MSC 49J20, 93E20, 68M20, 90B22, 68Q85 DOI: 10.14529/mmp190405

RESOURCE ALLOCATION IN CLOUD COMPUTING VIA OPTIMAL CONTROL TO QUEUING SYSTEMS

A. Madankan[1]***, A. Delavarkhalafi***[1]***, S.M. Karbassi***[1]***, F. Adibnia***[1]
[1]Yazd University, Yazd, Iran
E-mails: A.madankan@stu.yazd.ac.ir, Delavarkh@yazd.ac.ir, Smkarbassi@yazd.ac.ir, Fadib@yazd.ac.ir

We consider resource allocation problem in the cloud computing. We use queuing model to model the process of entering into the cloud and to schedule and to serve incoming jobs. In this paper, the main problem is to allocate resources in the queuing systems as a general optimization problem for controlled Markov process with finite state space. For this purpose, we study a model of cloud computing where the arrival jobs follow a stochastic process. We reduce this problem to a routing problem. In the case of minimizing, cost is given as a mixture of an average queue length and number of lost jobs. We use dynamic programming approach. Finally, we obtain the explicit form of the optimal control by the Bellman equation.

Keywords: cloud computing; multiple queueing system; multiple job classes; stochastic control policy.

Introduction

Resource Allocating (RA) is the process of assigning available resources to the applicants. There are several techniques in parallel computing and sharing the computing resources where cloud computing is the latest technique that shares resources. Cloud computing provides distributed computing over a network that allocates resources to the end-users. Since users are distributed geographically over a wide area, the resources are also distributed. Therefore the cloud management needs to manage resources to allocate resources efficiently. A comprehensive survey on the cloud computing can be found in [1–3].

In [4], resource allocation strategy is a considered as an integrating cloud provider activity for utilizing and allocating resources of cloud environment and a classification of resource allocation strategy is presented. The paper [5] analyses the architecture of the cloud with considering the Quality of Service (QoS). Also, a sequence of M/M/1 and M/M/m queues is considered to model the cloud architecture. The paper [6] proposes different workload types with different characteristics that should be supported by cloud computing, but there is no single solution where it can allocate resources to all imaginable demands optimally.

Stolyar A. [7] uses resource pooling to a generalized switch model, where it was used to study the heavy traffic optimality of the Max-Weight algorithm. Stolyar A. considered a scaled version of queue lengths and time to obtain these results, where it leaded to a regulated Brownian motion. In [8], the authors propose another method in unscaled time for heavy traffic optimality. In addition, this method directly obtains heavy-traffic optimality in steady state. The problem reduces to a routing problem, which is well studied in [9–13].

The optimization and the parameter for evaluating the service in cloud computing are studied in [14]. Also, [14] uses a queuing model to study the performance of services and

develops a method to optimize it. The paper [15] also considers queuing system for cloud computing and studies a routing problem regarding to reduced workload, response time and the average queue length.

In [16], authors use a stochastic model for load balancing and scheduling in cloud computing clusters where jobs arrive according to a stochastic process and request resources such as CPU, memory, and storage space. They show that the performance of JSQ[1] routing and two-choice routing algorithms with Max-Weight scheduling policy have optimal throughput. In [17], the authors introduce a cloud resource allocating algorithm called CRAA/FA[2], which creates a market for cloud resources and makes the resource agents and service agents bargain in that market.

Winston [18] proves that under some conditions, the JSQ strategy is optimal. These conditions are as follows: (1) we do not have any prior knowledge about the job, (2) the jobs are not allowed to preempt, (3) the router sends job immediately when it comes, (4) First-Come-First-Served policy is the only policy for each servers, (5) the job size has exponential distribution. Although, the JSQ is optimal under these conditions, the model that we considered does not meet one or more conditions, therefore we need to study a methodology over these conditions.

There are several types of networks, such as computer multiprocessor networks and communications data networks in the queuing model. Queues are the main part of various network components, such as the input and output buffers of packets. We often want to find the optimal performance of queuing systems, for example, queue length, waiting time, workload, and probabilities of certain states. Since the performance parameters are nonlinear as functions of the arrival and service rates , finding the optimal performance of queues is a difficult problem, whereas efficient control to work loud is one of the highlighted problems of the computer networks.

In this paper, we consider the problem of providing resources in cloud computing. Costumers request resources and set the size of their request, such as a memory, processor power, ..., upon the arrival. The entry point (EP) of the cloud computing puts these requests into a queue and then distribute them to a processing server (PS). These PSs have several kinds of resources where there is a limit of each kind in each PS. Due to this limitation, we have finite number of jobs of each kind on a PS. The simplest architecture of cloud is parallel queues with a router at the front to schedule incoming jobs to servers (see Fig.1).

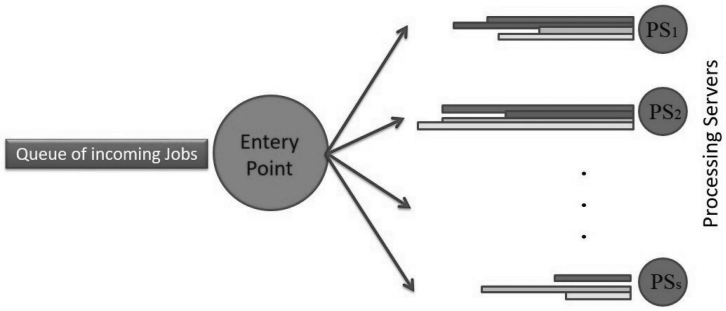

Fig. 1. Queuing model of cloud computing

[1]Join-the shortest-queue

[2]Cloud Resource Allocating Algorithm via Fitness-enabled Auction

For a wide class of criteria, we demonstrate that the optimal control problem can be reduced to the solution of the system of ordinary differential equations. Moreover, the optimal control exists within the class of Markov strategies and therefore can be calculated for each possible state of controlled Markov chain. The existence of the optimal solution was proved in [19]. The papers [20] and [21] consider the general optimization problem for jump Markov process and propose the reduction to a problem with complete information for a wide class of optimal control problems.

