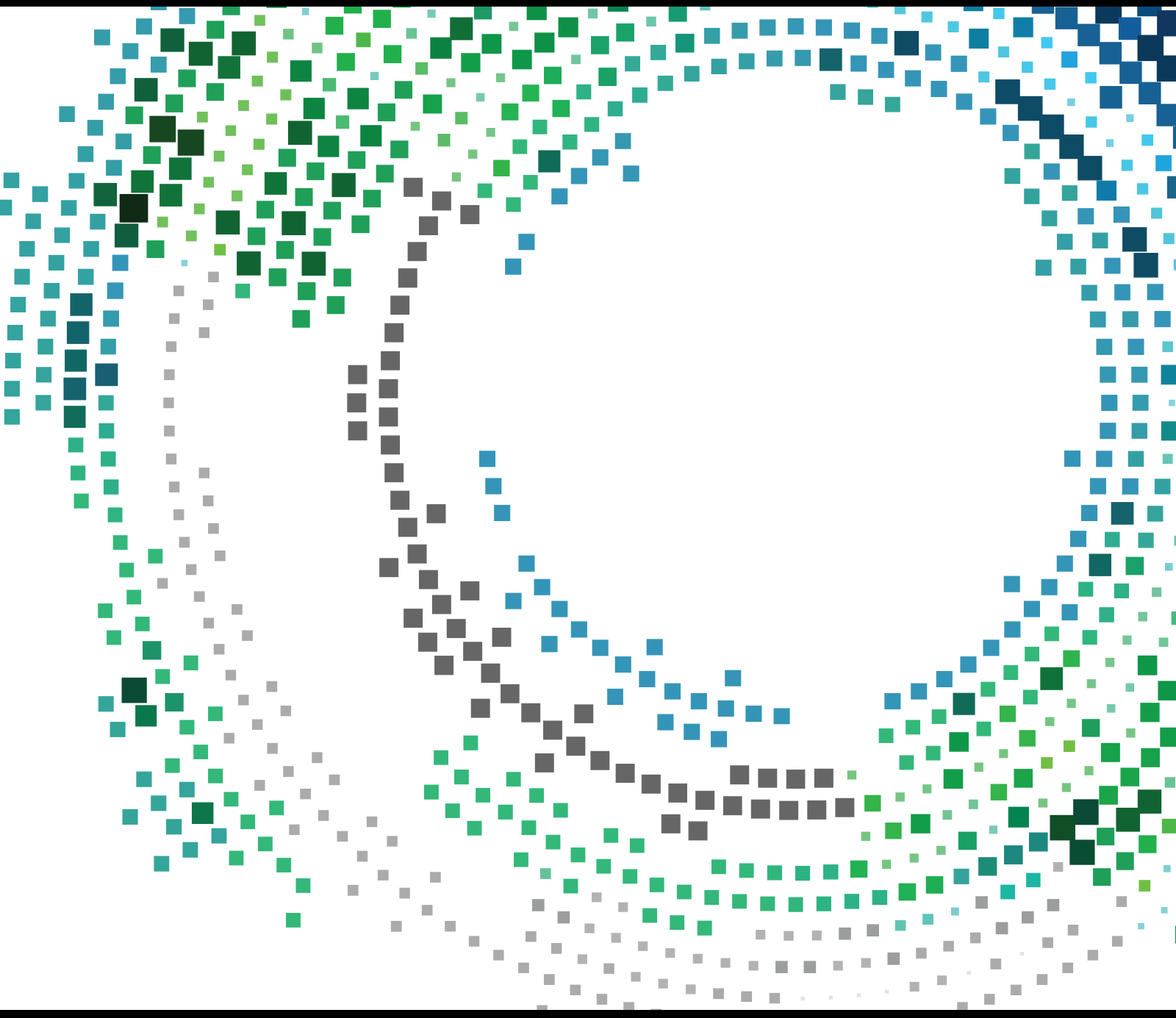
Mobile Information Systems

Mobile Data Cloud Storage in the Era of Intelligence

Lead Guest Editor: Shuai Zhao
Guest Editors: Wei Ni, Cheng Yao, and Li Duan



Mobile Data Cloud Storage in the Era of Intelligence

Mobile Information Systems

Mobile Data Cloud Storage in the Era of Intelligence

Lead Guest Editor: Shuai Zhao
Guest Editors: Wei Ni, Cheng Yao, and Li Duan

Contents

Preserving Resource Handiness and Exigency-Based Migration Algorithm (PRH-EM) for Energy Efficient Federated Cloud Management Systems
R. Karthikeyan, B. Sundaravadivazhagan, Robin Cyriac, Praveen Kumar Balachandran, and S. Shitharth
Research Article (11 pages), Article ID 7754765, Volume 2023 (2023)

A Composite Service Provisioning Mechanism in Edge Computing
Junna Zhang, Xiaoyan Zhao, Yali Wang, Peiyan Yuan, and Xinglin Zhang
Research Article (16 pages), Article ID 9031201, Volume 2022 (2022)

Hindawi
Mobile Information Systems
Volume 2023, Article ID 7754765, 11 pages
https://doi.org/10.1155/2023/7754765

Research Article

Preserving Resource Handiness and Exigency-Based Migration Algorithm (PRH-EM) for Energy Efficient Federated Cloud Management Systems

R. Karthikeyan,[1] **B. Sundaravadivazhagan**,[2] **Robin Cyriac**,[2] **Praveen Kumar Balachandran**,[3] **and S. Shitharth**[4]

[1]*Department of CSE (AI&ML), Vardhaman College of Engineering, Hyderabad, TS 501218, India*
[2]*Department of IT, University of Technology and Applied Sciences, Al Mussanah, Oman*
[3]*Department of EEE, Vardhaman College of Engineering, Hyderabad, TS 501218, India*
[4]*Department of CSE, Kebri Dehar University, Somali, Ethiopia*

Correspondence should be addressed to S. Shitharth; shitharths@kdu.edu.et

Received 30 August 2022; Revised 28 October 2022; Accepted 25 November 2022; Published 17 February 2023

Academic Editor: Shuai Zhao

On-demand computing ability and efficient service delivery are the major benefits of cloud systems. The limitation in resource availability in single data centers causes the extraction of additional resources from the cloud providers group. The federation scheme dynamically increases resource availability in response to service requests. The dynamic increase in resource count leads to excessive energy consumption, maximum cost, and carbon footprints emission. Hence, the reduction of resources is the major requirement to construct the optimized cloud source models for profit maximization without considering energy mix and CO_2. This paper proposes the novel migration method to reduce carbon emissions and energy consumption. The initial stage in the proposed work is the categorization of data centers based on the MIPS and cost prior to job allocation offers scalable and efficient services and resources to the cloud user. Then, the job with the maximum size is allotted to the VM only if its capacity is less than the cumulative capacity of data centers. A novel migration based on overutilized and underutilized levels provides the services to the user even if the particular VM fails. The proposed work offers efficient maintenance of resource availability and maximizes the profit of the cloud providers associated with the federated cloud environment. The comparative analysis of the proposed algorithm with the existing methods regarding the response time, accuracy, profit, carbon emission, and energy consumption assures the effectiveness in a confederated cloud environment.

1. Introduction

Cloud computing is an architectural framework that helps to construct service-based applications and provide the cost-effective outsourcing to dynamic service environments. Among these service models, IAAS allows resource sharing to the customers in virtual machine (VM) form. The resource provisioning scheme greatly impacts on increasing the profit of IAAS providers. The assurance of quality of service (QoS) constraints with the service level agreement (SLA) agreed upon by customers depends on the effective resource management policies. Depending on the resource performance level, the relaxation of QoS constraints is required to process a number of requests simultaneously. The trade-off between the QoS guarantee and the minimum number of requests require a dynamic increase of available resources. The increasing scenario requires the attainment of coordination between the providers that refers to cloud federation. During situations such as the reception of on-demand requests and the absence of idle resources in a federated cloud environment, the provider makes the decision between the spot price increment or spot VM termination. Research studies address the several policies to govern the decision-making process.

The extension of cloud computing introduces the cloud data center as the new research area regarding minimum energy consumption and revenue improvement. The rise of service demand by the data center management unit increases the complexity, size, and energy. The virtualization concept evolution in research studies reduces energy consumption effectively. The social and political pressure forces the analysis of CO_2 emissions to protect the environment from excessive energy consumption. The processor time and the energy consumption are interrelated and such relationship plays a major role in energy consumption measurement. Research studies introduce energy-aware approaches such as power-efficient server prediction, optimal workload placement, and application scheduling. During the implementation of energy-aware models, the estimation of energy consumption is based on the assumption such that the power consumption is proportional to the processor time. But the energy-aware VM model depicts that the processor time is not a required criterion to estimate the power accurately.

The VM management with the established rules maintains the energy and footprint as low as possible. Based on the energy objectives, the ad-hoc VM placement algorithm includes the data center and workload properties to reduce the consumption level. The workload characteristics and the hardware refreshing make the data center as unique which is not appropriate in an ad-hoc VM placement algorithm since the ad-hoc nature prevents upgrading according to the data center properties. Alternatively, the VM allocation algorithm is flexible to address the complexities and dynamic changes. Based on multispecific hardware and SLAs, the energy-aware algorithm provides additional energy savings by using the fine-tuning process. The requirements for the implementation of fine-tuning process are as follows: (i) VM is extensive with new SLA declaration for new service and (ii) adaptability with the data center. The utilization of constraint programming (CP) and its relevant algorithms addresses the user requirements, namely, performance and fault tolerance. But the lack of energy models counteracts the explicit energy-related concerns that play a major role in data center upgradation and capacity improvement.

