Autonomous Workload Balancing in Cloud Federation Environments with Different Access Restrictions

A. Amjad*, M. Sharma†, R. Abozariba†, Md. Asaduzzaman*, E. Benkhelifa* and M. Patwary†
*School of Creative Arts and Engineering, Staffordshire University, Stoke-on-Trent, UK
†School of Computing and Digital Technology, Birmingham City University, Birmingham, UK
a.amjad@staffs.ac.uk, r.abozariba@ieee.org, {mak.sharma}, {mohammad.patwary}@bcu.ac.uk, {md.asaduzzaman}, {e.benkhelifa}@staffs.ac.uk

***Abstract*—Although federated cloud computing has emerged as a promising paradigm, autonomous orchestration of resource utilization within the federation is still required to be balanced on the basis of workload assignment at a given time. Such potential imbalance of workload allocation as well as resource utilization may lead to a negative cloudburst within the federation. The analytical models found in the literature do not provide explicit framework to provide dynamic measure of workload requirement within a cloud federation environment. An additional challenge is the adoption of operational restrictions from regulatory body, the federation, or the federation participants. The analytical models presented in this paper have addressed workload balancing within a federated cloud environment under the access control restrictions agreed between federation members. The proposed analytical models provide a closed form solution for access probability and resource utilization at a given time. The analytical results are evaluated at different degree of security within the cloud federation environment and efficiency of the proposed workload balancing models is demonstrated. The proposed models can be used for cloud services dimensioning to handle high computational demand.**

***Index Terms*—Cloud computing, queuing system, Markov process, performance modeling, workload balancing.**

I. Introduction

Cloud computing is currently one of the most popular computation platforms for business and personal use. During the past few years, enterprises have been increasingly fast in moving various applications to the cloud. The technology is bound to provide efficiency and continuous assurance for resiliency and security among other things. Based on the trend towards migrating various workloads to the cloud, service providers are expected to handle sharp surges of workloads to avoid what is called "Negative Cloudburst" [1], [2], [3]. The latter refers to a cloud-based application or infrastructure's inability to efficiently manage resource requirements in abnormal workload scenarios. In this context, many research studies have focused on cloud federation as a potential candidate to address this problem [4], [5]. Cloud federation is implemented to provide better outcomes within a cloud environment with respect to the grade of service (GoS), the risk of system failure in an unpredictable environment and the security issues. Typically, a cloud federation consists of two or more cloud service providers (CSPs). Each CSP in the federation is initially configured with a fixed amount of computational resources (e.g., CPU, memory and I/O). CSPs which participate in a cloud federation have access to more computational resources, which are temporarily underutilized by the other federation participants. However, access to resources vary from one federation arrangement to another according to the initial agreement between the federation participants. Thus, the gain and loss of performance, by being involved in cloud federation, also vary. Moreover, the design of performance models to accurately understand the behavior of various cloud federations is a difficult task.

To study the performance of a cloud data center, a number of performance metrics are proposed in [3]. The author considers different policies and cloud-specific strategies, based on stochastic reward nets (SRNs). An open Jackson network to model cloud platforms is adopted in [6] to determine and measure the quality of service (QoS) guarantees, which the cloud can offer regarding the response time. Resource utilization of consolidated and provisioned multiple Virtual Machines (VMs) are analyzed in [7]. Authors in [8] proposed an analytical model to evaluate the performance of virtualized cloud computing centers. The evaluation is based on several performance indicators that can be utilized under heavy workload scenarios. A trade-off model for cost-effectiveness and performance in cloud environments is proposed in [9] which considers a a system with multi-servers to establish a relationship between system controls and throughput variations. Service response time is considered in [10] to quantify the performance of cloud services within fault recovery context. In [11], the performance of green clouds is analyzed to study the trade-off between QoS metrics and energy consumption. A detailed study of various models for the performance evaluation of cloud services can be found in [12].

The majority of the analytical models presented above, do not provide a precise dynamic performance measure of cloud federation demand patterns. Although the security issues of cloud computing were largely studied [13], [14], their trade-off with respect to computational utilization and access probability

was not discussed in previous research. Moreover, various agreements between federation participants considering different levels of access control and the presence of secure servers were not studied within the context of performance analysis.

In contrast, based on queuing theory we develop two performance models to study the access probability and resource utilization of utility computing. The models allow the cloud federation to be scaled and respond to negative cloudburst by balancing the loads. To evaluate the performance of this approach, we mathematically modeled a cloud federation specifically to analyze the effect of workload fluctuation. In addition, the role of secure servers within a federation environment is also considered in the analysis. The main contributions of this paper are summarized as follows:

- Cloud federation-based models are proposed that seek to balance the workloads between service providers according to specific access control rules.
- Two algorithms are proposed to calculate the efficiency of autonomous workload balancing using the proposed models. A cloud federation is established with up to two levels of access control and the closed form solutions for access probability and utilization are derived.
- The federation models are mathematically developed for performance evaluation. The models provide a simple tool to analyze the cloud federation with different access restrictions in the presence of secure servers.

