Modeling and Performance Analysis of Interconnected Servers in a Cloud Computing System with Dynamic Load Balancing

Udo R. Krieger(✉)

Otto-Friedrich-Universität, Fakultät WIAI,
An der Weberei 5, 96047 Bamberg, Germany
udo.krieger@ieee.org

Abstract. We consider the efficiency of dynamic resource pooling and allocation in a cloud computing system offering infrastructure-as-a-service (IaaS). We assume that the demand for service computing by virtual machines (VMs) follows a Poisson load pattern and that the response times of the provided computing services can be classified into several service categories that are governed by exponential service time patterns. A hierarchical, dynamic, class-dependent balancing policy based on a least-loading scheme is applied to provide a uniform utilization among the servers. It is derived from cascaded mutual overflow routing using information on the utilization of VM clusters of similar type on adjacent servers within this resource pool. Regarding the allocation of virtual machines of these different service types to the user demand by a pool of physical servers, we derive a Markovian loss model with adaptive routing induced by cascaded mutual overflow as effective, state-dependent load balancing policy. We determine its basic performance characteristics applying a Markovian fixed-point model. Based on the latter we gain insight on the power of the proposed dynamic load balancing policy among service classes.

Keywords: Cloud computing · IaaS · Performance analysis · Randomized load balancing · Mutual overflow system

1 Introduction

In recent years modern Web services have been provided by cloud computing systems that are hosted by powerful infrastructures in modern data centers like Microsoft Windows Azure or Amazon Web Services. The latter environment constitutes an example of an infrastructure-as-a-service (IaaS) system where users can deploy their virtualized service computing systems on the physical resources of the infrastructure provider. Pooling the virtualized resources offers the chance to follow efficiently the dynamically changing demand curves of the Web services advertised by a service provider, and to satisfy the scalability, elasticity, and resilience requirements of service-oriented computing.

Effective load balancing and resource allocation schemes are an important ingredient of IaaS systems. Recently, new randomized resource assignment policies based on sampling the utilization of physical servers such as the power-d scheme have been studied (cf. [2–4]). Following Mukhopadhyay et al. [3], we can argue that the resource allocation of virtual machines (VMs) on an interconnected cluster of physical servers may be considered as a randomized routing among loss systems hosting the VMs as service units. Then a utilization-oriented resource allocation policy that first senses the status of the physical servers, allocates VMs and tries to optimize the load assignment subject to loss minimization and uniformization constraints can be translated into a state-dependent routing policy of VM requests to the coupled loss systems (cf. [2,3,5]).

Related research [3] has revealed that randomized power-d load balancing schemes for interconnected loss systems are very powerful mechanisms, but their technical realization requires some overhead. Therefore, we propose a much simpler load balancing mechanism that is derived from classical mutual overflow routing (MOR) with state-dependent load splitting (cf. [1]). Its superior performance has already been revealed in the context of circuit-switched networking and guided a worldwide deployment in the switching equipment of two big European manufacturers.

We suppose that a cluster of interconnected physical machines is given which hosts groups of VMs as virtualized computing resources. The latter are divided into different service classes. VM groups of the same class on two neighboring servers are coupled by a MOR scheme. In this way we construct a hierarchical binary load balancing tree among groups of VMs of the same class. Considering a binary tree component, the load is balanced is such a way that a VM request is first offered to the least loaded component with the option to overflow to the other one in case of a fully loaded structure. In this way an effective adaptive local load balancing scheme can be extended to a tree structure.

In this paper we first analyze the derived performance model of a basic component of this IaaS system and describe a fixed-point model of the underlying coupled loss systems with dynamic load balancing in Sect. 2. Then we investigate its basic performance metrics in Sect. 3. Finally, some general conclusions are drawn.