The adaptability of hosts in data centers with another host in VM depends on the resource profile creation based on load management schemes. Each load management scheme should include heterogeneity, hardware diversity, fluctuations in load patterns, and energy consumption. The centralized approaches proposed in traditional research studies are not enough scalable since they require multiple distributed host monitoring which is complex during the stressful condition of the data center. The increase in resource utilization level efficiently reduces the cloud operational cost. The lack of proper management during virtualization within cloud data centers degrades the cloud operating performance. The rise of the VM migration concept supports effective resource management with the elimination of human supervision. The major categories of migration patterns are live and nonlive. VM live migration enables dynamic load management on data centers and allows the migration of one VM to another VM without suspension of application services compared to nonlive migration. The energy consumption of the VM during the idle state and the data degradation are the major issues in a federated cloud environment. Hence, the research work proposed in this paper reduces energy consumption and improves profit with the novel VM migration model.

(i) The grouping of data centers based on their MIPS and cost prior to job allocation supports the scalable service delivery to the cloud users

(ii) The migration based on the VM status (overutilized/underutilized) proposed in this work is responsible for service provision during the failure condition

(iii) The novel migration and federated cloud environment creation in this paper maintain the resource availability to increase the profit

This paper is organized as follows. Section 2 describes the related works on energy-efficient scheduling and migration techniques for the federated cloud environment. Section 3 discusses the proposed workload-based VM placement/migration and priority-based job allocation algorithm. Section 4 presents the performance analysis of the proposed algorithm over the existing migration techniques. Finally, Section 5 presents the conclusion.

2. Related Work

Efficient service delivery by using the resource provisioning on demand conditions governed by the paradigm called cloud computing. The coordination between the service providers increases resource availability dynamically and it plays a major role in the construction of a federated cloud environment. Toosi et al. [1] illustrated the policies to increase resource utilization and profit. The providers who participated in the federated cloud environment had diverse choices that make the decision policies not suitable. The large-size data transmitted in the federated cloud environment consumed more time. Celesti et al. [2] analyzed the potentiality of cloud federated architectures and proposed a more efficient cloud service provisioning strategy with the consideration of the WEB TV test case. The challenges observed from satellite applications are the consideration of application deployment and the migrating services for spanning clouds. Paraiso et al. [3] presented the federated multicloud environment that addressed three foundations: open service model, configurable architecture, and infrastructure service. Cloud federations depend on the aggregation of IAAS providers with their own capabilities. Kertesz et al. [4] introduced the cloud management solution based on the utilization level of cloud brokers. They utilized an integrated monitoring approach that enabled provider selection enhancement and cloud service executions. The provision of elastic on-demand computing ability to support the service delivery. Petri et al. [5] evaluated the federation establishment and determined the impact of policies on the system status by using comet cloud-based implementation. The utilization of special gateways supported the comet cloud implementation. The large deployment of data centers

to deliver various services to the user consumed more energy and power. Hence, development techniques were required to construct eco-friendly cloud computing.

Wadhwa and Verma [6] proposed the carbon-di-oxide (CO_2) emission rate-based new technique to reduce energy consumption in data centers. The distributed data centers used in this technique have different energy sources and carbon footprint rates that affected the VM placement. Wadhwa and Verma [7] proposed the carbon-efficient VM placement and migration (CEPM) algorithm for the optimization of the VM placement problem. The current utilization level of servers was considered to improve efficiency. The metrics such as efficiency/scalability of data center and the performance of hosted applications are dependent on resource allocation. Ferdaus et al. [8] presented the necessary background and characteristics of various components for VM placement and migration techniques. The VM management by rule-based approaches was responsible for the minimization of energy consumed by each data center. The rule-based approaches were not applicable for additional energy saving. Dupont et al. [9] proposed the energy-aware VM placement algorithm called Plug4Green that is used to compute the suitable location for the VM and the state of the servers depending on large size constraints. The QoS constraints ensure caused maximum energy consumption. Hence, a suitable technique was required to provide the balance between the minimum energy consumption and an accurate assurance of QoS constraints. Horri and Dastghaibyfard [10] proposed a novel study of QoS-aware VM consolidation to provide the necessary trade-off between QoS constraints and energy consumption. They adopted the new method based on historical data existing in CPU and VM memory. The increase in scalability of cloud data centers maximized energy consumption, and a suitable resource allocation strategy was required to reduce the energy consumption.

The traditional data center models have not included energy consumption as a key parameter for their configuration. Dupont et al. [11] presented the energy-aware framework for VM reallocation. They employed the constraint programming (CP) and entropy-based approaches to achieve the flexibility of the model. The evolution of high-throughput computing resource consolidation models effectively reduced energy consumption with idle resource problems. Ding [12] discussed the particle swarm optimized tabu (PSO-T) search for resource utilization level improvement in order to reduce energy consumption. The PSO-T algorithm turn-off the sparse servers to reduce the power and time consumption level. The sustainability of the cloud model was the major concern to develop energy efficiency and CO_2 emissions. Wajid et al. [13] considered the usefulness of sustainability of the cloud models to allow the conception and development of new techniques. The optimization of assets associated with sustainability cloud models was an important requirement to improve energy efficiency. Volk et al. [14] discussed the energy-efficient approach to Eco2Clouds projects under minimum energy consumption and CO_2 emissions. The cloud environment monitoring was based on data gathering regarding energy consumption with workload variations. Based on the processor's time and VM instances, the cost of virtualization varied since the cooling and energy costs for data centers exceeded the purchasing cost. Kim et al. [15] measured the energy consumption level based on the processing events. They utilized a scheduling algorithm to provide the necessary resources with an energy budget basis. The cloud service utilization level improvement required the resource optimization approach in addition to scheduling.

The simultaneous multiple cloud service requests were handled by parallel processing and required suitable resource allocation and scheduling of tasks. Li et al. [16] proposed resource allocation algorithms with preemptable tasks that adjusted the allotted resource dynamically. The periodical update regarding the task execution was required for the dynamic operation. The makespan and energy consumption were more in pre-emptive models. Nesmachnow et al. [17] introduced meta-broker (level I) and local providers (level II) to schedule all the received tasks. They investigated the energy consumption and capacity of the resources from the multicore processing systems. The advertisement of the illusion of resources to the customers required higher quality and reliability levels leads to excessive energy consumption. Benyi et al. [18] proposed the pliant system-based VM scheduling approach to reducing energy consumption. They designed the cloudsim-based simulation environment to evaluate the performance of the pliant-system. The nonadaptability of scheduling algorithms to uncertain and the dynamic nature made the job scheduling as a problematic task. Miranda et al. [19] presented the scheduling problems and reviewed the scheduling algorithms to provide the solution to uncertainty. The evolution of VM migration techniques from one physical machine to another in order to reduce time and service degradation. Soni and Kalra [20] discussed the various migration techniques (offline and live migration) to reduce the overall time consumption. They also presented the various live migration scenarios for better performance with minimum bandwidth. The distinct participants with their own objectives required the multiagent system with negotiation capabilities.

Leite et al. [21] proposed the server consolidation approach called federated application provisioning (FAP) strategy to manage the power consumption level in virtualized federated cloud environments. The major objective of the consolidation approach was to provide the trade-off between the QoS constraints satisfaction and minimum energy consumption compared to the trivial approach. The increase in processors provided the chance for failure occurrence. The application running in a cloud environment was represented by the workflow. But the inclusion of link failure and service provision failure violated the robustness of the computing environment. Singh and Kinger [22] discussed the failure removal with the fault tolerant mechanism (FTM). Failure detection was considered as an important stage in the virtualization process, and it required an effective control scheme with parametric guidance. The balance and consolidation of workloads with energy minimization required problem-solving techniques.

Garcia and Nafarrate [23] proposed the novel load-balancing heuristic algorithm to migrate the VM from the overutilized host to the underutilized host. The increase in cloud resource utilization level reduced the operational cost effectively. Liaqat et al. [24] surveyed the migration techniques and policies to represent future challenges in the VM domain. They proposed the queue-based migration model for memory page migration. The incorporation of virtualization and consolidation approaches were not provided the necessary balance between minimum energy consumption and better execution performance. Kaur and Chana [25] proposed the green cloud scheduling model (GCSM) that exploited the heterogeneity property of tasks and resources by using the scheduler unit. The GCSM facilitated energy efficiency and prevented the degradation on provider perspective. Alternatively, the execution of tasks within the time limit was achieved by GCSM under the client perspective. Efficient resource utilization was the challenging issue in the task execution performance improvement. Rathor et al. [26] provided the best fit and worst fit techniques to improve the resource utilization level at a high cost. Kruekaew and Kimpan [27] discussed the artificial bee colony algorithm with three scenarios first come first serve (FCFS), shortest job first (SJF), and longest job first (LJF) to improve the resource management performance. The migration and placement techniques in traditional approaches were not considered the utilization scenarios (overutilization/underutilization) and the job priority levels that lead to excessive energy consumption and less profit. Recent advances in cloud data centers and their energy-efficient methods were discussed in the following references [28–31].