The rest of the paper is organized as follows: Section II provides the system model. The workload balancing models and the mathematical approach for performance evaluation are presented in Section III. The simulation results are discussed in Section IV. Finally, the conclusions are made in Section V.

II. System Model

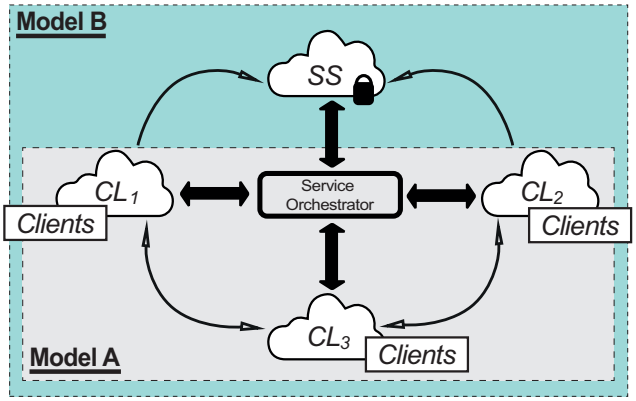

Fig. 1: Autonomous workload balancing models comprising federated cloud providers, secure servers and a service orchestrator.

Two models, Model A and Model B, for workload balancing between service providers in a federation setting are shown in Fig. 1. Model A includes three service providers, CL_1, CL_2 and CL_3, each having its own clients. In order to respond to negative cloudbursts, a service orchestrator is introduced to automate the sharing of resources between the federation participants and to provide interconnectivity. Generally, the federation settings depend on the agreement between the participating cloud providers. A single level of access control is considered in Model A to form a federation. It is assumed that an agreement is made between the federation participants such that CL_3 can access the resources of both CL_1 and CL_2. Moreover, CL_1 and CL_2 have access to the resources of CL_3 in the periods of high workload. It is also assumed in Model A that CL_1 and CL_2 do not have any direct resource sharing agreement, i.e. they cannot access the resources of one another. This is a realistic assumption because some service providers within the federation may not be willing to engage in resource sharing with particular federation participants [15].

The capabilities of Model A are extended by introducing secure servers (represented by *SS*) in Model B to provide a dual level of access control, as shown in Fig. 1. In addition to the aforementioned access control rules considered in Model A, it is assumed that only CL_1 and CL_2 can access the secure servers in Model B; whereas, the secure servers are inaccessible to CL_3. The advantages of Model B compared to Model A are two-fold. Firstly, the access probability of CL_1 and CL_2 is expected to be improved in periods of high workload fluctuations due to the presence of additional resources of secure servers. Secondly, as the secure servers are inaccessible to particular service providers, a secure platform can be facilitated for federation participants to efficiently handle sensitive data. The next section presents the performance analysis of the proposed workload balancing Model A and Model B considering different levels of access control.

III. Performance analysis

Given the system model described in the previous section, our aim is to find the access probability and the resource utilization of the cloud federation. We assume that each of the cloud providers has finite available resources $\mathcal{R}_j$, and requests from clients with Poisson distributed arrivals and rate $\mathcal{A}_j$ and the service rate $\mathcal{S}_j$ $\forall j = 1, 2, 3$. Moreover, the secure servers capacity is denoted by $\mathcal{R}_s$. The access probability is defined by the likelihood of a client to access computational resources provided by the CSP. The overall access probability (P_a) can be written as

$$P_a = 1 - \left[\sum_{j=1}^{n} P_{(b_j)} \times f(\mathcal{A}_j/\mathcal{S}_j) \right] \tag{1}$$

where $f(\cdot)$ is function given by

$$f(\mathcal{A}_j/\mathcal{S}_j) = (\mathcal{A}_j/\mathcal{S}_j) / \sum_{j=1}^{n} (\mathcal{A}_j/\mathcal{S}_j) \tag{2}$$

$P_{(b_j)}$ is the service denial probability at the jth CSP and n is the number of total CSPs. The computational resource utilization of the federation is defined as the ratio of the average number of busy resources and the overall available resources for each CSP in the federation. We calculate this quantity by

Algorithm 1 and 2 for Model A and Model B respectively. In this context, we analyze two access mechanisms as follows.