1. The Mathematical Modelling and State Equation

In this paper, the considered system is a multi-server queuing network where there is a single entry point (Entry Server) at the front. This entry server works as a load balancer and sends the user job requests to one of the processing servers PS_s, where $s = 1, ..., S$. The purpose of this server is to chose a PS_s to deliver the job requests.

A processing server PS_s can be a node, core or processor representing the physical computational resources of cloud architecture where the services are computed. The selected processing server executes all services that are demanded by an assigned job. The processing servers are identical and are modelled as a M/M/m queuing system. The system has S processing servers (PS) indexed by $s = 1, ..., S$.

Let L be the number of different kinds of resources of each PSs. These resources can be processing power, disk space, memory, etc. Let $R_{l,s}$ be the amount of the resource l of PS_s for $l = 1, ..., L$ and $s = 1, ..., S$. We suppose that each job needs r_l, $l = 1, ..., L$, units of the resource l to complete its service.

Let $\Gamma(t)$ be the set of jobs that arrive at the time slot t and $\gamma(t) = \sum_{i \in \Gamma(t)} D_i$ be the size of $\Gamma(t)$ which means the total time needed for the jobs in Γ. It is clear that $\gamma(t)$ is a stochastic process which is i.i.d. over time slots, with $E[\gamma(t)] = \lambda$.

In each time slot, the Entering Server disputes the arrivals to one of the processing servers. In the time slot t, we use q_s to show the queue length at the server s. For a given server m, denote by Y_s the state of the queue. If we define Y_s^i as the size of the i-th job at the server s then the total backlogged job size is given by $q_s(t) = \sum_i Y_s^i$ which is a function of the state Y_s.

In a Markov chain, one can control arrivals to each queues. This control is called load restriction. For load restriction, control parameter u_s is a probability of accepting the new arrival to the queue of the server s. Now, let all processes be defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ with right-continuous set of complete σ-algebras $\mathcal{F} = \{\mathcal{F}_t\}_{t \geq 0}$ generated by $X_s(t)$. Let the process $\{X_s(t)\}$ be the random variable of queue length of queue of the server s. Similar to [22] and [23], we assume that trajectories of the random processes $X_s(t)$ are piecewise right continuous nonnegative jump-like Markov processes and have finite left limits. Suppose that the state is described by a vector X_s. We write the standard problem statement equation as follows:

$$X_s(t) = \int_0^t \gamma_s(\tau) d\tau - \int_0^t w_s(\tau) d\tau, \tag{1}$$

where

$$\gamma_s(t) = \begin{cases} \gamma(t) & if\ s = s^*(t), \\ 0 & otherwise, \end{cases}$$

and s^* is the server that router chooses it.

2. Optimal Control Problems

2.1. Optimality Criterion

We assume that the classical performance criterion takes the form

$$\Psi(a) = E^a \left\{ c_0(T, X_T) + \int_0^T c(t, X_t, a)dt \right\}, \tag{2}$$

where E^a is the expectation of its argument and the cost functions $c_0(.)$ and $c(.)$ are two continuous nonnegative functions. In the cost functional (2) the time T is the time horizon. The optimal control problem is to find controls $a_t \in A$, where achieves the extremum of the functional (2).

Boel R. and Varaia P. [22] propose a method for using Bellman equation to control a jump-like process. We follow the model presented in [22, 24] to drive Bellman equations.

The function $V(t, x)$ is called the value function of the control problem. The function $V(t, x)$ satisfies the Hamilton–Jacobi–Bellman (HJB) equation (also called the optimality equation)

$$V(t, x) = \min_{\{a_s \in A\}_{t \le s < T}} E^u \left\{ c_0(T, X_T) + \int_t^T c(s, X_s, a_s)ds | X_{t-} = x \right\}. \tag{3}$$

One can find the derivation of equation (3) for a Markov queuing network in [24]. If we define the cost function's difference derivatives as at point (τ, x) $\Upsilon_s^+(\tau, x) = V(t, x + e_s) - V(t, x)$ and $\Upsilon_s^-(\tau, x) = V(t, x - e_s) - V(t, x)$, then the value function $V(t, x)$ can also be presented as the following equation:

$$-\frac{\partial V}{\partial t}(t, x) = \min_{a \in A} \left\{ c(t, x, a) + \sum_{k \in S} a_k \lambda_k \Upsilon_s^+(\tau, x) + \sum_{k \in S} a_k \Upsilon_s^-(\tau, x) \mathbf{1}_{(Q_k > 0)} \right\}.$$

where the initial condition in (4) is

$$V(T, x) \equiv c_0(T, x). \tag{4}$$

3. Controlling Servicing Intensity

In order to find the control to classical queuing system, let us consider the state X_t be queue length which is a jump-like Markov process. Obviously, the state X_t is a birth and death process. Our goal is to minimize the following functional:

$$\Psi(a) = E^a \int_0^T X_t dt, \tag{5}$$

Therefore, we need to find the optimal strategy $\hat{a}_{kt} = \hat{a}(X_t)|X_t = k$ to optimize this functional considered as an accumulated delay functional. Regarding to this functional, the value function (4) takes the following form:

$$-\frac{\partial V}{\partial t}(t, x) = \min_{a \in [0, \mu]} \left\{ x + a \left[V(t, x - 1) - V(t, x) + \lambda \left[V(t, x + 1) - V(t, x) \right] 1_{(x>0)} \right\}, \tag{6}$$

under the initial condition $V(T, x) \equiv 0$.