3. Preserving Resource Handiness and Exigency-Based Migration (PRH-EM)

This section discusses the implementation of the novel techniques to maintain resource availability in order to deliver services to cloud users efficiently. Figure 1 shows the flow of the proposed preserving resource handiness and exigency-based migration (PRH-EM) algorithm for profit maximization with less energy consumption.

The (PRH-EM) algorithm contains successive processes such as federated environment creation, grade-based VM placement, job allocation, and exigency-based migration. Initially, the data centers are federated into four groups based on MIPS and cost. Then, the workload on the FDC and the capacity of the VM are measured. The capacity of the VM is compared with the FDC workload and place the VM into FDC if its capacity is less than the FDC. Then, the jobs are allotted to the respective VM in the updated host/VM list within the same and different FDC. When the jobs are allotted to the VM in the next FDC, the threshold value to indicate the utilization level is computed. Then, based on the threshold, the overutilized and underutilized VM are separated. Finally, the jobs are migrated from one VM to another according to their capacity.

3.1. Federation. The service provider is assumed to be autonomous with its own customers. During the demand conditions, the federation mechanism helps the providers identify the overloads. The federation model contains the cloud exchange service as the center. The providers send the necessary query to the exchange service to identify the available resources. The cloud exchange service generates the list of providers with the MIPS and cost value. The redirection of the requests helps to identify the suitable providers that use the MIPS and price list.

The component responsible for the decision-making regarding the allocation of additional resources is called the cloud coordinator. The simultaneous measurement of MIPS and cost for each data centers available in a cloud environment is responsible for federated environment creation. For each data center, the proposed algorithm computes the cost for million instruction (MI) execution. The overall cost of the MIPS for each data center is measured. The computed cost and MIPS are separated into three limits: low, medium, and high (L, M, and H). Depending on the MIPS and cost, the federation comprises four types as follows:

(i) Type I–More than high MIPS and less than low cost

(ii) Type II–More than high MIPS and less than medium cost

(iii) Type III–More than medium MIPS and less than medium cost

(iv) Type IV–More than low MIPS and less than medium cost

Algorithm 1 for the federation scheme is listed as follows:

The federated data centers are assumed to have enough resources to handle the various jobs. But the changes in resource availability and the virtualization process makes the job assignment as uncertain. The proposed algorithm addresses this issue to maintain resource availability and deliver services during failure conditions.

3.2. Grade-Based VM Placement. The VM placement in federated data centers is the second stage of the proposed PRH-EM algorithm. After all the data centers are federated into geographic locations, the carbon footprint for each data center is computed as follows:

$$\mathrm{CF} = \sum_{t=1}^{T} \sum_{j=1}^{d} \mathrm{PUE}_i \times \sum_{j=1}^{c} \left(cf_j \times \sum_{k=1}^{h} P\left(vm_{i,j,k,t} \times ht \right) \right), \quad (1)$$

where CF represents the carbon footprint of the cloud and PUE_i describes the power usage effectiveness which is the ratio of the overall power consumption of the data centers (d) to the power consumed by IT devices within holding time (ht). The arrival of new vm request initiates its allocation to the host depending on the carbon level. The estimated carbon level is considered as the workload for federated cloud data centers. Algorithm 2 proposed in this paper is referred to as grade-based VM placement since it follows the descending order of workload.

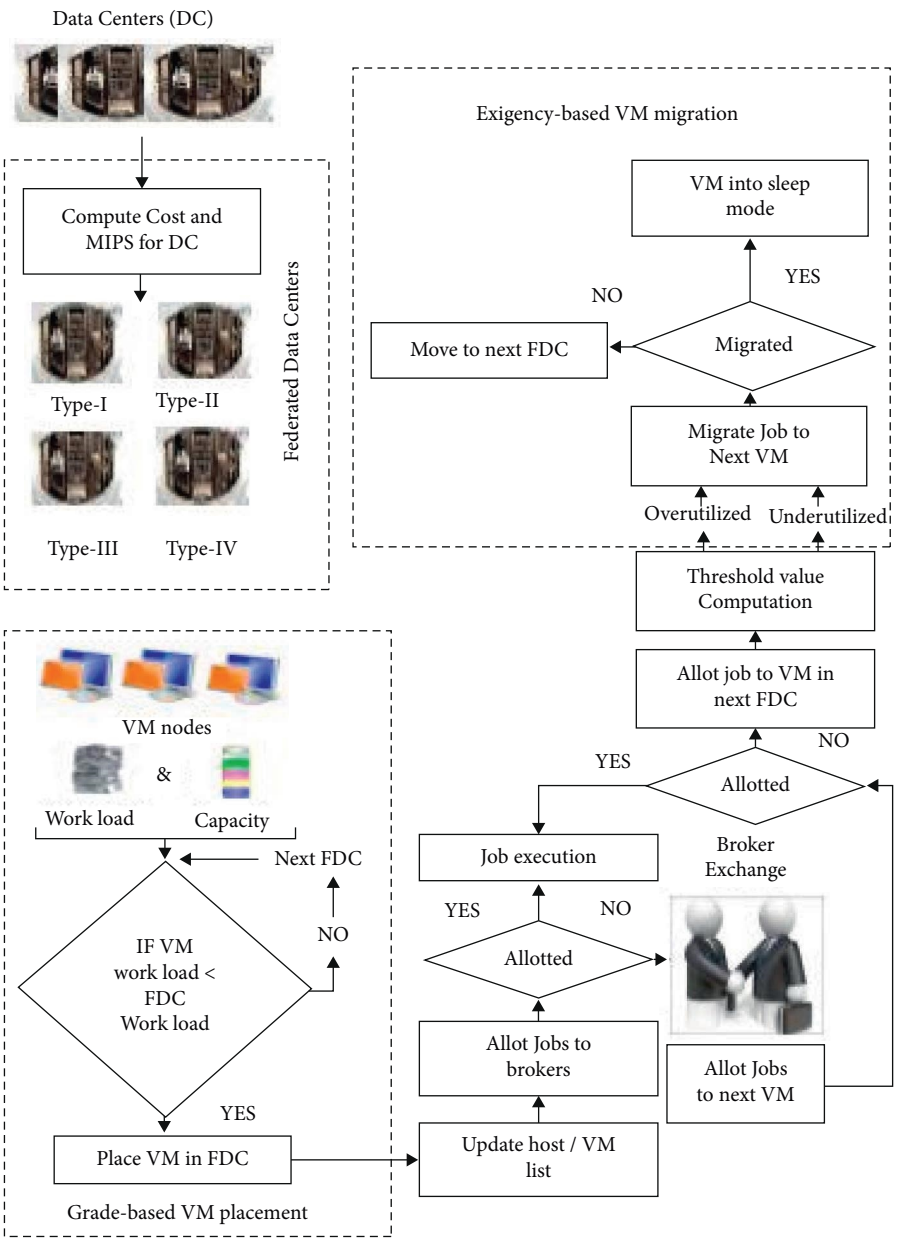


Figure 1: Overall flow of PRH-EM method.

(1) Collect the data centers (DC)
(2) Compute MIPS of each data center and cost/one MI
(3) Compute the cost of MIPS of each data center
(4) Split the MIPS and cost as follows:
$$\text{MIPS} = \{\text{MIPS}_L, \text{MIPS}_M, \text{MIPS}_H\}$$
$$\text{Cost} = \{\text{Cost}_L, \text{Cost}_M, \text{Cost}_H\}$$
(5) If (MIPS executed $\geq \text{MIPS}_H$ && Cost $< \text{Cost}_L$)
(6) FDC 1 $\leftarrow$ Type 1 *DCs*
(7) If (MIPS executed $\geq \text{MIPS}_H$ && Cost $\leq \text{Cost}_M$)
(8) FDC 2 $\leftarrow$ Type 2 *DCs*
(9) If MIPS executed $\geq \text{MIPS}_M$ && Cost $\leq \text{Cost}_M$
(10) FDC 3 $\leftarrow$ Type 3 *DCs*
(11) If (MIPS executed $\geq \text{MIPS}_L$ && Cost $\leq \text{Cost}_M$)
(12) FDC 4 $\leftarrow$ Type 4 *DCs*

Algorithm 1: Federation.

(1) Compute the workload for all the FDCs
(FDC 1, FDC 2, FDC 3, FDC 4)
(2) Load the VM_list and compute the carbon level using (1)
(3) Arrange the Host_list in descending order
(4) If ($cap_{vm} < cap_{FDC1}$)
(5) Place *vm* into host
(6) Set $status_{vm} \leftarrow$ Ready
(7) Compute the remaining capacity of VM
(8) Update Host_list and VM_list
(9) Else
(10) Select the next data center (FDC)
(11) Repeat from step 2
(12) Update the VM_list and Host_list
(13) Place all *vm* to the host

Algorithm 2: Grade-based VM placement.