A. Workload balancing: Model A

Let us assume that CL_1 and CL_2 allow access to their resources from CL_3 when there are resources available. Similarly, when CL_1 (CL_2) receive a resource request and all its $\mathcal{R}_1$ ($\mathcal{R}_2$) are busy it may access resources of CSP CL_3 if there are resources available at the CSP CL_3. This model is formally written as

$$CL_1\left(\mathcal{A}_1, \mathcal{S}_1, (\mathcal{R}_1 + (\mathcal{R}_3 - t_3))\right) \tag{3}$$

$$CL_2\left(\mathcal{A}_2, \mathcal{S}_2, (\mathcal{R}_2 + (\mathcal{R}_3 - t_3))\right) \tag{4}$$

$$CL_3\left(\mathcal{A}_3, \mathcal{S}_3, (\mathcal{R}_3 + (\mathcal{R}_1 + \mathcal{R}_2 - t_1 - t_2))\right) \tag{5}$$

B. Workload balancing: Model B

In this section, in addition to Model A, we consider that CL_1 and CL_2 have a shared secure resources, denoted as SS where it is disconnected from CL_3. This is a practical assumptions since many cloud providers are considering firewall policies for additional security [16], [17].

The model can be formally written as

$$CL_1\left(\mathcal{A}_1, \mathcal{S}_1, (\mathcal{R}_1 + (\mathcal{R}_3 - t_3) + (\mathcal{R}_s - t_2))\right) \tag{6}$$

$$CL_2\left(\mathcal{A}_2, \mathcal{S}_2, (\mathcal{R}_2 + (\mathcal{R}_3 - t_3) + (\mathcal{R}_s - t_1))\right) \tag{7}$$

$$CL_3\left(\mathcal{A}_3, \mathcal{S}_3, (\mathcal{R}_3 + (\mathcal{R}_1 + \mathcal{R}_2 - t_1 - t_2))\right) \tag{8}$$

C. Markov chain and transition rates for Model A and B

Let the state of the Markov chain for model A consist of $\Phi = \{\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3\}$ where $\mathbf{t}_1 = \{t_j | j = 1, 2, 3\}$, $\mathbf{t}_2 = \{t_{3j} | j = 1, 2\}$ and $\mathbf{t}_3 = \{t_{j3} | j = 1, 2\}$. Such that, $(t_3 \leq \mathcal{R}_3)$; $(t_j + t_{3j} \leq \mathcal{R}_j), j = 1, 2$ and $(t_3 + t_{3j} \leq \mathcal{R}_3), j = 1, 2$. Similarly, the state of the Markov chain for model B consist of $\Phi = \{\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \mathbf{t}_4\}$ where $\mathbf{t}_1 = \{t_j | j = 1, 2, 3\}$, $\mathbf{t}_2 = \{t_{3j} | j = 1, 2\}$, $\mathbf{t}_3 = \{t_{j3} | j = 1, 2\}$ and $\mathbf{t}_4 = \{t_{js} | j = 1, 2\}$ where s denotes the secure server. Such that, $(t_3 \leq \mathcal{R}_3)$; $(t_j + t_{3j} \leq \mathcal{R}_j), j = 1, 2$, $(t_3 + t_{3j} \leq \mathcal{R}_3), j = 1, 2$ and $t_{3s} \leq \mathcal{R}_s$.

We now describe how the transition rates for each model are defined. The transition rates of the process for Model A can be defined as $q = \mathcal{A}_3$ satisfies

$$\begin{cases} \mathbf{t}' = \mathbf{t} + \mathbf{w} \text{ or } \mathbf{t}' = \mathbf{t} + \mathbf{w}_{31} \text{ if} \\ (t_3 - \mathcal{R}_3 = 0) \cap (t_1 + t_{31} - \mathcal{R}_1 < 0) \\ \\ \mathbf{t}' = \mathbf{t} + \mathbf{w}_{32} \text{ if} \\ (t_3 - \mathcal{R}_3 = 0) \cap (t_1 + t_{31} - \mathcal{R}_1 = 0) \cap \\ (t_2 + t_{32} - \mathcal{R}_2 < 0) \end{cases}$$

We also have $q = \mathcal{A}_j$ satisfies $\mathbf{t}' = \mathbf{t} + \mathbf{w}_j$ or $\mathbf{t}' = \mathbf{t} + \mathbf{w}_{j3}, j = 1, 2$ if $(t_j - \mathcal{R}_j = 0) \cap (t_3 + t_{13} + t_{23} - \mathcal{R}_3 = 0)$, $q = t_j \mathcal{A}_j$ satisfies $\mathbf{t}' = \mathbf{t} - \mathbf{w}_j, j = 1, 2, 3$, $q = t_{3j} \mathcal{A}_j$ satisfies $\mathbf{t}' = \mathbf{t} - \mathbf{w}_{3j}, j = 1, 2$ and $q = t_{j3} \mathcal{A}_j$ which satisfies $\mathbf{t}' = \mathbf{t} - \mathbf{w}_{j3}, j = 1, 2$.