We consider in the region $[0,T]\times\mathbb{W}$, and $\pi_k(t)=V(t,k)$. With this consideration, (6) reduces to an infinite system of ODEs with respect to the functions $\pi_k(t)$:

$$\begin{aligned}
\frac{d\pi_0(t)}{dt} &= \lambda[\pi_0(t)-\pi_1(t)],\\
\frac{d\pi_1(t)}{dt} &= \min_{a_1\in[0,\mu]}\{a_1[\pi_0(t)-\pi_1(t)]\}\lambda[\pi_1(t)-\pi_2(t)]+1,\\
&\vdots\\
\frac{d\pi_i(t)}{dt} &= \min_{a_i\in[0,\mu]}\{a_i[\pi_{i-1}(t)-\pi_i(t)]\}\lambda[\pi_i(t)-\pi_{i+1}(t)]+i,\\
&\vdots
\end{aligned}\tag{7}$$

where the initial conditions are $\pi_i(T)=0$ for each i.

Also, if we consider $\nabla_i(t)=\pi_i(t)-\pi_{i-1}(t)$ then we can represent the optimal control by the following equation:

$$\hat{a}_i(t)=\mu\mathbf{1}_{(i>0)}a_{-1}\nabla_i(t).\tag{8}$$

And one can find the following expression for $P_0(\tau,i)$ in queueing theory in [25]:

$$\begin{aligned}
P_0(\tau,i) &= e^{-(\lambda+\mu)\tau}\Bigg[\rho^{-i/2}I_i(2\tau\sqrt{\lambda\mu})+\rho^{-(i+1)/2}I_{i+1}(2\tau\sqrt{\lambda\mu})\\
&\quad+(1-\rho)\sum_{k=i+2}^{\infty}\rho^{-k/2}I_k(2\tau\sqrt{\lambda\mu})\Bigg],
\end{aligned}\tag{9}$$

where $\rho=\lambda/\mu$.

Definition 1. *We use the notation $\otimes$ to define the convolution of the functions $\Delta(x)$ and $\Lambda(x)$ as*

$$\Delta(x)\otimes\Lambda(x)=\int_0^t\Delta(x-y)\Lambda(y)dy.$$

Remark 1. By solving the Chapman–Kolmogorov equations

$$\begin{aligned}
\frac{d}{d\tau}P_o(\tau,i) &= \lambda P_{o-1}(\tau,i)-(\mu+\lambda)P_o(\tau,i)+\mu P_{o+1}(\tau,i),\quad o>0,\\
\frac{d}{d\tau}P_0(\tau,i) &= \mu P_1(\tau,i)-\lambda P_0(\tau,i),
\end{aligned}\tag{10}$$

under the initial condition $P_o(0,i)=1_{(o=i)}$, one can find that

$$p^+(\tau,i)=P\{\hat{X}_0=i\}=1-P_0(\tau,i),P_0(\tau,i)=P\{\hat{X}_\tau=0|\hat{X}_0=i\}.$$

Theorem 1. *Let $\hat{X}_\tau$ be a birth and death process with constant birth and death rates λ and μ with respect to $\tau=T-t$. For the problem with functional (5), the optimal control is a Markov control and can be reprresented as*

$$\hat{a}_{it}=\mu 1_{(i)},\tag{11}$$

where the value function is

$$V(t,x) = \sum_{i=0}^{\infty} \pi_i(\tau) 1_{(x=e_i)}, \tag{12}$$

and

$$\pi_i(t) = i\tau - \mu\tau \otimes p^+(\tau, i) + \lambda \frac{\tau^2}{2}. \tag{13}$$

Proof. In order to show the Laplace transform of the function $\Lambda(x)$

$$\Lambda^*(s) = [\Lambda]^* = \int_0^{\infty} e^{-sx} \Lambda(x) dx,$$

we use the symbol $*$. We use backward method of dynamic programming to find the optimal control. Let us define a new time variable τ as $\tau = T - t$ which takes values from the time T to 0. We consider the function $\eta_i(\tau) = \pi_i(t)$, therefore we have

$$\frac{d\eta_i(\tau)}{d\tau} = -\frac{d\pi_i(t)}{dt}, \eta_{i+1}(\tau) - \eta_i(\tau) = \pi_{i+1}(t) - \pi_i(t)$$

and $\eta_i(0) = \pi_i(T)$. Let $\tau = 0$. We present optimal control (8) as follows

$$\hat{a}_i(0) = \mu, \quad \forall i > 0.$$

Hence, system (7) can be rewritten as follows:

$$\begin{aligned}
\frac{d\eta_0(\tau)}{d\tau} &= \lambda(\eta_0(\tau) - \eta_1(\tau)), \\
\frac{d\eta_1(\tau)}{d\tau} &= \mu(\eta_1(\tau) - \eta_0(\tau)) - \lambda(\eta_2(\tau) - \eta_1(\tau)), \\
&\vdots \\
\frac{d\eta_i(\tau)}{d\tau} &= \mu(\eta_i(\tau) - \eta_{i-1}(\tau)) - \lambda(\eta_{i+1}(\tau) - \eta_i(\tau)), \\
&\vdots
\end{aligned} \tag{14}$$

where $\eta_i(0) = 0$. In order to solve (14) we follow the method used in [25].