The workload computation for each data center is the initial stage of the VM placement algorithm. Then, the carbon level for each VM is estimated for each machine in the VM list. Based on the workload and the carbon level estimation, the host list is created with the high-capacity host in the first place and the low-capacity host in the last place. Then, compare the capacity of the VM with the workload of the selected FDC to locate VM to the specific host. If the capacity of the selected VM is less than the FDC workload, then the corresponding VM is allotted to the selected host. Otherwise, the next FDC is selected and repeated from the list formation. During the comparison, Host_list and VM_list are periodically updated to know the specific host for each VM. The grade-based VM placement proposed in this paper is responsible for the effective service delivery on failure conditions. But the execution of the job is also a major concern in the energy-efficiency cloud environment.

3.3. Job Allocation. The major assumption for an energy-efficient cloud model is to allocate the job to one cloud only without replication. If there are multiple jobs arise under the specific instance, high-priority jobs are to be executed initially. Then, according to the capacity of the VM, the other jobs are allotted to the other VM. The particular data center includes the routers and switches responsible for the transportation of traffic between the servers and the outside world. The interconnection of processors is static in nature, and their utilization levels are periodical changes depending on the size of a job executed. The large size of I/O data transfer between them causes the overload condition. The brokers associated with VM monitor the capacity and job size to allocate the job to the maximum capacity VM. During the overload conditions (either VM capacity limit is reached or not), the broker will exchange their monitoring control with each other to provide constant resource availability.

The algorithm 3 for job allocation is described as follows:

Initially, the jobs to be executed are collected and their size is estimated. Then, the jobs are arranged in descending order with maximum size as the first one and minimum size as the last one. Then, the priority level is assigned to each job based on the size value. The capacity of the VM is estimated parallel to the priority assignment. Then, the job size is compared with the VM capacity level in the initial round of execution. If the capacity of the VM is greater than the job size, then the corresponding job is allotted to the VM from the VM_list. Then, the remaining capacity and jobs are estimated and placed to the nonallotted list for further processing. The job allocation algorithm 4 for a nonallotted list is described as follows:

The jobs from the nonallotted list are collected and the corresponding broker and host are identified. Then, switch the job to the next broker within the same FDC if there is any VM available with the necessary capacity until the last job on the list. If there is any job available in the list for the capacity limit reached state, then the next FDC is selected for job allocation. Finally, VM_*list* and *J* are updated after allocation. During the demand conditions (capacity limit is reached) for job execution, the jobs executed on VM is migrated from one VM to another VM in order to provide the service delivery to the user effectively.

3.4. Exigency-Based Migration. The traditional VM migration schemes violated resource availability. Alternatively, the exigency-based migration algorithm proposed in this paper estimates the utilization level of VM by the job. Based on the utilization level, two conditions arise such as overutilization and underutilization. The migration of jobs in underutilized status converts the idle VM into a sleep mode that reduces unnecessary energy consumption. Similarly, the jobs in overutilized status are migrated to the maximum capacity VM provides the immediate response that reduces the time delay. Algorithms 5 and 6 for overutilized and underutilized conditions are described as follows:

The VMs allotted for the execution of the job are extracted from the VM_list. The threshold values for the categorization of VM are computed to identify the status whether it is overutilized or underutilized. If the capacity of the VM is exceeded the maximum threshold value, then the corresponding VM is placed in overutilized list. Then, the job executed by the corresponding VM is migrated to the

(1) Collect the job list $J = \{J_1, J_2, \ldots, J_N\}$
(2) Compute the size of each job (size)
(3) Arrange the jobs in descending order based on size
(4) Compute $\max_{\text{size}}$ and $\min_{\text{size}}$
(5) Assign $(J_{\max_{\text{size}}}) \leftarrow \text{High}_{\text{priority}}$ and $(J_{\min_{\text{size}}}) \leftarrow \text{Low}_{\text{priority}}$
(6) Compute VM capacity (cap_{vm})
(7) Load high priority job to the VM with maximum capacity
(8) For $J_i i = 1, 2, \ldots, N$
(9) Collect the VMs from VM_list
(10) If $(\text{cap}_{vm} > J_{\max_{\text{size}}}$ && status)
(11) Allocate the job to VM
(12) Compute the remaining capacity of VM
(13) Update the VM_list and J
(14) Else
(15) Allocate the job to nonallotted list

Algorithm 3: Job Allocation.

(1) Collect the nonallotted jobs
(2) Identify the broker correspond to the job
(3) Identify the host from the Host_list associated to broker in same FDC (B_{FDC_i})
(4) $B_{\text{FDC}_i} \leftarrow \text{non} - \text{allotted Job}$
(5) Allot the job to VM monitored by B_{FDC_i}
(6) If (Job exist in list)
(7) Move the job to next FDC (FDC_{i+1})
(8) Update VM_list and J
(9) Else
(10) Execute the job

Algorithm 4: Job allocation for non-allotted list.

(1) Load VM_list
(2) Compute the threshold values (min, max) for each VM
(3) If $(\text{cap}_{vm} > \max)$
(4) Assign VM as overutilized
(5) List the remaining VMs with sufficient capacity (VM_R)
(6) If $(\text{cap}_{vm_R}! > \max)$
(7) Migrate the over utilized job to the VM_R
(8) Else
(9) Migrate job to next type FDC_{i+1}
(10) Compute remaining capacity of VM
(11) Update VM_list and migration list

Algorithm 5: Overutilized.

(1) Identify underutilized VMs
(2) List the remaining VMs with sufficient capacity in FDC (VM_R)
(3) If $((\text{cap}_{vm_R} > \min)$ && (VM not in underutilized list))
(4) Migrate the underutilized job to the VM_R
(5) Else
(6) Migrate job to next type FDC_{i+1}
(7) Set VM free and sleep mode
(8) Update VM_list and migration list

Algorithm 6: Underutilized.

remaining VM for the specific FDC. Alternatively, if the capacity of the selected VM is less than the minimum threshold value, it can be regarded as underutilized. The jobs executed by the underutilized VM are migrated to the next VM with sufficient capacity. The migration based on the utilization level maintains resource availability and reduces energy consumption effectively. VM failure detection and service failover are the two main principles in failure handling.

> In general, failure detection is the process of identifying when something went wrong and informing the appropriate administrator so they can address it.
>
> System monitoring and failure detection are two different things. There are frequent and quick failure detection checks (for example, every 5 seconds). They are typically more constrained in what they check as a result.

Failure tests in API Connect simply examine the availability of the web server and database. Contrarily, monitoring checks are often performed less regularly and are more likely to examine factors like CPU, RAM, and disc space utilization. These checks' outcomes can then be kept track of for historical trend analysis to find memory leaks.

4. Performance Analysis

This section discusses the performance of proposed preserving resource handiness and exigency-based migration (PRH-EM) regarding the accuracy, response time variations for various jobs, and VMs. The comparative analysis of proposed PRH-EM with the best/worst fitness, greedy [26], and ABC approaches [27] assures the effectiveness in energy-efficient cloud environment. Besides, the profit variations for nonfederated totally in-house (NFTI), federation aware outsourcing oriented (FAOO), federated aware profit oriented (FAPO), [1] and proposed PRH-EM also presented.

4.1. Response Time. The overall time consumption from the rise of a new request to the response from the particular machine refers to the response time. The comparison between the proposed PRH-EM with the existing best/worst fitness and greedy [26] regarding the response time depicts that the hybrid processes such as VM placement, scheduling, and migration in the proposed approach reduce the response time efficiently.

Figure 2 shows the response time variations for a various number of VMs. The increase in VM reduces the response time effectively in traditional approaches. But grade-based placement and exigency-based migration provide a constant resource availability level that reduces the response time compared to other. The comparison shows that the PRH-EM offers a 28 and 50% reduction in response time compared to greedy approaches.

4.2. Accuracy. The measure of how the data centers are utilized for the overall time limit refers to the accuracy of the proposed system. The accuracy of the FDCs in the proposed

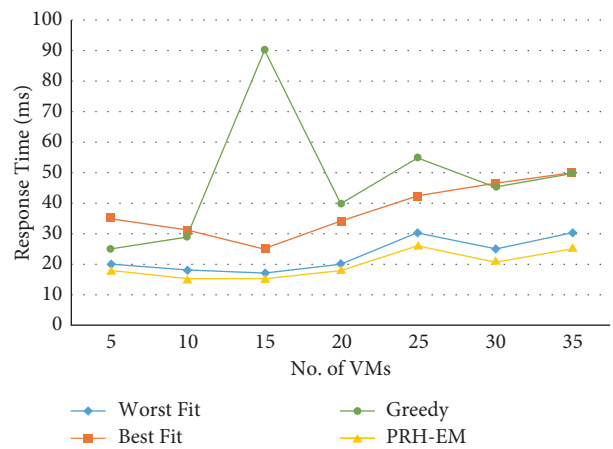


Figure 2: Response time analysis.

PRH-EM validated for various VM shows better performance compared to traditional approaches. Table 1 shows the accuracy variations for each FDC over the various VM count.

For the minimum value of VM (25), the FDC 4 shows a better performance than the other FDCs. Similarly, the FDC 3 provides better performance for higher VMs. The increase in VMs reduces the accuracy of FDC within the considerable level of the constant resource availability and efficient delivery to the user during the failure conditions are achieved by PRH-EM.