Algorithm 1: Calculate service denial probability and utilization of federated cloud resources (Model A).

Initialization: Get $\mathcal{A}_j$, $\mathcal{S}_j$, $\mathcal{R}_j$, $\forall\ j = 1, 2, 3$

Phase I — Define the unique invariant distribution $\mathcal{D}_A$ using the transition rates of CL_j, $\forall\ j = 1, 2, 3$

$$\mathcal{D}_A = \mathcal{P}^{-1}(\Omega_A), \text{ where } \mathcal{P} = \sum_{\mathbf{t} \subseteq \Phi} \Omega_A \tag{9}$$

and

$$\Omega_A = \gamma \times \left[\frac{a_1^{(t_1 + t_{13})} \times a_2^{(t_2 + t_{23})}}{(t_1 + t_{13})! \times (t_2 + t_{23})!} \right] \ \forall\ \mathbf{t} \subseteq \Phi \tag{10}$$

where

$$\gamma = \frac{a_3^{(t_3 + t_{31} + t_{32})}}{(t_3 + t_{31} + t_{32})!} \tag{11}$$

Phase II — Calculate the service denial probabilities for CL_j, $\forall j = 1, 2, 3$.

$$P_{(b_j)} = \frac{\sum\limits_{\mathbf{t} \subseteq \mathbf{V}} \Omega_A}{\mathcal{P}} \tag{12}$$

(comment: where the restricted state space $V = \Psi_{A_1}$ for CL_3 and Ψ_{A_2} for CL_1 and CL_2 defined as $\Psi_{A_1} = \{\mathbf{t} \subseteq \Phi \mid (t_3 - \mathcal{R}_3 = 0) \cap (t_{31} + t_1 - \mathcal{R}_1 = 0) \cap (t_{32} + t_2 - \mathcal{R}_2 = 0)\}$ and $\Psi_{A_2} = \{\mathbf{t} \subseteq \Phi \mid (t_3 + t_{j3} - \mathcal{R}_3 = 0) \cap (t_j + t_{3j} - \mathcal{R}_j = 0)\ j = 1, 2.)$

Phase III — Calculate the utilization of computational resources $\mathcal{U}_{(t_j)}$

$$\mathcal{U}_{(t_j)} = \sum_{\mathbf{t}_j \subseteq \Phi} \left[\sum_{i=1}^{3} \mathcal{R}_i \right]^{-1} \left[t_j \times \sum_{\mathbf{t} \subseteq \{\Phi \setminus t_j\}} \mathcal{P}(\mathbf{t}) \right], \tag{13}$$

where $i \in \{j, 3j, j3\}, j = 1, 2$.

The transition rates of the process for Model B can be defined as $q = \mathcal{A}_3$ satisfies

$$\begin{cases} \mathbf{t}' = \mathbf{t} + \mathbf{w} \text{ or } \mathbf{t}' = \mathbf{t} + \mathbf{w}_{31} \text{ if} \\ (t_3 - \mathcal{R}_3 = 0) \cap (t_1 + t_{31} - \mathcal{R}_1 < 0) \\ \\ \mathbf{t}' = \mathbf{t} + \mathbf{w}_{32} \\ \text{if } (t_3 - \mathcal{R}_3 = 0) \cap (t_1 + t_{31} - \mathcal{R}_1 = 0) \cap \\ (t_2 + t_{32} - \mathcal{R}_2 < 0) \end{cases}$$

$q = \mathcal{A}_j$ satisfies

$$\begin{cases} \mathbf{t}' = \mathbf{t} + \mathbf{w}_j \text{ or } \mathbf{t}' = \mathbf{t} + \mathbf{w}_{j3}, j = 1, 2 \text{ if} \\ (t_j - \mathcal{R}_j = 0) \cap (t_3 + t_{13} + t_{23} - \mathcal{R}_3 = 0) \\ \\ \mathbf{t}' = \mathbf{t} + \mathbf{w}_s \text{ if} \\ (t_j - \mathcal{R}_j = 0) \cap (t_3 + t_{13} + t_{23} - \mathcal{R}_3 = 0) \cap \\ (t_{1s} - t_{2s} - \mathcal{R}_s < 0) \end{cases}$$