2 Performance Modeling of Dynamic Load Balancing Among Virtualized Server Units in Cloud Computing

We consider a dynamic resource allocation of virtual machines hosted on the interconnected physical servers that is governed by a dynamic load balancing scheme. The latter uses both information about the utilization of the C virtual machines of M different service types and the number of different machines on the interconnected cluster of the $N = 2^k, k \in \mathbb{N}$, physical servers. Motivated by the approach of Mazumdar et al. [3] and references therein, we assume that the server $i \in \mathbb{S} = \{1, \ldots, N\}$ accomodates C_{ij} virtual machines as service units of M different service types $j \in \mathbb{T} = \{1, \ldots, M\}$. Then we suppose that the resulting number $C_{\cdot j} = \sum_{i \in \mathbb{S}} C_{ij}$ of virtualized servers of a certain type $j \in \mathbb{T}$ can be arranged such that $C_{\cdot 1} \leq C_{\cdot 2} \leq \ldots \leq C_{\cdot M}$ holds.

2.1 Dynamic Load Balancing by Mutual Overflow Routing

We propose to apply a dynamic load balancing scheme among two adjacent virtualized clusters of the same type $j \in \mathbb{T}$ on pairs $(i, i+1), i = 2l-1 \in \{1, \ldots, N-1\}, l = 1, \ldots N/2$ of adjacent servers that is derived from a mutual overflow scheme (see Fig. 1, cf. [1]). Then the scheme can be applied in a hierarchical way to the compound of two clustered servers $(i, i+1)$ and $(i+2, i+3)$ as single load balancing block of the mutual overflow scheme and so on. In this way a hierarchical, binary load balancing tree is formed among the service units of each service type.

The offered traffic to service category $j \in \mathbb{T}$ is resulting from a splitting of the overall load to all interconnected servers with rate λ_S by a splitting ratio $p_j^{(T)}$. Each server $i \in \mathbb{S}$ gets a conditional splitting ratio $p_i^{(S|T=j)}$ of the overall traffic of type j. We assume that the offered load is determined by a Poisson stream, hence, all traffic of service demand for virtual machines of a certain class $j \in \mathbb{T}$ at different servers $i \in \mathbb{S}$ is governed by Poisson processes with rates $\lambda_{ij} = \lambda_S \cdot p_j^{(T)} \cdot p_i^{(S|T=j)}$.

2.2 Performance Model of a Two-Server System

Let us now consider two adjacent servers $(i, i+1), i \in \{1, \ldots, N-1\}$ within the server farm as building block of the service infrastructure with hierarchical, dynamic load balancing among a certain service class. For simplicity we assume $i = 1$. We call the latter service computing system 1 and 2, respectively (see Fig. 1). Then we look at a fixed service category $j \in \mathbb{T}$, e.g. $j = 1$, and apply the mutual overflow scheme as basic load balancer among the $N_1 = C_{1j}$ and $N_2 = C_{2j}$ virtual machines on system 1 and 2, respectively, that are serving the incoming service requests of type j. In this case we interpret the system of parallel virtual machines on each server as fully available trunk groups 1 and 2 with Poisson arrival streams and rates λ_1, λ_2 as offered traffic 1 and 2, and exponential service times. Without loss of generality, we assume a common service time with rate $\mu = 1$. If all virtual machines in both service groups are busy, an arriving VM service request is lost in this combined server system. In the binary tree structure this portion of the traffic will overflow to the neighboring block of the server cascade. Thus, a coupled system of two isolated Erlang loss models can be used to describe the basic performance behavior of the coupled virtualized server cluster of fixed type j (cf. [1]).