4.3. Makespan. The time consumption for the number of tasks executed within FDC is represented by makespan. The increase in tasks requires more VM that consumes more time for service delivery. But the migration based on exigency conditions reduces the makespan effectively.

Figures 3 and 4 show the makespan performance for a number of jobs and VMs, respectively. The increase in job size and VM size will increase the makespan value. But the provision of migration and grade-based VM placement reduces the makespan effectively. The comparative analysis shows that the PRH-EM scheme reduces the makespan by 40 and 8.3% of minimum and maximum jobs respectively compared to the ABC-SJF approach. Similarly, the PRH-EM reduces the makespan by 11.76 and 25% compared to the ABC-LJF approach respectively.

4.4. Energy Consumption. The energy model creation in this paper is based on the assumption that processor utilization and energy consumption are directly proportional to each other. The utilization level for the particular resource is expressed as

$$U_i = \sum_{j=1}^{n} u_{i,j}, \tag{2}$$

where n–number of VMs, $u_{i,j}$–utilization level for particular VM_j

Table 1: Accuracy analysis.

No. of VMs	Accuracy (%)			
	FDC1	FDC2	FDC3	FDC4
25	91	89	92	93
40	84	90	85	80
55	80	82	86	81

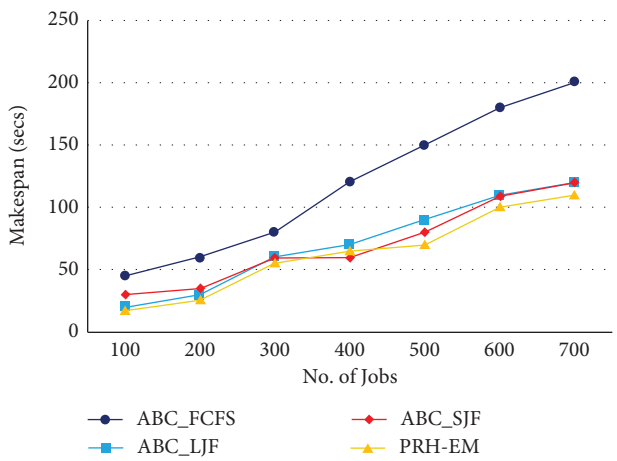


Figure 3: Makespan analysis for no. of jobs.

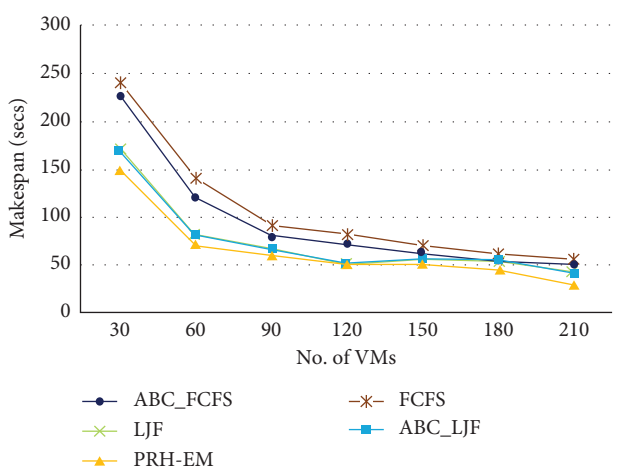


Figure 4: Makespan analysis for no. of VMs.

The energy consumption for the particular VM depends on the utilization level, power consumption (P_{max}) at the peak overload (100% utilization), and power consumption (P_{min}) at the active mode (1% utilization), and it is expressed as

$$E_i = (P_{max} - P_{min}) \times U_i + P_{min}. \quad (3)$$

The energy variations with the utilization level of CPU for proposed PRH-EM and existing MD_MMT [28] show the effective reduction of energy consumption in PRH-EM.

Figure 5 shows the comparative analysis of PRH-EM with the MD_MMT regarding energy consumption. The proposed PRH-EM provides the reduction of energy consumed by the group of data centers by 15.79% for minimum (20%) utilization levels.

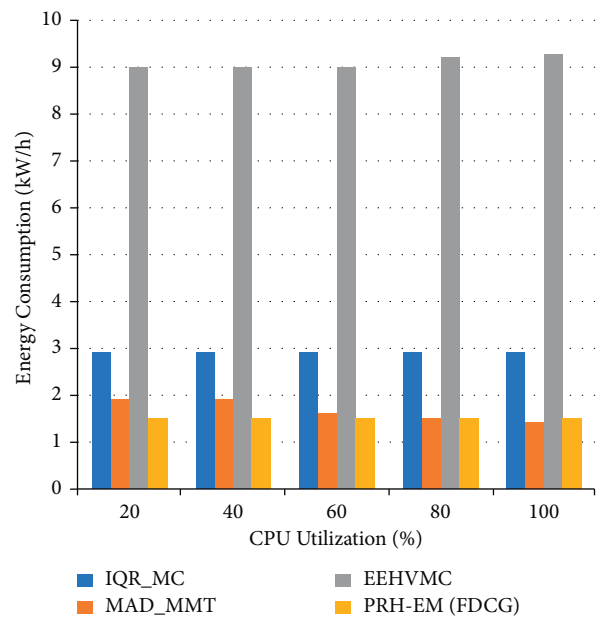


Figure 5: Energy consumption analysis.

4.5. Carbon Emission. The carbon emission level depends on the PUE level and the number of VMs used according to equation (1). The analysis of carbon utilization level for proposed PRH-EM and existing carbon efficient placement and migration (CEPM) and round-robin (RR) migration algorithms [6]. Among these methods, the CEPM offers less carbon emission level compared to RR.

Figure 6 shows the comparative analysis of carbon emission level variations for each VM request. The increase in the number of requests increases the carbon footprint level linearly. The comparative analysis shows the PRH-EM reduces the utilization level by 20% for a maximum number of VM requests (200) compared to CEPM, respectively.

4.6. Profit Analysis. The deviation of revenue achievement and the cost required to achieve the revenue at the same time refers to profit. The mathematical formulation of profit is expressed as

$$\text{profit}(\Delta t) = \text{Revenue}(\Delta t) - \text{cost}(\Delta t). \quad (4)$$

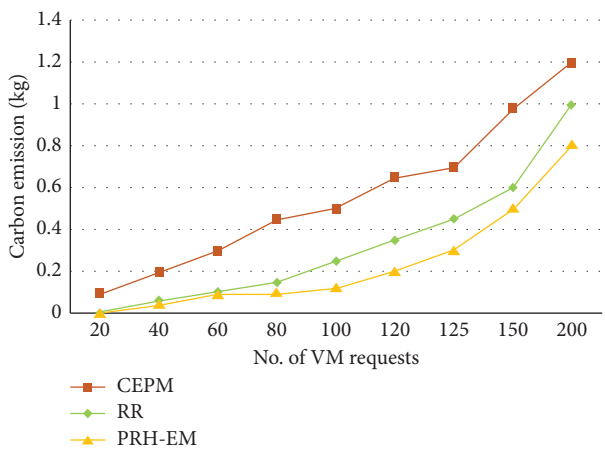


Figure 6: Carbon emission analysis.

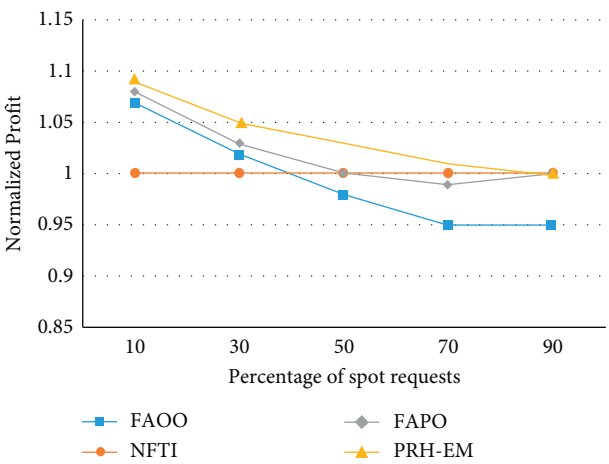


Figure 7: Profit analysis.

The experimental analysis of profit variation for the trivial, FAPO strategies [1] shows that the FAPO models provide more profit.

Figure 7 shows the comparative analysis of the profit variation for existing FAPO and PRH-EM models with the different percentages of spot request count. The comparative analysis shows that PRH-EM increases the profit by 0.9% compared to FAPO due to the exigency-based migration and grade-based VM placement for the minimum percentage of spot requests.

5. Conclusion

This paper discussed the various issues in the scheduling/migration schemes during the maintenance of resource availability in a confederated cloud environment. A novel VM migration algorithm is proposed to provide the trade-off between the reduction in energy consumption and profit maximization. Initially, the available data centers are categorized based on MIPS and cost value. The prior categorization to job allocation has a great impact on resource availability maintenance. The comparison between the cumulative workload capacity of data centers and the individual VM capacity offers immediate migration during the demand conditions that prevented resource loss. The overutilized and underutilization-based migration reduced the number of resources that directly reduced energy consumption and carbon emissions. The proposed work maintained the constant resource availability and service delivery to the users during the VM failure conditions that increased the profit level. The comparative analysis of the proposed algorithm with the existing methods assured the effectiveness of the proposed work in a federated cloud environment.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. N. Toosi, R. N. Calheiros, R. K. Thulasiram, and R. Buyya, "Resource provisioning policies to increase IaaS provider's profit in a federated cloud environment," in *Proceedings of the IEEE 13th International Conference on High-Performance Computing and Communications (HPCC)*, pp. 279–287, Banff, Canada, September 2011.