Consider the generator function to be:

$$\eta(z,t) = \sum_{i=0}^{\infty} \eta_i(\tau) z^i.$$

multiply both sides of (14) by z_i for all $i \geq 0$ and take the sum of the equations:

$$\sum_{i=1}^{\infty} \frac{d\eta_i(\tau)}{d\tau} z^i = \sum_{i=1}^{\infty} \Big(i + \mu\eta_{i-1}(\tau) - (\lambda + \mu)\eta_i(\tau) + \lambda\eta_{i+1} \Big) z^i,$$

i.e.

$$\frac{\partial}{\partial\tau}[\eta(z,\tau)-\eta_0(\tau)]=\sum_{i=0}^{\infty}iz^i+\left[\mu z+\frac{\lambda}{z}-(\lambda+\mu)\right]\eta(z,\tau)-\left[\frac{\lambda}{z}-(\lambda+\mu)\right]\eta_0(\tau)-\lambda z\eta_1(\tau).$$

With respect to the generator function, $\sum_{i=0}^{\infty}iz^i=\frac{z}{(1-z)^2}$, and the first equation of system (14), we have $\frac{d\eta_0(\tau)}{d\tau}=\lambda(\eta_1(\tau)-x\eta_0(\tau))$, therefore we have following linear equation:

$$z\frac{\partial}{\partial\tau}\eta(z,\tau)=[\lambda-z(\lambda+\mu)+\mu z^2]\eta(z,\tau)+[\mu z-\lambda]\eta_0(\tau)+\frac{z^2}{(1-z)^2}. \tag{15}$$

By using the Laplace transform of (15) for τ, we have

$$z[s\eta^*(z,s)-\eta(z,0+)]=[\lambda-z(\lambda+\mu)+\mu z^2]\eta^*(z,s)+[\mu z-\lambda]\eta_0^*(s)+\frac{z^2}{s(1-z)^2}. \tag{16}$$

With regards to the initial condition, we have

$$\eta(z,0+)=\sum_{i=0}^{\infty}\eta_i(0+)z^i=0$$

where (16) means that

$$\eta^*(z,s)=\frac{z^2-(\lambda-\mu z)s(1-z)^2\eta_0^*(s)}{(1-z)^2(zs-\lambda(1-z)+\mu z(1-z))}. \tag{17}$$

If we set the denominator of fraction (17) to be zero then we have the following roots:

$$\begin{aligned}\beta_0 &= 1, \text{ which ia a double root,}\\ \beta_1 &= \frac{1}{2\mu}(s+\lambda+\mu+\sqrt{(\lambda-\mu+s)^2+4\mu s}),\\ \beta_2 &= \frac{1}{2\mu}(s+\lambda+\mu-\sqrt{(\lambda-\mu+s)^2+4\mu s}).\end{aligned}$$

According to the Rouche's theorem [25], only $\beta_2(s)$ is inside the unit circle on the complex plane. Under the condition that poles of function (17) correspond to a root of top of the fraction i.e., numerator turns to 0, the function (17) has bounds. Since for $x=\beta_2(s)$ the top of fraction (17) turns to 0, we have the following equality:

$$\eta_0^*(s)=\frac{\beta_2^2(s)}{s\Big(\lambda(1-\beta_2(s))^2-\mu\beta_2(s)(1-\beta_2(s))^2\Big)}.$$

After some algebra and considering $\varsigma=\mu-\lambda$, we obtain the final representation for the Laplace transform:

$$\eta^*(z,s)=\frac{1}{(1-z)^2s^2}-\frac{1}{(1-z)s^2}-\frac{\vartheta^*(s)}{(\beta_1(s)-z)s^2}-\frac{\varsigma}{(1-z)s^3}, \tag{18}$$

where

$$\vartheta^*(s) = \frac{2\lambda}{s + \varsigma - \sqrt{(s-\varsigma)^2 + 4\mu s}}.$$

It is easy to inverse equation (18) with regarding to table of generator functions [25], to get the following representation for the Laplace transforms $\eta_k(\tau)$:

$$\eta_i^*(s) = \frac{i+1}{s^2} - \frac{1}{s^2} - \frac{\varsigma}{s^3} - \frac{1}{s^2}\vartheta^*(s)[\beta_1(s)]^{-(i+1)}. \tag{19}$$

It can be easily seen that

$$\vartheta^*(s) = \frac{\rho}{\beta_2(s) - \rho},$$

$\beta_1(s)\beta_2(s) = \rho$ follows that

$$\vartheta^*(s)[\beta_1(s)]^{-(i+1)} = \frac{\rho[\beta_1(s)]^{-(i+1)}}{\beta_2(s) - \rho} = \frac{[\beta_1(s)]^{-k}}{\beta_1(s) - 1}. \tag{20}$$

Next, we show that

$$P_0^*(s,i) = \frac{[\beta_1(s)]^{-i}}{\mu(\beta_1(s) - 1)}, \tag{21}$$

and $P_i(\tau, i)$s are solutions to system of equations (10), which these are widely known [25] in queueing theory. We consider $\kappa = 2\tau\sqrt{\lambda\mu}$ and use equation (9) for $P_0(\tau, i)$, which is

$$P_0(\tau,i) \;=\; e^{-(\lambda+\mu)\tau}\left[\rho^{-i/2}I_i(\kappa) + \rho^{-(\lambda+\mu)/2}I_{i+1}(\kappa) + (1-\rho)\sum_{k=i+2}^{\infty}\rho^{-k/2}I_k(\kappa)\right],$$

where $I_k(x)$ is a modified Bessel function for which we have the following simple relation:

$$I_{k+1}(x) = I_{k-1}(x) - \frac{2k}{x}I_k(x).$$

After some simple algebra, we have

$$P_0(\tau,i) = e^{-(\lambda+\mu)\tau}\sum_{k=i}^{\infty}\rho^{-k/2}[I_k(\kappa) - I_{k+2}(\kappa)],$$

and, therefore,

$$P_0(\tau,i) = e^{-(\lambda+\mu)\tau}\frac{1}{\mu}\sum_{k=i+1}^{\infty}k\tau^{-1}\rho^{-k/2}I_k(\kappa).$$