We propose to use both the information on the relative server capacities N_1, N_2 and the VM utilizations ρ_1, ρ_2 on the different servers 1 and 2 to allocate the incoming requests to the least loaded server. This state-dependent dynamic routing policy within the mutual overflow system between system 1 and 2 is modelled by an adaptive routing with a random splitting of the offered Poisson traffic of class j with common rate:

$$\lambda = \lambda_{ij} + \lambda_{i+1j} = \lambda_S \cdot p_j^{(T)} \cdot (p_i^{(S|T=j)} + p_{i+1}^{(S|T=j)}) \quad i = 1, j \in \mathbb{T}.$$

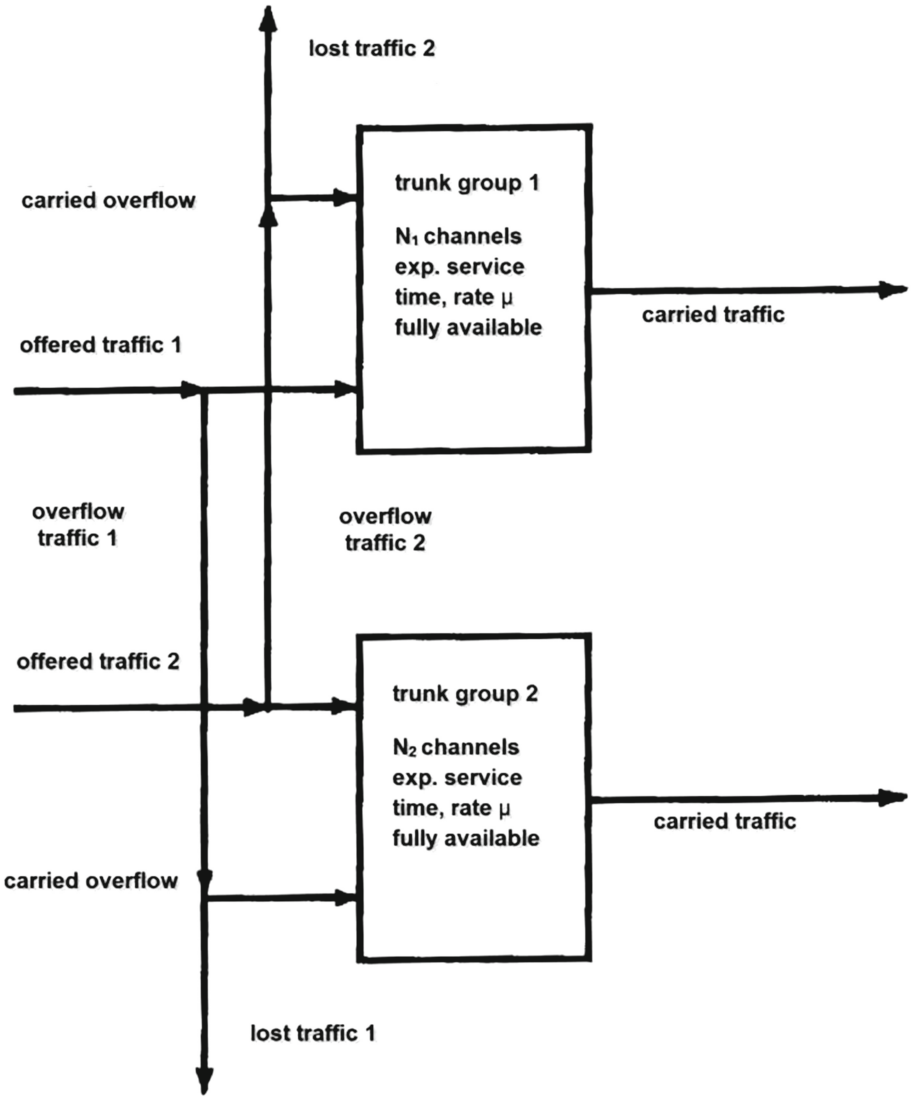


Fig. 1. Principle of mutual overflow routing between two coupled loss systems of virtualized machines that are interpreted as trunk group 1 and 2 (cf. [1]).