$q = t_j\mathcal{A}_j$ satisfies $\mathbf{t}' = \mathbf{t} - \mathbf{w}_j, j = 1,2,3$, $q = t_{3j}\mathcal{A}_j$ satisfies $\mathbf{t}' = \mathbf{t} - \mathbf{w}_{3j}, j = 1,2$, $q = t_{j3}\mathcal{A}_j$ satisfies $\mathbf{t}' = \mathbf{t} - \mathbf{w}_{j3}, j = 1,2$ and $q = t_{js}\mathcal{A}_j$ satisfies $\mathbf{t}' = \mathbf{t} - \mathbf{w}_{js}, j = 1,2$. where $\mathbf{w}$ is computed by

$$w_j = \begin{cases} 1, & \text{if } j = i, 1k, 3k \text{ or } ks \\ & \text{where } i = 1,2,3; k = 1,2 \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Given the transition rates of the federation participants when considering Model A and Model B (as described above), let us now present Algorithm 1 and Algorithm 2 that can calculate the service denial probabilities ($P_{(b_j)}$ $\forall j = 1,2,3.$) for each CSP in the federation, respectively. In this context, the service denial probability is defined as the probability that all servers available to the clients of a particular CSP are busy. The service denial probabilities are used to compute the overall access probability (P_a) of the federation, as given in (1), and the utilization of computational resources.

Algorithm 2: Calculate service denial probability and utilization of federated cloud resources (Model B).

Initialization: Get $\mathcal{R}_s$, $\mathcal{A}_j$, $\mathcal{S}_j$, $\mathcal{R}_j$, $\forall\, j = 1,2,3$

Phase I — Define the unique invariant distribution $\mathcal{D}_B$ using the transition rates of CL_j, $\forall\, j = 1,2,3$

$$\mathcal{D}_B = \mathcal{P}^{-1}(\Omega_B), \text{ where } \mathcal{P} = \sum_{\mathbf{t}\subseteq\Phi} \Omega_B \tag{15}$$

and

$$\Omega_B = \gamma \times \left[\frac{a_1^{(t_1+t_{13}+t_{1s})} \times a_2^{(t_2+t_{23}+t_{2s})}}{(t_1+t_{13}+t_{1s})! \times (t_2+t_{23}+t_{1s})!} \right] \;\forall\, \mathbf{t} \subseteq \Phi \tag{16}$$

Phase II — Calculate the service denial probabilities for CL_j, $\forall j = 1,2,3.$

$$P_{(b_j)} = \frac{\sum\limits_{\mathbf{t}\subseteq\mathbf{V}} \Omega_B}{\mathcal{P}} \tag{17}$$

(comment: where the restricted state space $V = \Psi_{B_1}$ for CL_3 and Ψ_{B_2} for CL_1 and CL_2 defined as
$\Psi_{B_1} = \{\mathbf{t} \subseteq \Phi \mid (t_3 - \mathcal{R}_3 = 0) \cap (t_{31} + t_1 - \mathcal{R}_1 = 0) \cap (t_{32} + t_2 - \mathcal{R}_2 = 0)\}$
$\Psi_{B_2} = \{\mathbf{t} \subseteq \Phi \mid (t_3 + t_{j3} - \mathcal{R}_3 = 0) \cap (t_j + t_{3j} - \mathcal{R}_j = 0) \cap \left(\sum_{j=1}^{2} t_{js} - \mathcal{R}_s = 0\right)\; j = 1,2.)$

Phase III — Calculate the utilization of computational resources $\mathcal{U}_{(t_j)}$

$$\mathcal{U}_{(t_j)} = \sum_{\mathbf{t}_j\subseteq\Phi} \left[\left(\sum_{i=1}^{3} \mathcal{R}_i \right) + \mathcal{R}_s \right]^{-1} \left[t_j \times \sum_{\mathbf{t}\subseteq\{\Phi\setminus t_j\}} \mathcal{P}(\mathbf{t}) \right], \tag{18}$$

where $i \in \{j, 3j, j3, js\}, j = 1,2.$

IV. Results

This section demonstrates the performance of the proposed workload balancing models for the aforementioned access control rules. In order to evaluate the utilization behavior of federated cloud resources, a simulation platform is developed by considering the resource capacity of each CSP (i.e. CL_1, CL_2 and CL_3) to be 3 for Model A. On the other hand, the resource capacity of each CSP is assumed as 2 and the resource capacity of secure servers i.e. *SS* is considered as 5 for Model B. A constantly increasing computational demand at CL_1, CL_2 and CL_3 is considered, which is modeled as $\{4, 6, 8, \ldots, 26\}$ for each CSP; resulting in total computational demand at all the federation participants to be $\{12, 18, 24, \ldots, 78\}$. It is pertinent to mention that in the proposed workload balancing Model B which is shown in Fig. 1, *SS* does not have any direct clients. Therefore, in the simulation settings, the workload at *SS* is considered to be the demand from CL_1 and CL_2, where the service orchestrator facilitates the sharing of resources. Considering this simulation setup, Fig. 2 depicts comparative analysis of Model A and Model B in terms of federated cloud resources utilization with growing total computational demand. As a result of increasing computational demand at the federation participants, an exponential growth in the utilization of federated resources is observed. Such utilization behavior justifies the significance of federated cloud resources orchestration for 5G wireless networks, where a continuous growth in computational demand is expected. Moreover, the results show that, initially, the utilization of federated resources in Model A is 2.67% higher than Model B. However, as the total computational demand increases, the difference in resource utilization between Model A and Model B decreases resulting in approximately similar resource utilization behavior at high computational demand scenarios due to increased saturation of resources.