The following Laplace transform is known:

$$[k\tau^{-1}\rho^{-k/2}e^{-(\lambda+\mu)\tau}I_k(\kappa)]^* = [\beta_1(s)]^{-k},$$

where

$$\beta_1(s) = \frac{1}{2\mu}\left(s + \lambda + \mu + \sqrt{(\lambda - \mu + s)^2 + 4\mu s}\right).$$

Then

$$P_0^*(s,i) = \frac{1}{\mu}\sum_{k=i+1}^{\infty}[\beta_1(s)]^{-k}.$$

Comparing (19), (20), and (21), we have

$$\eta_i^*(s) = \frac{i}{s^2} + \frac{\mu}{s^2} - \frac{\varsigma}{s^3}. \tag{22}$$

After taking inverse Laplace transform, we obtain equation (12) for the value function $v_i(t) = \eta_i(\tau)$.

In order to complete the proof, we need to show that the optimal control does not switch for all $t > 0$. To show this, with regards to (8) it is enough to show $\pi_{i+1}(t) - \pi_i(t) > 0$ for $t < T$. To this end, we prove that $\Upsilon_i(\tau) = \eta_{i+1}(\tau) - \eta_i(\tau)$ is an increasing function. Now by applying both (21) and (22), we find a representation for the Laplace transform of the first differentiation of $\Upsilon_i(\tau)$:

$$\Upsilon_i^*(s) = \frac{1}{s^2}\left(1 - [\beta_1(s)]^{-(i+1)}\right).$$

By inverting the expression $[\beta_1(s)]^{-(i+1)}$, we find

$$p_i(\tau) = (i+1)\tau^{-1}\rho^{(i+1)/2}e^{-(\lambda+\mu)\tau}I_{i+1}(\kappa),$$

while the expression for $p_i(\tau)$ for $i = 0$,

$$p_0(\tau) = \frac{1}{\tau\sqrt{\rho}}e^{-(\lambda+\mu)\tau}I_1(\kappa),$$

is known [25]. For $i > 0$ one can easy find that $p_i(\tau)$ is a i-fold convolution of the density function for the busy period

$$p_i(\tau) = \underbrace{p_0(\tau) * \ldots * p_0(\tau)}_{i},$$

i.e., from [25], it is the distribution density function for the generalized busy period, defined as the busy period starting from a moment when the queuing system has i claims, one of which is starting to be serviced. Therefore, we have shown that

$$\frac{d}{d\tau}\Upsilon_i(\tau) = 1 - F_i(\tau), \tag{23}$$

where $F_i(\tau) = \int_0^\tau p_i(x)dx$ is the distribution function for the generalized busy period. $F_i(\tau) < 1$, hence $\Upsilon_i(\tau)$ are strictly increasing functions. Since $\Upsilon_i(0) = 0$, we get that $\Upsilon_i(\tau) > 0$ for $\tau > 0$, and this completes the proof. □

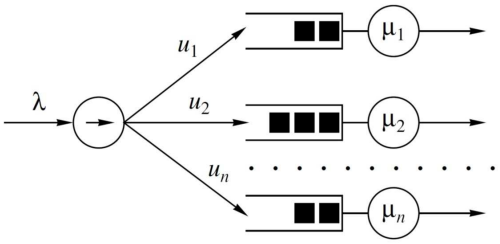

Fig. 2. Router diagram

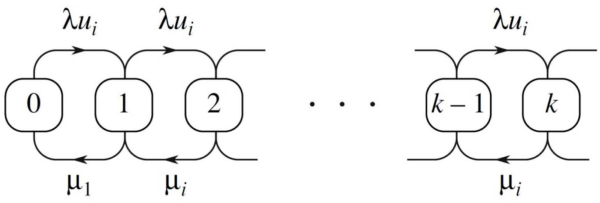

Fig. 3. Transition intensity diagram on the i-th queue

4. Optimal Load Balancing (OLB)

Let us consider the EP which sends arrival jobs with the probability $a_s \in [0,1]$, $\sum_{s=1}^{S} a_s = 1$ to the PS_s. Denote the S-tuple of queue lengths by $X^t = (X_1^t, ..., X_S^t)$ with the set of states $x = (x_1, ..., x_S)$, $x \in [\mathbb{W}]^S$. Under these assumptions, each of the processing servers individually is a queuing system with load restriction. In [26], these processes are studied and it proved that the processes X_s^t are controllable Markov processes. The scheme of these processes is given in Fig. 3.