[2] A. Celesti, M. Fazio, M. Villari, and A. Puliafito, "Virtual machine provisioning through satellite communications in federated cloud environments," *Future Generation Computer Systems*, vol. 28, no. 1, pp. 85–93, 2012.

[3] F. Paraiso, N. Haderer, P. Merle, R. Rouvoy, and L. Seinturier, "A federated multi-cloud PaaS infrastructure," in *Proceedings of the IEEE 5th International Conference on Cloud Computing (CLOUD)*, pp. 392–399, Honolulu, HI, USA, June 2012.

[4] A. Kertész, G. Kecskemeti, M. Oriol et al., "Enhancing federated cloud management with an integrated service monitoring approach," *Journal of Grid Computing*, vol. 11, no. 4, pp. 699–720, 2013.

[5] I. Petri, T. Beach, M. Zou, J. D. Montes, O. Rana, and M. Parashar, "Exploring models and mechanisms for exchanging resources in a federated cloud," in *Proceedings of the IEEE International Conference on Cloud Engineering (IC2E)*, pp. 215–224, Boston, MA, USA, March 2014.

[6] B. Wadhwa and A. Verma, "Carbon efficient VM placement and migration technique for green federated cloud datacenters," in *Proceedings of the International Conference on Advances in Computing Communications and Informatics (ICACCI)*, pp. 2297–2302, Delhi, India, September 2014.

[7] B. Wadhwa and A. Verma, "Energy and carbon efficient VM placement and migration technique for green cloud datacenters," in *Proceedings of the Contemporary Computing (IC3), 2014 Seventh International Conference on*, pp. 189–193, Noida, India, August 2014.

[8] M. H. Ferdaus, M. Murshed, R. N. Calheiros, and R. Buyya, "Network-aware virtual machine placement and migration in cloud data centers," *Emerging Research in Cloud Distributed Computing Systems*, vol. 42, 2015.

[9] C. Dupont, F. Hermenier, T. Schulze, R. Basmadjian, A. Somov, and G. Giuliani, "Plug4Green: a flexible energy-aware VM manager to fit data centre particularities," *Ad Hoc Networks*, vol. 25, pp. 505–519, 2015.

[10] A. R. A. Horri and GH. Dastghaibyfard, "Energy and performance-aware virtual machine consolidation in cloud computing A two-dimensional approach," *Turkish Journal of Engineering, Science and Technology*, vol. 1, pp. 20–35, 2015.

[11] C. Dupont, G. Giuliani, F. Hermenier, T. Schulze, and A. Somov, "An energy aware framework for virtual machine placement in cloud federated data centres," in *Proceedings of the Third International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet (e-Energy)*, pp. 1–10, Madrid, Spain, May 2012.

[12] F. Ding, "Energy-aware and revenue-enhancing combinatorial scheduling in virtualized of cloud datacenter," *JCIT*, vol. 7, no. 1, pp. 62–70, 2012.

[13] U. Wajid, B. Pernici, and G. Francis, "Energy efficient and CO2 aware cloud computing: requirements and case study," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 121–126, Manchester, UK, October 2013.

[14] E. Volk, A. Tenschert, M. Gienger, A. Oleksiak, L. Sisó, and J. Salom, "Improving energy efficiency in data centers and federated cloud environments: comparison of CoolEmAll and Eco2Clouds approaches and metrics," in *Proceedings of the Third International Conference on Cloud and Green Computing (CGC)*, pp. 443–450, Karlsruhe, Germany, September 2013.

[15] N. Kim, J. Cho, and E. Seo, "Energy-credit scheduler: an energy-aware virtual machine scheduler for cloud systems," *Future Generation Computer Systems*, vol. 32, pp. 128–137, 2014.

[16] J. Li, M. Qiu, Z. Ming, G. Quan, X. Qin, and Z. Gu, "Online optimization for scheduling preemptable tasks on IaaS cloud systems," *Journal of Parallel and Distributed Computing*, vol. 72, no. 5, pp. 666–677, 2012.

[17] S. Nesmachnow, B. Dorronsoro, J. E. Pecero, and P. Bouvry, "Energy-aware scheduling on multicore heterogeneous grid computing systems," *Journal of Grid Computing*, vol. 11, no. 4, pp. 653–680, 2013.

[18] A. Benyi, J. D. Dombi, and A. Kertész, "Energy-aware VM scheduling in IaaS clouds using pliant logic," *Closer*, vol. 74, pp. 519–526, 2014.

[19] V. Miranda, A. Tchernykh, and D. Kliazovich, "Dynamic communication-aware scheduling with uncertainty of workflow applications in clouds," in *High Performance Computer Applications*, pp. 169–187, Springer, 2015.

[20] G. Soni and M. Kalra, "Comparative study of live virtual machine migration techniques in cloud," *International Journal of Computer Application*, vol. 84, no. 14, pp. 19–25, 2013.

[21] A. F. Leite and A. C. M. A. De Melo, "Energy-aware multi-agent server consolidation in federated clouds," in *Cloud Computing*, pp. 72–81, Springer, 2012.

[22] A. Singh and S. Kinger, "Virtual machine migration policies in clouds," *International Journal of Science and Research (IJSR), India Online ISSN*, pp. 2319–7064, 2013.

[23] J. O. Gutierrez-Garcia and A. Ramirez-Nafarrate, "Collaborative agents for distributed load management in cloud data centers using live migration of virtual machines," *IEEE Transactions on Services Computing*, vol. 8, no. 6, pp. 916–929, 2015.

[24] M. Liaqat, S. Ninoriya, J. Shuja, R. W. Ahmad, and A. Gani, "Virtual machine migration enabled cloud resource management: a challenging task," 2016, http://arxiv.org/abs/1601.03854.

[25] T. Kaur and I. Chana, "Energy-aware scheduling of deadline-constrained tasks in cloud computing," *Cluster Computing*, vol. 19, no. 2, pp. 679–698, 2016.

[26] V. S Rathor, R. K Pateriya, R. K Gupta, M. Shelar, S. Sane, and V. Kharat, "An efficient virtual machine scheduling technique in cloud computing environment," *International Journal of Modern Education and Computer Science*, vol. 7, no. 3, pp. 39–46, 2015.

[27] B. Kruekaew and W. Kimpan, "Virtual machine scheduling management on cloud computing using artificial bee colony," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, pp. 12–14, 2014.

[28] R. Karthikeyan and P. Chitra, "Novel power reduction framework forenhancing cloud computing by integrated GSNN scheduling method," *Cluster Computing*, vol. 21, no. 1, pp. 755–766, 2018.

[29] R. Karthikeyan and V. Balamurugan, "Energy-aware and SLA-guaranteed optimal virtual machine swap and migrate system in cloud-Internet of Things," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 10, Article ID e6171, 2021.

[30] M. Shaukat, W. Alasmary, E. Alanazi, J. Shuja, S. A. Madani, and C.-H. Hsu, "Balanced energy-aware and fault-tolerant data center scheduling," *Sensors*, vol. 22, no. 4, p. 1482, 2022.

[31] S. Bashir, S. Mustafa, and R. W. Ahmad, "Multi-factor nature inspired SLA-aware energy efficient resource management for cloud environments," *Cluster Computing*, vol. 28, p. 188p. 8, 2022.

Hindawi
Mobile Information Systems
Volume 2022, Article ID 9031201, 16 pages
https://doi.org/10.1155/2022/9031201

Research Article

A Composite Service Provisioning Mechanism in Edge Computing

Junna Zhang,[1,2] **Xiaoyan Zhao**,[1] **Yali Wang**,[1] **Peiyan Yuan**,[1] **and Xinglin Zhang**[3]

[1]*College of Computer and Information Engineering, Henan Normal University, Xinxiang, Henan, China*
[2]*Engineering Lab of Intelligence Business & Internet of Things, Henan Province 453000, China*
[3]*South China University of Technology, Guangzhou, Guangdong, China*

Correspondence should be addressed to Junna Zhang; jnzhang@htu.edu.cn

Received 20 April 2022; Accepted 6 July 2022; Published 18 August 2022

Academic Editor: Li Duan

Users can invoke various composite services in Edge Computing (EC) with the development of Service Computing. Generally, users cannot invoke the optimal composite service due to the migration of component service or failure in the edge device. In this study, a Composite Service Provisioning mechanism in EC (CSP-EC) is proposed. It starts when the location of component service changes and terminates when the optimal composite service is obtained. CSP-EC employs dynamic optimization process to meet the real-time requirements. It first caches m optimized intermediate solutions from the first m users, which can be reused by other users to get the final optimal solution. Moreover, a multipopulation Estimation of Distribution Algorithm is used to reduce the probability of falling into a local optimum, and the roulette mechanism is used to accelerate the optimization process. Extensive simulations are conducted based on the real-world dataset. The results show that the proposed mechanism achieves higher quality, better stability, and shorter execution time compared with other schemes.