As the assurance of desirable access probability is one of the primary aims within cloud federation environments, it is instructive to analyze the access probability of the proposed workload balancing models with growing computational de-

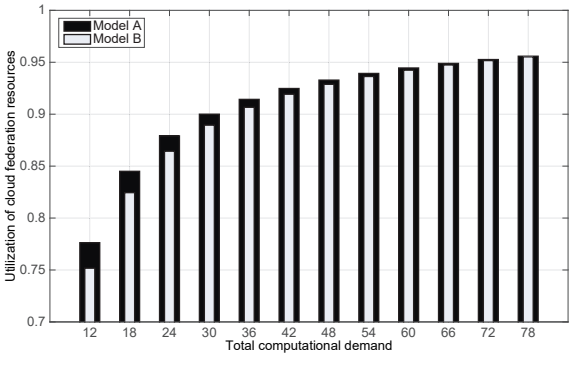

Fig. 2: A comparison of federated cloud resources utilization for Model A and Model B with increasing total computational demand at CSPs.

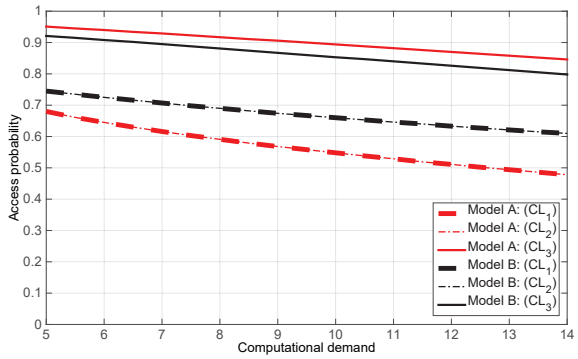

Fig. 3: A comparison of the access probability of CL_1, CL_2 and CL_3 for Model A and Model B with increasing computational demand at CL_3.

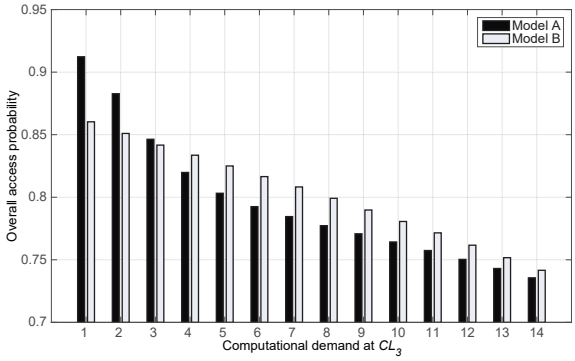

Fig. 4: A comparison of the overall access probability for Model A and Model B with increasing computational demand at CL_3.

mand. Depending on the expected demand in periods of high workload, the proposed analytical models can be utilized to find the amount of resources required within the federation to avoid Service Level Agreements (SLAs) violations. A simulation platform is developed by setting the resource capacity of CL_1, CL_2 and CL_3 to 4, 4 and 3 respectively for Model A. Furthermore, for Model B, the resource capacity of each CSP i.e. CL_1, CL_2, CL_3 is set to 3 and the resource capacity of *SS* is considered as 2. It is worthwhile to mention that the total resource capacity of the federation (i.e. sum of all the federation participants' resource capacity) in the simulation for both Model A and Model B is the same, which leads to a fair comparison. In the simulation setup for both models, a fixed computational demand i.e. 3 is assumed at CL_1 and CL_2, whereas, an increasing computational demand at CL_3 is modeled as $\{5, 6, 7, \ldots, 14\}$. Fig. 3 illustrates the access probability of each CSP for Model A and Model B with increasing computational demand. The results demonstrate superior access probability of CL_3 as compared to other federation participants due to its resource sharing agreements with both CL_1 and CL_2 in Model A and Model B. Moreover, the access probability of CL_3 in Model A is greater than Model B due to the higher resource capacity of CL_1 and CL_2 in Model A compared to Model B. The results of the access probability of CL_1 and CL_2 demonstrate that Model B, with dual level of access control, provides better performance compared to Model A due to the presence of secure servers which are accessible only to CL_1 and CL_2. The results also reveal that both CL_1 and CL_2 have an equal access probability as they both have equal resource capacity and have shared access to the resources of secure servers *SS*.