Our problem is to find optimal strategy $\hat{a}(t,x) = \hat{a}(t, X_t)|X_t = x$ to minimize

$$J(a) = E^a \int_0^T \sum_{s=1}^{S} X_s^t dt. \tag{24}$$

In order to find the optimal Markovian control, the value function is

$$\begin{aligned} \tfrac{\partial V}{\partial t}(t,x)| = \min_{a\in A} \Big\{ & \textstyle\sum_{s=1}^{S} (x_s + \lambda a_s [V(t, x+e_s) - V(t,x)] + \\ & + \mu_s [V(t, x-e_s) - V(t,x)] 1_{(x_s>0)}) \Big\}, \end{aligned} \tag{25}$$

where the initial condition is $V(T,x) \equiv 0$. The other conditions are as follows:

$$A = \left\{ a_s \in [0,1], \sum_{s=1}^{S} a_s = 1 \right\}. \tag{26}$$

Since value function (25) and also conditions (26) are linear with respect to a_s, we can present the optimal strategy by a cost function. We define the set of indices s as

$$J_0(t,x) = \{ j : j = \operatorname{argmin}_{1\le s\le S} \Upsilon_s^+(t,x) \} \tag{27}$$

in which the indices represent minimum of derivatives at the given point (τ, x) by

$$\Upsilon_s^+(\tau,x) = V(t, x+e_s) - V(t,x). \tag{28}$$

Now we can rewrite the optimal control as follows:

$$\hat{a} = 1(m = j_0(t,x)). \tag{29}$$

The idea given in [24] to find the explicit expression of the cost function:

$$V(t,x) = \int_t^T \int_t^s \lambda \sum_{s=1}^{S} \hat{a}_s(\zeta,x) d\theta d\zeta + \int_t^T \sum_{s=1}^{S} x_i^s ds - \sum_{s=1}^{S} \mu_s \int_t^T \int_t^\zeta .p_s^+(\theta - t, x) d\theta d\zeta.$$

Regarding to (26), one can easy find that the vectors of optimal control are vectors with unity as the $j_0(x)$-th element and zeros elsewhere. As a result, the cost function with some replacement such as $\tau = T - t$, $z = \theta - t$, $y = \zeta - t$ can be formed as follows:

$$V(t,x) = \lambda \frac{\tau^2}{2} + \tau \sum_{j=s}^{S} x_m - \sum_{j=s}^{S} \mu_j [\tau \otimes p_s^+(\tau, x)], \tag{30}$$

where $p_s^+(\eta, x) = P\{x_{s\eta} > 0 | x_0 = x, a_\eta = \hat{a}(\eta, x_\eta)\}$ are probabilities of non-zero queue lengths at the given time η, whereas at the initial time, x is the vector of queue lengths. These probabilities depend on the optimal control.

Now if we can write the equations for $p_s^+(\tau, x)$ then we can find the closed representation for the cost function:

$$p_s^+(\tau,x) = 1 - e^{-(\lambda+\mu_s)\tau} \left[\rho_s^{x_s/2} I_{x_s}(\varsigma_s) + \rho_s^{-(x_s+1)/2} I_{x_s+1}(\varsigma_s) + (1-\rho_s) \sum_{s=x_i+2}^{\infty} \rho_i^{-s/2} I_s(\varsigma_s) \right].$$

As a result, the optimal control strategy is to route arrival jobs from users at the given time t to the processing server such that the difference derivative $\Upsilon_j^+(\tau, x)$ is minimal. The Markovian optimal strategy is not uniform. However, the properties of $\Upsilon_j^+(\tau, x)$ for large τ imply the optimal controls depend on time only near the optimal control horizon, i.e., for values of t that are close to T.

5. Simulation

The system that we consider to simulate is a sequence of M/M/1 and M/M/K queuing sytem. The first server, which is called EP, routes incoming jobs to the servers (load balancer). Each server has its own queue to accept the coming new jobs when the server is currently busy. To illustrate the effectiveness of this control, we consider the following assumptions for the system.

- The arrival process is a Poisson process with the arrival intensity λ.
- Servers are considered to have the same service rate μ.
- The router sends jobs immediately as they arrive.

To evaluate the presented methodology, we compare it with Join Shortest Queue (JSQ); which is a common strategy of load balancing. We use Matlab Smulink to simulate and to compare these strategies. We compare them with regard to the waiting time for routing policies for a range of different job-size distributions such as Deterministic and Exponential distributions. The results are shown in Figs 4 – 7.

We consider several strategies to compare OLB method versus JSQ method. We consider two different queuing system, M/M/N and M/D/N, and then for each of them

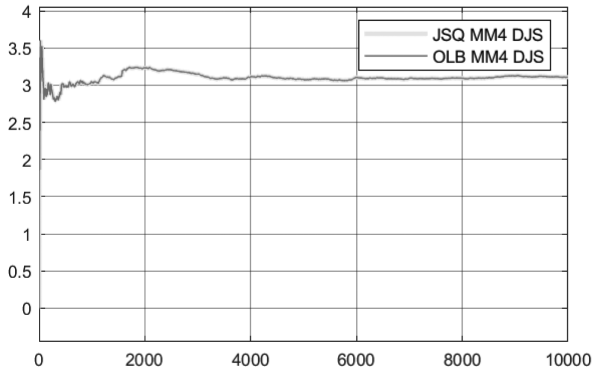

Fig. 4. The waiting time of sampling 10000 entities to the system where the load balancers are **JSQ** and **OLB** and the service rate depends on the service time needed for entities for all servers and the job size follows deterministic distribution

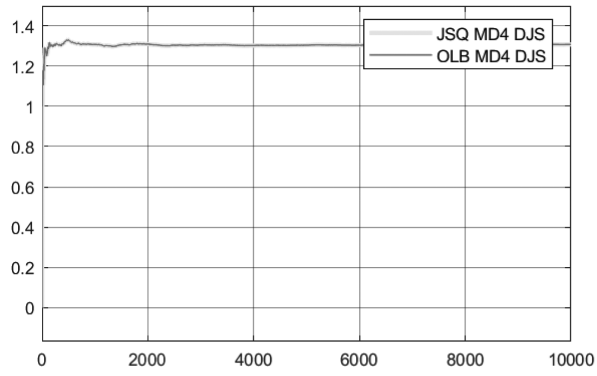

Fig. 5. The waiting time of sampling 10000 entities to the system where the load balancers are **JSQ** and **OLB** and the service rate is constant and equals 1 for all servers and the job size follows deterministic distribution