Due to the dynamic MOR-type load balancing subject to the server capacities and utilization states, we assume that this offered traffic with rate λ is randomly split up into two portions for system 1 and 2 with rates λ_1 and λ_2. Then the described policy can be modelled by the following splitting probabilities

$$p = \frac{\frac{N_1}{N_1+N_2}(1-\rho_1(N_1))}{1-\rho(N_1+N_2)}, \quad 1-p = \frac{\frac{N_2}{N_1+N_2}(1-\rho_2(N_2))}{1-\rho(N_1+N_2)} \tag{1}$$

of the Poisson stream with rate λ offered to the consecutive virtualized server groups on system 1 and 2. Here $\rho_1(N_1) = E(X_1)/N_1, \quad \rho_2(N_2) = E(X_2)/N_2$ denote the utilization of a single virtual machine on server 1 and 2, respectively, where the state variable X_k records the number of active virtual machines on server $k \in \{1, 2\}$.

$$\rho(N_1 + N_2) = E(X_1 + X_2)/(N_1 + N_2) = \frac{N_1}{N_1 + N_2} \cdot \rho_1(N_1) + \frac{N_2}{N_1 + N_2} \cdot \rho_2(N_2)$$

is the utilization of a single machine of the same type on two adjacent servers that are coupled by mutual overflow load balancing. Here we assume that the systems 1 and 2 will not be fully utilized. The latter case may be simply incorporated by adopting the modified splitting probability

$$p = \frac{\frac{N_1}{N_1+N_2}(1 - \rho_1(N_1) + \varepsilon)}{1 - \rho(N_1 + N_2) + \varepsilon}, \quad 1 - p = \frac{\frac{N_2}{N_1+N_2}(1 - \rho_2(N_2) + \varepsilon)}{1 - \rho(N_1 + N_2) + \varepsilon}$$

for sufficiently small $\varepsilon > 0$. It yields a static splitting according to the relative number of VMs for fully utilized service blocks in the coupled VM clusters.

Then the resulting fresh Poisson stream of type j to system 1 has the rate

$$\lambda_1 = \lambda \cdot p = \lambda \cdot \frac{\frac{N_1}{N_1+N_2}(1 - \rho_1(N_1))}{1 - \rho(N_1 + N_2)} \tag{2}$$

and that one offered to system 2 has the rate

$$\lambda_2 = \lambda \cdot (1 - p) = \lambda \cdot \frac{\frac{N_2}{N_1+N_2}(1 - \rho_2(N_2))}{1 - \rho(N_1 + N_2)} \tag{3}$$

The flow analysis of the interacting server systems coupled by the adaptive routing which is induced by the mutual overflow load balancing now yields the following offered traffic rates L_1 and L_2 on system 1 and 2, respectively:

$$L_1 = L_1(p) = \lambda_1 + \lambda_2 \cdot B_2 = \lambda \cdot p + \lambda \cdot (1 - p) \cdot B_2 = \lambda[1 - (1 - p) \cdot (1 - B_2)] \tag{4}$$

$$L_2 = L_2(p) = \lambda_2 + \lambda_1 \cdot B_1 = \lambda \cdot (1 - p) + \lambda \cdot p \cdot B_1 = \lambda[1 - p \cdot (1 - B_1)] \tag{5}$$

Here the terms B_1 and B_2 denote the arrival-stationary blocking probabilities of system 1 and 2, respectively. If we make the simplifying assumption that the Markov-modulated overflow traffic is again approximated by Poisson streams, the latter coincide with the time-stationary blocking probabilities due to the PASTA property.