The simulation parameters presented above for generating the results shown in Fig. 3 are used to analyze the overall access probability of the federated cloud resources. Fig. 4 provides a comparative analysis of the overall access probability for Model A and Model B with growing computational demand at CL_3 and constant demand at CL_1 and CL_2. The results show that Model A has a higher overall access probability compared to Model B at scenarios with lower workload. However, as the computational demand increases beyond a particular level i.e. $(\mathcal{A}_3/\mathcal{S}_3) > 3$, Model B outperforms Model A and results in superior overall access probability of the federated resources. The results also reveal that with growing computational demand, the overall access probability for Model A and Model B has exponential and linear decaying characteristics respectively; thus, justifying the robustness of Model B.

Within the federation, CSPs may experience resource failures due to increased complexity and functionality of the overall system resulting in performance degradation and SLAs violations [18], [19]. Therefore, it is important to analyze the overall access probability of the federated cloud resources by simultaneously considering growing computation demand and resource failures. A simulation platform is created to perform such an analysis for Model A and Model B by assuming decreasing resource capacity of CL_1 and CL_2 i.e. $\{12, 11, 10, \ldots, 1\}$ and fixed workload which is assumed to be 2. In the simulation settings, the resource capacity of CL_3 is assumed as 2 and 1 for Model A and Model B respectively, whereas, the resource capacity of *SS* is considered as 3 for Model B. An increasing computational demand at CL_3 is considered in the simulation setup which is modeled as $\{1, 2, 3, \ldots, 12\}$. Fig. 5 and Fig. 6 show the overall access probability for Model A and Model B respectively with decreasing resources of CL_1 and CL_2 and increasing computational demand at CL_3. The results demonstrate reduction in overall access probability in case of failures and increased computational demand. Moreover, it is observed from the results that Model B shows superior performance compared to Model A in scenarios with resource failures and high workload. The analysis presented in this paper can assist CSPs in gaining insight into access probability and utilization of cloud resources to avoid SLAs violations within federated cloud environments.

V. Conclusion

Cloud computing is becoming increasingly popular for reliable service provision in future generation of wireless

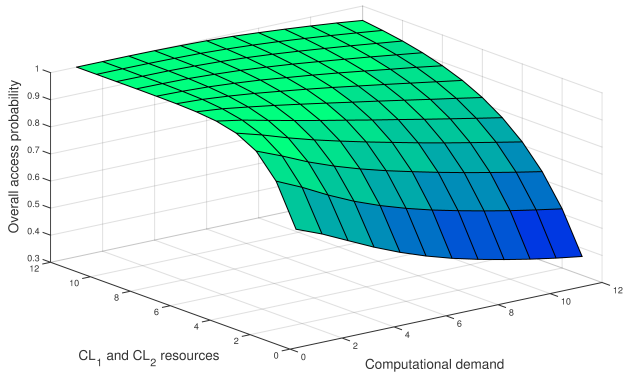

Fig. 5: Overall access probability for the proposed Model A with decreasing resource capacity of CL_1 and CL_2 as well as increasing computational demand at CL_3.

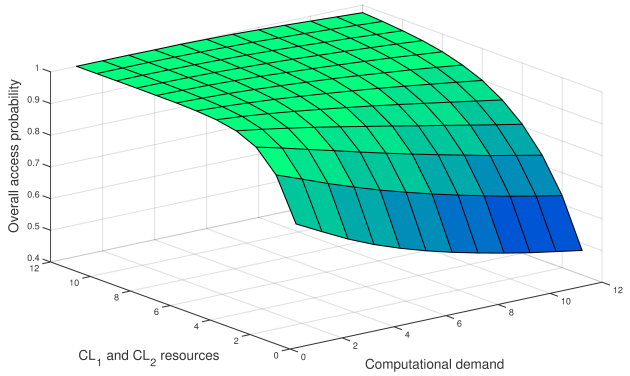

Fig. 6: Overall access probability for the proposed Model B with decreasing resource capacity of CL_1 and CL_2 as well as increasing computational demand at CL_3.

networks due to its performance and low monetary cost. In order to exploit the full potential of cloud computing, the notion of federated cloud resources has emerged as an appealing way to avoid performance degradation and SLAs violations at periods of sharp surges of workload by facilitating resource sharing among different service providers. One of the primary objectives in this context is to balance the workload between federation members for smooth functionality. This paper proposed two models for autonomous workload balancing between service providers, considering different access restrictions. A service orchestrator is incorporated which facilitates single level of access control in Model A, whereas, a dual level of access control is supported in Model B by introducing secure servers. In the proposed models, different resource sharing agreements are considered between the federation participants for performance analysis. Mathematical models are developed to analyze the utilization and access probability of the federated cloud resources. The performance of the proposed workload balancing models is thoroughly analyzed and the findings of the investigation are reported. Simulation results demonstrate the suitability of the proposed models in assisting federation participants to evaluate the effect of workload fluctuations and respond dynamically through autonomous workload balancing. In future work, the authors plan to extend this work by incorporating diverse network authentication layers to support heterogeneous access requests within 5G wireless networks.