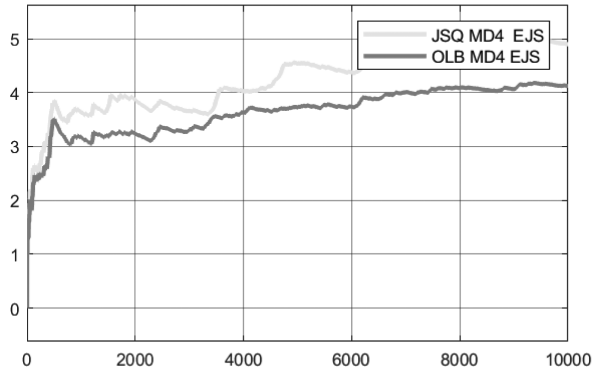

Fig. 6. The waiting time of sampling 10000 entities to the system where the load balancers are **JSQ** and **OLB** and the service rate is constant and equals 1 for all servers and the job size follows exponential distribution

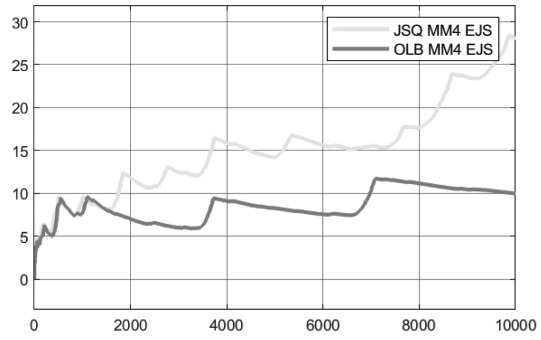

Fig. 7. The waiting time of sampling 10000 entities to the system where the load balancers are **JSQ** and **OLB** and the service rate depends on the service time needed for entities for all servers and the job size follows exponential distribution with mean 8

we consider two different job size distributions, i.e. deterministic job size distribution and exponential job size distribution. The result of comparison for M/D/N queuing system when job size follows deterministic distribution is shown in Fig. 5 and when job size follows exponential distribution is shown in Fig. 6. Also, the result of comparison for M/M/N queuing system when job size follows deterministic distribution is shown in Fig. 4 and when job size follows exponential distribution is shown in Fig. 7.

Figs. 4 and 5 show that the performance of the OLB strategy is similar to the JSQ strategy when the job size distribution is deterministic and the queue length is the number of jobs in the queue. But the OLB strategy shows better performance rather than the JSQ strategy when job size distribution is exponential as shown in Figs. 7 and 6.

Regarding to the results provided by simulation, the OLB strategy has less waiting time and as a consequence less queue length than the JSQ strategy. The reason is because of that the OLB picks the queue to send the new job based on the value function of queue lengths if either queue length is the number of jobs in the queue or the queue length equals to the sum of the job sizes. Despite of OLB, JSQ picks the queue that has the minimum number of jobs between queues. As a result, the OLB strategy is insensitive to the job size distribution, while the JSQ routing policy is sensitive.

Conclusion

We study the resource allocation in a private cloud computing via queuing theory and optimal control. In this paper, in order to analyse the resources allocation in cloud computing, we propose a queueing model for cloud computing and develope a synthetical optimal control method to optimize the performance of services. The considered system is a stochastic system for load balancing and scheduling in cloud computing clusters. We consider the state vector of system to be available to observe at any time. We pose and solve the problem of constructing joint control strategies for the queuing system state. For shortness, we consider a setting with the same type jobs and find the optimal control strategy. In the future we will study our findings in a region containing private and public cloud.

References

1. Foster I., Zhao Y., Raicu I., Lu S. Cloud Computing and Grid Computing 360-Degree Compared. *Grid Computing Environments Workshop*, 2008, vol. 32, pp. 1–10.

2. Armbrust M., Fox A., Griffith R., Joseph A., Katz R., Konwinski A., Lee G., Patterson D., Rabkin A., Stoica I. Above the Clouds: a Berkeley View of Cloud Computing. *Technical Reports*, 2009, p. UCB/EECS-2009-28.

3. Menasce D.A., Ngo P. Understanding Cloud Computing: Experimentation and Capacity Planning. *Computer Measurement Group Conference*, 2009, Dallas, 7–11 December, 11 p.

4. Vinothina V., Sridaran R., Padmavathi Ganapathi, Resource Allocation Strategies in Cloud Computing, *International Journal of Advanced Computer Science and Applications*, 2012, vol. 3, no. 6, pp. 18–22.

5. Vilaplana J., Solsona F., Teixidy I., Mateo J., Abella F., Rius J. A Queuing Theory Model for Cloud Computing. *The Journal of Supercomputing*, 2014, vol. 69, pp. 492–507.

6. Salehpour M., Shahbahrami A., Alleviating Dynamic Resource Allocation for Bag of Tasks Applications in Cloud Computing. *International Journal of Grid and Distributed Computing*, 2012, vol. 5, no. 3, pp. 95–110.

7. Stolyar A. Maxweight Scheduling in a Generalized Switch: State Space Collapse and Workload Minimization in Heavy Traffic. *Applied Probability Journals*, 2004, vol. 14, no. 1, pp. 1–53.

8. Eryilmaz A., Srikant R. Asymptotically Tight Steady-State Queue Length Bounds Implied by Drift Conditions. *Queueing Systems*, 2012, vol. 1, pp. 1–49.