1. Introduction

Edge Computing (EC) is being standardized by the European Telecommunications Standards Institute (ETSI) and has received increased attention in recent years. Some IT giants, including Huawei, IBM, and Intel, are pushing EC technology at an unprecedented speed [1]. EC provides IT services and cloud computing capabilities at the edge mobile network, within the radio access network and in close proximity to mobile subscribers. It reduces latency, ensures highly efficient network operation and service delivery, and offers an impressed user experience [2–7].

For example, people's daily life is more and more dependent on cars. As a result, the road capacity of many cities tends to be saturated, and the traffic jams occur frequently which seriously affects people's normal work and life. Therefore, it is important to integrate EC and automobile industry to improve transportation efficiency by coordinating the traffic volume of each road.

Figure 1 shows an EC environment which includes Road Side Units (RSU) and on-board unit. The two kinds of devices communicate with each other via the 5G infrastructure. A software system is predeployed in EC; it coordinates the traffic volume of roads and makes path plans for drivers, so as to alleviate the traffic jam. Here the reason that we deploy the software system in the edge instead of the traditional cloud data center is that it is data-intensive, computation-intensive, and highly real-time sensitive. The delay from the remote cloud data center cannot meet the real-time requirement [8, 9]. Otherwise, it needs to implement more functions, such as real-time road condition perception, logical reasoning, decision-making, and path planning. On the other hand, the storage and computing capabilities of devices in the EC environment are limited, making it difficult to implement the entire software system on one device. Therefore, the functions (i.e., component services) required by the software system can be implemented and deployed in different edge devices, and multiple component services can be combined to achieve the functions of software systems to alleviate traffic jam (i.e., composite service) through the Service Oriented Architecture [10, 11].

The response time of component services should be as short as possible, because of the small coverage of edge

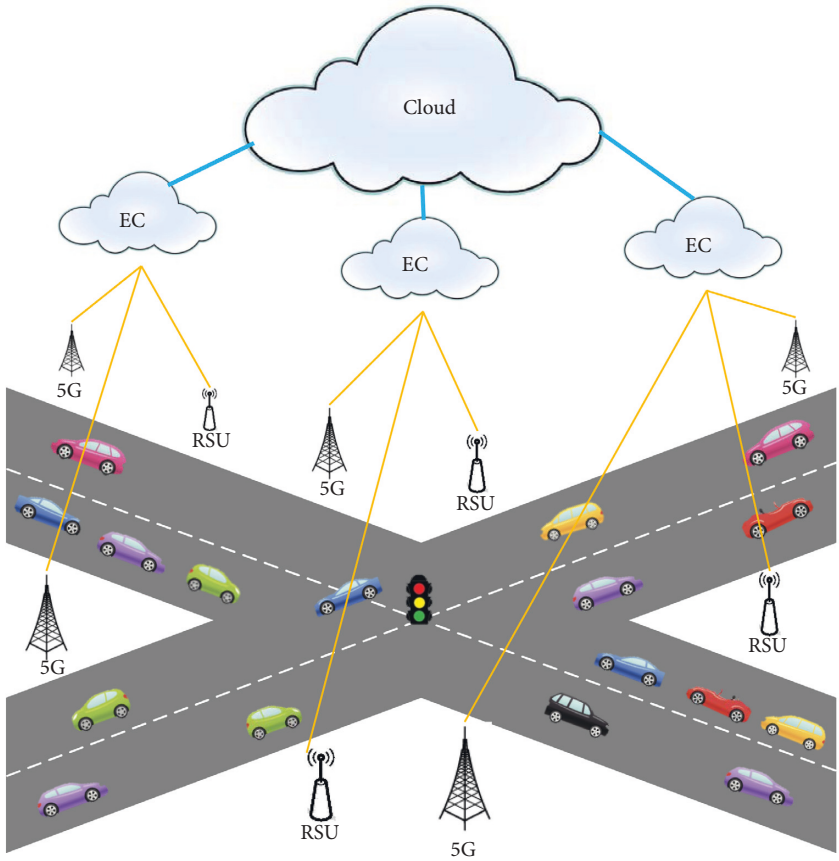


Figure 1: Realize composite services based on the EC environment.

devices, the fast velocities of vehicles, and the high real-time requirement of transportation systems. Meanwhile, component services with the same function must be evenly deployed on different edge devices to prevent vehicles from driving out of coverage or for fault tolerance. Therefore, vehicles at different locations need to call a set of component services deployed on different edge devices subject to a response time constraint on the composite service.

There are many tasks to be processed in EC environment, and some edge devices may be crash because of overload. Tasks on edge devices with overload can be migrated to other devices with lighter load [12, 13]. That is, component services deployed on one edge device may also be migrated to other edge devices, so the deployment position of a component service is dynamically changed. Therefore, component services invoked by vehicles passing the same location at different times may also be different (if deployed on different edge devices, even if they are the same component services, we also call them different component services). The reliability of composite service that is composed of them must be the highest due to the dynamic nature of component services, so as to improve the probability of the final successful execution of composite services.

In summary, the composite service of alleviating traffic jam has high reliability requirements and needs to meet a response time constraint. At the same time, it is a dynamic multiobjective optimization problem that requires location awareness and high real-time requirements. The essence of dynamic multiobjective optimization problem is to find the optimal solution dynamically in the search space. The optimal composite services are found according to different locations for the above research scenario. However, the real-time requirement for alleviating traffic jam is high, and it is impossible to find the optimal solution for users passing through at first when component service deployment is changed somewhere. Only the optimized intermediate solution to meet the real-time requirements can be found, but it can be provided for users to realize the function of alleviating traffic jam as soon as possible. The optimal solution can be further optimized to provide the optimal alleviate traffic jam function for users passing through the location.

A Composite Service Provisioning mechanism in Edge Computing (CSP-EC) is proposed. It will be triggered when the component service deployment changes at a certain location and terminated when the optimal composite service is obtained. In general, the mechanism is a progressive optimization process. Firstly, it decomposes the complex dynamic optimization problem into several static optimization problems. Each static optimization problem gets the solution at a certain location within a certain time limit; i.e., an optimized intermediate solution is obtained. Moreover, the multipopulation Estimation of Distribution Algorithm (EDA) is used to implement the static optimization process, so that the obtained

intermediate solution is closer to the optimal solution. And then, in order to speed up the progressive optimization process, the strategy of saving the optimized intermediate solutions is adopted. However, it may make the final optimization process fall into local optimum if only saving one optimized intermediate solution. Therefore, m optimized intermediate solutions are memorized to improve the probability of finding the optimal solution. Finally, the roulette mechanism is used to increase the reuse probability of the optimized intermediate solution with high fitness. The reason behind this is that the memorized m optimized intermediate solutions can be classified according to their fitness values, and the higher fitness value should be reused with a high probability to accelerate the optimization process. Therefore, these optimized intermediate solutions cannot be reused in a uniform way. The reused probability of each optimized intermediate solution should be proportional to the fitness value, so the roulette mechanism is selected.

The main contributions of this paper are summarized as follows:

(1) A multipopulation EDA is proposed in our study. The probability of falling into a local optimum is limited by increasing population diversity.

(2) Using limited storage space stores intermediate solutions for users to reuse, which is used to reduce the time to obtain the optimal solution.

(3) The experimental results show that the proposed mechanism can achieve higher quality, better stability, and shorter execution time compared with other mechanisms.

The remainder of this paper is organized as follows: Section 2 formalizes the research problems. Section 3 details the implementation of the proposed mechanism. Section 4 provides simulation experiments. Section 5 overviews related work. The last section concludes the paper and also summarizes the limitations of the proposed mechanism and the direction of our efforts in the future.

2. Problem Formulation

Composite service is composed of component services according to a certain composite structure to realize users' complex functional requirements. This paper adopts the key concepts and definitions of composite services from our previous work [14–24].

Definition 1 (Component Service). A component service is independent service that implements simple functions, and it is a 3-tuple $s = \{Id, \text{Fun}, \text{QoS}\}$, where Id is the unique identifier of the component service. Fun is the set of functions provides, and every function includes the input, output, and result of the service. QoS is Quality of Service (QoS) which represents a set of quality attributes $\text{QoS} = \{q_1, q_2, \ldots, q_M\}$, such as response time, reliability, cost, and availability.

Definition 2 (Service Class). A service class is a set of component services that implement the same functions and have different QoS attributes.

Definition 3 (Composite Service Plan). A composite service plan is a 2-tuple $CSP = (T, St)$, where $T = \{t_1, t_2, \ldots, t_N\}$ is a set of abstract components, that is, a set of tasks. S_t describes the structural information of the composite service plan, such as serial, parallel, conditional, and loop structure.

A composite service plan is only an abstract description of a complex software system. Each task must be realized by invoking a component service in reality. It can be selected from a service class which has multiple component services with different QoS, and all of them can be used to fulfill the task. However, users often have certain constraints on composite services; for example, the reliability cannot be lower than 0.9, and the response time cannot exceed 1 second. Therefore, the problem of composite service optimization is to select a set of component services in the service classes that satisfy user constraints, and the selected ones compose the optimal composite service.