2.3 Fixed-Point Approximation of the Blocking Probabilities

The time-stationary blocking probabilities B_1, B_2 are determined by Erlang's formula $B = E(N, A)$ for a pure loss system with offered load A and N service units if we suppose that the overflow streams are Poisson flows. Then they are given by the following quantities

$$B_1 = f_1(N_1, L_1, B_2, p) = {L_1}^{N_1}/[N_1!(\sum_{i=0}^{N_1} {L_1}^i/i!)] = E(N_1, \lambda \cdot [1 - (1-p) \cdot (1 - B_2)])$$

$$B_2 = f_2(N_2, L_2, B_1, p) = {L_2}^{N_2}/[N_2!(\sum_{j=0}^{N_2} {L_2}^j/j!)] = E(N_2, \lambda \cdot [1 - p \cdot (1 - B_1)])$$

if we assume without loss of generality that the service rates of all classes are uniformly given by $\mu = 1$. In [1] it has been revealed that this approximation of the Markov-modulated overflow streams by Poisson flows with identical overflow rates yields a very accurate approximation of the blocking behavior.

For any fixed splitting probability $p \in (0,1)$ this overall Erlang loss model yields a system of fixed-point equations $F = (f_1 \circ f_2, f_2 \circ f_1) : I \to I$ on the compact unit square $I = [0,1]^2$ to determine the blocking probabilities $(B_1(p), B_2(p)) \in [0,1]^2$ by

$$B_1 = B_1(p) = E\left(N_1, \lambda \cdot [1 - (1-p) \cdot (1 - E(N_2, \lambda \cdot [1 - p \cdot (1 - B_1)]))]\right) \quad (6)$$

$$B_2 = B_2(p) = E\left(N_2, \lambda \cdot [1 - p \cdot (1 - E(N_1, \lambda \cdot [1 - (1-p) \cdot (1 - B_2)]))]\right). \quad (7)$$

The existence of a fixed point $B^*(p) = (B_1^*(p), B_2^*(p)) \in [0,1]^2$ is guaranteed by Brower's fixed-point theorem. In [1] it was shown that for fixed $p \in (0,1)$ the independent offered Poisson streams with the positive rates λ_1, λ_2 in (2), (3) determine even a unique fixed point $B^*(p)$ due to the monotonicity of Erlang's loss formula. Then they can be computed by a simple power iteration, e.g. $B_1^{(n)} = [f_1 \circ f_2]^n(B_1^{(0)}), B_1^{(0)} = E(N_1, L_1(p))$. Both blocking terms B_1, B_2 arising as fixed point $B^*(p) = (B_1, B_2)$ in (6), (7) are coupled by the common splitting probability $p = g(B_1, B_2)$ in (1) which depends in a nonlinear manner on the individual server utilizations

$$\rho_1(N_1) = g_1(N_1, L_1, B_1) = E(X_1)/N_1 = \frac{1}{N_1} \cdot L_1 \cdot (1 - B_1)$$

$$\rho_2(N_2) = g_2(N_2, L_2, B_2) = E(X_2)/N_2 = \frac{1}{N_2} \cdot L_2 \cdot (1 - B_2)$$

and, hence, blocking probabilities in both loss systems 1 and 2, and on the server utilization

$$\rho(N_1 + N_2) = \frac{E(X_1 + X_2)}{(N_1 + N_2)} = \frac{N_1}{N_1 + N_2} \cdot \frac{L_1}{N_1} \cdot (1 - B_1) + \frac{N_2}{N_1 + N_2} \cdot \frac{L_2}{N_2} \cdot (1 - B_2)$$

in the overall system. Hence, the splitting probability $p \in (0,1)$ in (1) is determined by the resulting fixed-point equation:

$$p = h(B_1, B_2, p) = \frac{\frac{N_1}{N_1+N_2}(1 - \frac{L_1}{N_1}(1 - B_1))}{1 - [\frac{N_1}{N_1+N_2}\frac{L_1}{N_1}(1 - B_1) + \frac{N_2}{N_1+N_2}\frac{L_2}{N_2}(1 - B_2)]}$$

$$= \frac{N_1 - L_1(p)(1 - B_1(p))}{N_1 - L_1(p)(1 - B_1(p)) + N_2 - L_2(p)(1 - B_2(p))} \quad (8)$$

$$= \frac{N_1 - \lambda \cdot [1 - (1-p) \cdot (1 - B_2(p))](1 - B_1(p))}{\begin{array}{r}(N_1 - \lambda \cdot [1 - (1-p) \cdot (1 - B_2(p))](1 - B_1(p)) \\ +N_2 - \lambda \cdot [1 - p \cdot (1 - B_1(p))](1 - B_2(p)))\end{array}} \quad (9)$$