9. Mitzenmacher M. *The Power of Two Choices in Randomized Load Balancing*. PhD thesis, University of California at Berkeley, Harvard University, 1996.

10. Bramson M., Lu Y., Prabhakar B. Randomized Load Balancingwith General Service Time Distributions. *ACM SIGMETRICS Performance Evaluation Review*, 2010, vol. 8, no. 1, pp. 275–286.

11. Hong Chen, Heng-Qing Ye. Asymptotic Optimality of Balanced Routing. *Operation Research*, 2010, vol. 60, no. 1, pp. 163–179. DOI: 10.1287/opre.1110.1011

12. Yu-Tong He, Down D.G. Limited Choice and Locality Considerations for Load Balancing. *Performance Evaluation*, 2008, vol. 65, pp. 670–687. DOI: 10.1016/j.peva.2008.03.001

13. Vvedenskaya N.D., Karpelevich F.I., Queueing System with Selection of the Shortest of Two Queues: an Asymptotic Approach. *Problems of Information Transmission*, 1996, vol. 32, pp. 15–27.

14. Guo L., Yan T., Zhao S., Jiang C. Dynamic Performance Optimization for Cloud Computing Using M/M/m Queueing System. *Journal of Applied Mathematics*, 2014, vol. 2014, article ID: 756592, 8 p. DOI: 10.1155/2014/756592

15. Eisa M., Esedimy E.I., Rashad M.Z. Enhancing Cloud Computing Scheduling Based on Queuing Models. *International Journal of Computer Applications*, 2014, vol. 85, no. 2, pp. 17–23.

16. Siva Theja Maguluri, Srikant R., Lei Ying. Heavy Traffic Optimal Resource Allocation Algorithms for Cloud Computing Clusters. *Perfomance Evolution*, 2014, vol. 81, pp. 20–39. DOI: 10.1016/j.peva.2014.08.002

17. Zuling Kang, Hongbing Wang. A Novel Approach to Allocate Cloud Resource with Different Performance Traits. *Services Computing*, 2013, article ID: 13878874, 6 p. DOI: 10.1109/SCC.2013.109

18. Winston W. Optimality of the Shortest Line Discipline. *Journal of Applied Probability*, 1977, vol. 14, pp. 181–189.

19. Wan C., Davis M. Existence of Optimal Control for Stochastic Jump Processes. *SIAM Journal on Control and Optimization*, 1979, vol. 17, pp. 511–524.

20. Elliott R. A Partially Observed Control Problem for Markov Chains. *Applied Mathematics and Optimization*, 1992, vol. 25, pp. 151–169.

21. Elliott R., Aggoun L., Moore J. *Hidden Markov Models. Estimation and Control.* N.Y., Springer, 1995.

22. Boel R., Varaia P. Optimal Control of Jump Processes. *SIAM Journal on Control and Optimization*, 1977, vol. 15, pp. 92–119.

23. Miller B.M. Optimization of Queuing System via Stochastic Control. *Automatica*, 2009, vol. 45, pp. 1423–1430.

24. Solodyannikov Yu.V. Control and Observation for Dynamical Queueing Networks. *Automation and Remote Control*, 2014, vol. 75, pp. 422–446.

25. Kleinrock L. *Queueing Systems.* N.Y., Springer, 1976.

26. Miller A.B. Using Methods of Stochastic Control to Prevent Overloads in Data Transmission Networks. *Automation and Remote Control*, 2010, vol. 71, pp. 1804–1815.

Received May 5, 2019

УДК 519.872 **DOI: 10.14529/mmp190405**

РАСПРЕДЕЛЕНИЕ РЕСУРСОВ В ОБЛАЧНЫХ ВЫЧИСЛЕНИЯХ С ПОМОЩЬЮ ОПТИМАЛЬНОГО УПРАВЛЕНИЯ СИСТЕМАМИ ПЕРЕДАЧИ

А. Маданкан[1]***, А. Делаварxалафи***[1]***, С.М. Карбасси***[1]***, Ф. Адибния***[1]
[1]Йездский университет, г. Йезд, Иран

Рассматривается задача выделения ресурсов в облачных вычислениях. Мы используем модель очередей для моделирования процесса входа в облако, а также для планирования и обслуживания входящих заданий. Основной задачей, с которой мы сталкиваемся в данной статье, является задача распределения ресурсов в системах массового обслуживания как общая задача оптимизации для управляемого марковского процесса с конечным пространством состояний. Для этой цели мы изучаем модель облачных вычислений, в которой задания по прибытию следуют случайному процессу. Мы сводим эту задачу к задаче маршрутизации. В случае минимизации стоимость выражается через среднюю длину очереди и количество потерянных заданий. Мы используем подход динамического программирования и получаем явную форму оптимального управления по уравнению Беллмана.

Ключевые слова: облачные вычисления; система множественных очередей; несколько классов работы; стохастическая политика управления.

Али Маданкан, аспирант, кафедра «Математические науки», Йездский университет (г. Йезд, Иран), A.madankan@stu.yazd.ac.ir.

Али Делаварxалафи, PhD, доцент, кафедра «Математические науки», Йездский университет (г. Йезд, Иран), Delavarkh@yazd.ac.ir.

Сейед Медиа Карбасси, PhD, профессор, кафедра «Математические науки», Йездский университет (г. Йезд, Иран), Smkarbassi@yazd.ac.ir.

Фазлолла Адибния, PhD, доцент, кафедра «Математические науки», Йездский университет (г. Йезд, Иран), Fadib@yazd.ac.ir.

Поступила в редакцию 5 мая 2019 г.