10	9-10am	527	8	0	12	480
11	10-11am	699	1	0	20	479
12	11am-12pm	743	4	0	21	475
13	12-1pm	1458	26	0	25	449
14	1-2pm	1021	0	13	38	449
15	2-3pm	856	0	7	31	462
16	3-4pm	1672	25	0	31	444
17	4-5pm	923	0	23	33	444
18	5-6pm	467	0	15	18	467
19	6-7pm	584	3	0	18	479
20	7-8pm	992	16	0	21	463
21	8-9pm	642	0	15	22	463
22	9-10pm	592	0	0	22	478
23	10-11pm	855	8	0	22	470
24	11pm-12am	604	0	7	23	470

TABLE XII
OBJECTIVE FUNCTION AND CONSTRAINTS OF THE MATHEMATICAL MODEL.

I	Hour	Objective function	1 constraint LB	1 constraint	1 constraint UB
1	12-1am	350	140000	140000	5000000
2	1-2am	195	90000	90000	5000000
3	2-3am	146	70000	70000	5000000
4	3-4am	72	30000	30000	5000000
5	4-5am	85	40000	40000	5000000
6	5-6am	80	40000	40000	5000000
7	6-7am	130	60000	60000	5000000
8	7-8am	145	70000	70000	5000000
9	8-9am	265	120000	120000	5000000
10	9-10am	440	200000	200000	5000000
11	10-11am	425	210000	210000	5000000
12	11am-12pm	520	250000	250000	5000000
13	12-1pm	1150	510000	510000	5000000
14	1-2pm	799	380000	380000	5000000
15	2-3pm	641	310000	310000	5000000
16	3-4pm	1245	560000	560000	5000000
17	4-5pm	729	330000	330000	5000000
18	5-6pm	405	180000	180000	5000000
19	6-7pm	435	210000	210000	5000000
20	7-8pm	820	370000	370000	5000000
21	8-9pm	485	220000	220000	5000000
22	9-10pm	440	220000	220000	5000000
23	10-11pm	640	300000	300000	5000000
24	11pm-12am	481	230000	230000	5000000

I	Hour	2 constraint	3 constraint LB	3 constraint UB	4 constraint
1	12-1am	500	70	150	0
2	1-2am	500	15	150	0
3	2-3am	500	6	150	0
4	3-4am	500	12	150	0
5	4-5am	500	5	150	0
6	5-6am	500	0	150	0
7	6-7am	500	10	150	0
8	7-8am	500	5	150	0
9	8-9am	500	25	150	0
10	9-10am	500	40	150	0
11	10-11am	500	5	150	0
12	11am-12pm	500	20	150	0
13	12-1pm	500	130	150	0
14	1-2pm	500	39	150	0
15	2-3pm	500	21	150	0
16	3-4pm	500	125	150	0
17	4-5pm	500	69	150	0
18	5-6pm	500	45	150	0
19	6-7pm	500	15	150	0
20	7-8pm	500	80	150	0
21	8-9pm	500	45	150	0
22	9-10pm	500	0	150	0
23	10-11pm	500	40	150	0
24	11pm-12am	500	21	150	0

TABLE XIII
INPUTS VARIABLE THAT AFFECTS THE OBJECTIVE FUNCTION OF THE SYSTEM. THE PARAMETERS OF THE UNIFORM DISTRIBUTION ARE ALSO INDICATED.

Variable	Minimum value	Maximum value
Bandwidth per high quality user	500	2500
Bandwidth per low quality user	50	500
Cost of having 1 server ON	10	30
Cost Turn On	0	10
Cost Turn Off	0	10
Bandwidth per server	5000	20000
Price in low quality	5	30
Price in high quality	10	60
c weight	0	1
p weight	0	1

TABLE XIV
SENSITIVITY VARYING THE R PARAMETER

r	10	11	12	13	14	15
c	5	5	5	5	5	5
p	2	2	2	2	2	2
m	3	3	3	3	3	3
(r-c)+p	7	8	9	10	11	12
p+r+m	15	16	17	18	19	20
ratio	0.47	0.5	0.53	0.56	0.58	0.6
Difference		0.03	0.03	0.03	0.03	0.02

ISSN: 1390-4671