The corresponding fixed-point model (4), (5), (6), (7) and (9) of the combined splitting-blocking model $X = (B_1(p), B_2(p), p) \in [0,1]^3$ is simple, but analytically complex due to the ratio structure of the term p. It can be solved by a power iteration $p^{(n)} = h(B_1, B_2, p^{(n-1)})$, $n \in \mathbb{N}, p^{(0)} = 0.5$ whose local convergence to a fixed point $X^* = (B_1^*(p^*), B_2^*(p^*), p^*)$ is guaranteed by Brower's fixed-point theorem. Starting with the outcome of our previous analysis [1], we can reveal the dependence on the splitting parameter p^* by a

Steady-State Representation Result. *We consider the basic IaaS component of two servers that host N_1 and N_2 virtual machines (VMs) and serves Poisson arrival streams with offered loads $\lambda_1 = \lambda p^*$ and $\lambda_2 = \lambda(1-p^*)$, respectively. They are coupled by mutual overflow routing combined with a least-load balancing scheme with splitting probability p^* in (1). The steady-state distribution $\Pi = (\pi_{i,j}), \pi_{i,j} = \lim_{t\to\infty} \mathbb{P}(X(t) = (i,j))$ of the resulting ergodic loss model $X = (X_1, X_2) \in [0, N_1] \times [0, N_2] \subset \mathbb{N}^2$ that describes the number of active VMs in server 1 and 2 is determined by a perturbed product-form solution*

$$\pi_{i,j} = \sum_{k=0}^{N_1} c_k(p^*) g_i(\rho_k(p^*), \lambda_1, p^*) \cdot g_j(-\rho_k(p^*), \lambda_2, p^*) + \sum_{l=0}^{N_2} d_l(p^*) g_i(\gamma_l(p^*), \lambda_1, p^*) \cdot g_j(-\gamma_l(p^*), \lambda_2, p^*) \tag{10}$$

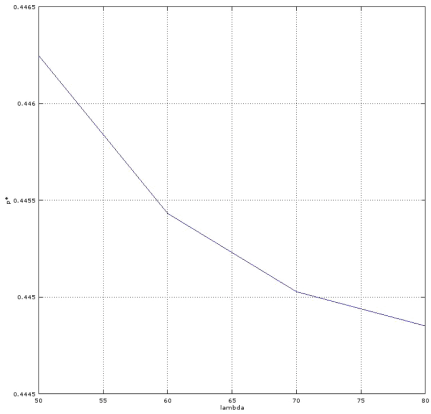
for $0 \le i \le N_1, 0 \le i \le N_2$. The parameters $\rho_k(p^*) \in (-\infty, -1), 1 \le k \le N_1$, are the distinct zeros of the Brockmeyer polynomial $g_{N_1}(1 + \rho(p^*), \lambda_1, p^*)$ and $\gamma_l(p^*) \in (1, \infty), 1 \le l \le N_2$, the distinct zeros of the Brockmeyer polynomial $g_{N_2}(1-\gamma(p^*), \lambda_2, p^*)$ and $\rho_0(p^*) = \gamma_0(p^*) = 0$. The $N_1 + N_2$ coefficients $0 \le k \le N_1, 0 \le l \le N_2$ are the unique solution of a linear system determined by these Brockmeyer polynomials. All terms dependent in unique manner on the existing fixed-point $p^* \in [0,1]$ of the non-linear system (9). They uniquely determine the IaaS blocking probability:

$$\pi_{N_1,N_2} = E(N_1, \lambda_1) \cdot E(N_2, \lambda_2) + \sum_{k=1}^{N_1} c_k(p^*) g_i(\rho_k(p^*), \lambda_1, p^*) \cdot g_j(-\rho_k(p^*), \lambda_2, p^*) + \sum_{l=1}^{N_2} d_l(p^*) g_i(\gamma_l(p^*), \lambda_1, p^*) \cdot g_j(-\gamma_l(p^*), \lambda_2, p^*) \tag{11}$$