REFERENCES

[1] H. Chen, B. An, D. Niyato, Y. C. Soh, and C. Miao, "Workload factoring and resource sharing via joint vertical and horizontal cloud federation networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 3, pp. 557–570, March 2017.

[2] J. Kroß and A. Wolke, "Cloudburst-simulating workload for IaaS clouds," in *IEEE 7th International Conference on Cloud Computing (CLOUD)*. IEEE, 2014, pp. 841–848.

[3] D. Bruneo, "A stochastic model to investigate data center performance and QoS in IaaS cloud computing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, pp. 560–569, 2014.

[4] N. Samaan, "A novel economic sharing model in a federation of selfish cloud providers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 12–21, 2014.

[5] K. Chard and K. Bubendorfer, "Co-operative resource allocation: Building an open cloud market using shared infrastructure," *IEEE Transactions on Cloud Computing*, 2016.

[6] J. Vilaplana, F. Solsona, I. Teixidó, J. Mateo, F. Abella, and J. Rius, "A queuing theory model for cloud computing," *The Journal of Supercomputing*, vol. 69, no. 1, pp. 492–507, 2014.

[7] X. Meng, C. Isci, J. Kephart, L. Zhang, E. Bouillet, and D. Pendarakis, "Efficient resource provisioning in compute clouds via VM multiplexing," in *Proceedings of the 7th international conference on Autonomic computing*. ACM, 2010, pp. 11–20.

[8] H. Khazaei, J. Misic, and V. B. Misic, "Performance of cloud centers with high degree of virtualization under batch task arrivals," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 12, pp. 2429–2438, 2013.

[9] Y.-J. Chiang, Y.-C. Ouyang, and C.-H. Hsu, "Performance and cost-effectiveness analyses for cloud services based on rejected and impatient users," *IEEE Transactions on Services Computing*, vol. 9, no. 3, pp. 446–455, 2016.

[10] B. Yang, F. Tan, and Y.-S. Dai, "Performance evaluation of cloud service considering fault recovery," *The Journal of Supercomputing*, vol. 65, no. 1, pp. 426–444, 2013.

[11] D. Bruneo, A. Lhoas, F. Longo, and A. Puliafito, "Modeling and evaluation of energy policies in green clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 11, pp. 3052–3065, 2015.

[12] Q. Duan, "Cloud service performance evaluation: status, challenges, and opportunities–a survey from the system modeling perspective," *Digital Communications and Networks*, vol. 3, no. 2, pp. 101–111, 2017.

[13] M. A. Khan, "A survey of security issues for cloud computing," *Journal of Network and Computer Applications*, vol. 71, pp. 11–29, 2016.

[14] D. A. Fernandes, L. F. Soares, J. V. Gomes, M. M. Freire, and P. R. Inácio, "Security issues in cloud environments: a survey," *International Journal of Information Security*, vol. 13, no. 2, pp. 113–170, 2014.

[15] J. Abawajy, "Determining service trustworthiness in intercloud computing environments," in *10th International Symposium on Pervasive Systems, Algorithms, and Networks*, Dec 2009, pp. 784–788.

[16] S. Yu, R. Doss, W. Zhou, and S. Guo, "A general cloud firewall framework with dynamic resource allocation," in *IEEE International Conference on Communications (ICC)*. IEEE, 2013, pp. 1941–1945.

[17] M. Liu, W. Dou, S. Yu, and Z. Zhang, "A decentralized cloud firewall framework with resources provisioning cost optimization," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 3, pp. 621–631, 2015.

[18] K. Alhamazani, R. Ranjan, K. Mitra, F. Rabhi, P. P. Jayaraman, S. U. Khan, A. Guabtni, and V. Bhatnagar, "An overview of the commercial cloud monitoring tools: research dimensions, design issues, and state-of-the-art," *Computing*, vol. 97, no. 4, pp. 357–377, 2015.

[19] B. Javadi, J. Abawajy, and R. Buyya, "Failure-aware resource provisioning for hybrid cloud infrastructure," *Journal of Parallel and Distributed Computing*, vol. 72, no. 10, pp. 1318 – 1331, 2012.