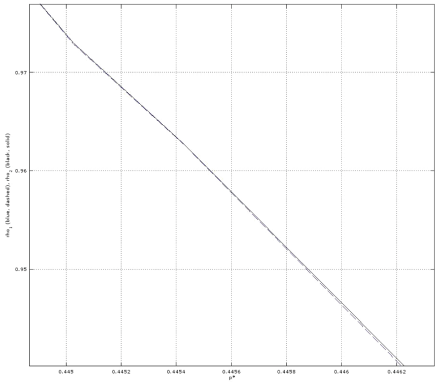
Combining all these coupled loss models arising from the hierarchical binary load balancing tree, we are able to determine by these analytic means the vector of the server-dependent blocking probabilities B_{ij} of class j on server i and the overall blocking probabilities $B_{i,*}, B_{*,j}$ of all servers and classes as fundamental performance metrics of the sketched interconnected service computing system.

3 Performance Study

We have investigated the loss and utilization performance of a basic building block of the proposed IaaS computing model. It consists of two loss systems



(a) Dependence relation of the splitting probability $p^* \in (0,1)$ on the overall rate λ of the arriving Poisson traffic.



(b) Dependence of the utilizationsρ_1, ρ_2 of system 1 and 2 on the fixed-point splitting probability $p^* \in (0,1)$.

Fig. 2. Dependence relations in two interconnected loss systems with $C_1 = 20$ and $C_2 = 25$ VMs and dynamic MOR-based load balancing.

with $C_1 = 20$ and $C_2 = 25$ virtual machines of the same class coupled by mutual overflow routing with a least-load balancing policy in (1). In Fig. 2(a) we depict the dependence relation of the adaptive splitting probability $p^* \in (0,1)$ on the overall rate λ of the arriving Poisson traffic for a heavily loaded system. In Fig. 2(b) we illustrate the dependence of the utilization ρ_1 of system 1 (blue, dashed line) and the corresponding utilization ρ_2 of system 2 (black, solid line) on the fixed-point of the splitting probability $p^* \in (0,1)$.

4 Conclusions

We have investigated the basic component of a new IaaS computing model with dynamic load balancing that is derived from mutual overflow routing (MOR) among two physical servers with virtualized computing resources and a state-dependent splitting of the offered traffic based on a least-load policy. We have first derived a fixed-point model for the MOR scheme on an underlying binary tree of Erlang loss systems that can reflect the state-dependent load balancing policy. Then we have developed an analysis method to compute the loss performance of the service system. The outcome has been demonstrated by a case study of a single class IaaS block illustrating the dependence relation of the traffic splitting and the utilization vector of the system.

References

1. Krieger, U.R.: Analysis of a loss system with mutual overflow. In: Proceedings of ITC-Seminar, Peking, September 1988
2. Maguluri, S.T., Srikant, R., Ying, L.: Stochastic models of load balancing and scheduling in cloud computing clusters. In: Proceedings IEEE INFOCOM (2012)
3. Mukhopadhyay, A., Mazumdar, R., et al.: The power of randomized routing in heterogeneous loss systems. In: Proceedings of ITC, Ghent (2015)
4. Mitzenmacher, M.: The power of two choices in randomized load balancing. IEEE Trans. Parallel Distrib. Syst. **12**, 1094–1104 (2001)
5. Turner, S.R.E.: The effect of increasing routing choice on resource pooling. Probab. Eng. Inf. Sci. **12**, 109–124 (1998)