According to the above definition, it is assumed that the composite service includes N tasks, which are represented by symbol $t_i\,(1 \le i \le N)$. t_i includes M component services $s_{i,1}, s_{i,2}, \ldots, s_{i,M}$, which form service class $S_i\,(1 \le i \le N)$. Each service class includes K QoS attributes, which are represented by K -dimension vector $Q'_{i,j} = (q'_{i,1}, q'_{i,2}, \ldots, q'_{i,K})\,(1 \le i \le N)$. In the composite service process, the user's preference weights for QoS attributes of component services are different, which are reflected by the K-dimensional vector $w = (w_1, w_2, \ldots, w_K)$, where $\sum_{h=1}^{K} w_h = 1$.

The service composition process is to select a component service in each service class to form a composite service with the best performance. The selection needs to meet/satisfy user constraints and QoS preference weights of composite service. Therefore, the primary issue is how to evaluate the performance of composite services based on the QoS attributes of component services.

Composite service performance can be evaluated based on overall QoS attributes, which are determined by its composite structure. In general, there are four basic composite structures, such as serial, parallel, conditional, and loop structure [25]. The overall QoS attributes vary according to their composite structures. The aggregation functions of three typical attributes are given in Table 1, i.e., reliability, response time, and throughput.

$q(s_i)$ are standardized QoS attributes of component service in a composite service. There are n component services in serial and parallel structure. For a loop structure, component service loops k times.

A simple weighting method [25] can be used to calculate the overall QoS attributes of a composite service according to Table 1. Formally, the overall QoS of a composite service cs is computed as

$$U(cs) = \sum_{i=1}^{K} w_i q_i. \tag{1}$$

Table 1: Aggregation function in different composite structures.

Attribute	Serial	Parallel	Conditional	Loop
Reliability	$\prod_{i=1}^{n} q(s_i)$	$\prod_{i=1}^{n} q(s_i)$	$q(s_i)$	$q(s_i)^k$
Response time	$\sum_{i=1}^{n} q(s_i)$	$\max_{i=1}^{n} q(s_i)$	$q(s_i)$	$k * q(s_i)$
Throughput	$\min_{i=1}^{n} q(s_i)$	$\sum_{i=1}^{n} q(s_i)$	$q(s_i)$	$k * q(s_i)$

$U(cs)$ is the overall QoS of composite service. q_i is the aggregation function for ith QoS attribute.

According to formula (1), the optimal composite service is defined as follows.

Definition 4 (Optimal Composite Service) An optimal composite service is composed of the selected component services from each candidate service class and its aggregated QoS is optimal among all the composite services, while satisfying all the QoS constraints.

For the scenario described in this research, it is to find a set of component services so that the resulting composite service has the optimal aggregated QoS and satisfies the response time constraint.

$$\begin{aligned} &\max U(cs) \\ &\quad s.t. \\ &q_{\text{res}} <= C, \end{aligned} \tag{2}$$

where q_{res} is response time of composite service, q_{rel} represents the reliability of composite service, and C represents a constraint on response time. Our scenario only focuses on the reliability; i.e., in formula (1), $K = 1$, $w_1 = 1$, and q_1 is reliability of composite service. Then the scenario in our study can be formalized as the following optimization problem:

$$\begin{aligned} &\max q_{\text{rel}} \\ &s.t. \qquad . \\ &q_{\text{res}} <= C \end{aligned} \tag{3}$$

Based on Table 1, the above q_{res} and q_{rel} of composite service can be calculated according to the different structures. The above optimization problem is a NP-hard problem. The optimal solution cannot be found in polynomial time, and only a heuristic algorithm can be used to find an approximate optimal solution [26–28].

3. Composite Service Provisioning Mechanism in EC

The users need to call component services deployed on different edge devices at different locations for the scenario described in this research. Due to component service migration or edge device failure, component services called by users passing through the same location at different times may also be different. Moreover, because the users are mobile, solving the optimal composite service problem is a dynamic optimization process. To simplify the problem, we first make some assumptions. (1) Dynamic optimization process start time: the dynamic optimization process is started when the component service deployment location changes somewhere (for example, when a component service is migrated, or when an edge device fails and component services on it are unavailable). (2) Dynamic optimization process end time: the optimal solution at a certain position tends to be stable. (3) The time when the users pass the same location obeys the Poisson distribution.

In general, users invoke the optimal composite service when the component service deployment locations do not change in the same location, while the optimization process is restarted when component service deployment locations change. Because of real-time requirements, users invoke better composite service (i.e., optimized intermediate solution) during the optimization process. The users continue to invoke the optimal composite service when the optimization process is completed, i.e., the new optimal composite service is found. We focus on the gradual optimization process when component service deployment locations change.

The flowchart of CSP-EC is illustrated in Figure 2. When the CSP-EC method is triggered, the users passing this location recalculate the optimal composite service in the current EC environment; that is, the static optimization process (for details, see Section 3.2) is started. A user should get a composite service within a certain time limit due to real-time requirements. Therefore, a user can only perform part of the optimization process and has to stop the optimization execution and execute the resulting composite service. At this time, only the intermediate solution in the optimization process can be obtained, but it can also complete all the required functions according to the characteristics of the composite service. However, it is not the optimal composite service, and we call it the optimal intermediate solution. In order to allow users who pass by this location to have better user experience, the optimization process should be continued. CSP-EC saves a certain number of optimal intermediate solutions and uses roulette selection mechanism to determine the reused optimized intermediate solutions and finally complete the entire optimization process (for details, see Section 3.3).

3.1. Fitness Function Construction and Coding. Similar to traditional evolutionary algorithms, EDA also uses fitness functions to evaluate the pros and cons of individuals in the population during the optimization process. The ultimate optimization goal of this paper is to select a set of component services among the candidate component services to achieve the maximum reliability aggregation value while meeting the user's constraints on response time. It can be known from Table 1 that the reliability aggregation value calculation method is different under different composite structures, and different composite services have different composite structures due to different functional requirements. Therefore, the calculation formula of the reliability aggregation value of different composite services is different.

Literature [29] proposed a method for converting the parallel, conditional, and loop structure into sequence structure to use a unified formula to calculate the composite

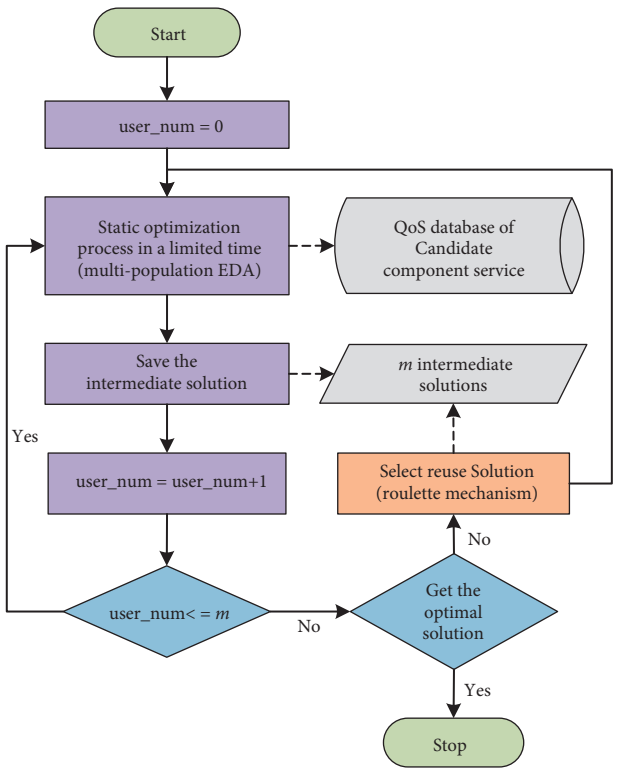


Figure 2: The flowchart of CSP-EC.

service QoS aggregation value. It also proved that the aggregated QoS value after conversion is the same as that of before conversion. Therefore, we use the method in [29] to calculate the composite service reliability aggregation value. Assume that, after conversion, the composite service has n component services, and the following formula can be used as the fitness function.

$$f = \prod_{i=1}^{n} rel_i. \tag{4}$$

EDA needs to encode the individuals in the population. The individuals in this research are composite schemes composed of candidate component services in the service class. In order to distinguish K candidate component services in each service class, an index is established for each candidate component service. The solution of composite service in CSP-EC is coded as chromosomes (i.e., individuals). According to the number of tasks N, the chromosomes are divided into N genes. The value of each gene represents the index of the selected candidate component service. The chromosomes use binary coding [25], and each gene represents the index value of the selected component service in the service class $S_i\,(1 \le i \le N)$. For example, the gene task 1 value is 00001010 in Figure 3, which means that the component service with the index number 10 in S_1 is selected.

3.2. Static Optimization Process. Each static optimization process uses improved EDA. The concept of EDA was first proposed in 1996 [30], which is an evolutionary algorithm based on group search. However, EDA does not generate the next generation of population through crossover and mutation like the Genetic Algorithm (GA). Firstly, EDA uses statistical analysis method to establish a probability model for the better individuals. Then it generates the next population by sampling according to the probability model. Next it updates the probability model based on the new population. Finally, it gets the optimal solution by continuously sampling and updating the probability model. In general, GA simulates biological evolution at the "micro" level, while EDA simulates biological evolution at the "macro" level. Thus, EDA can be considered as a combination of statistical learning and evolutionary computation.

Similar to the traditional service composition optimization process, static optimization process selects a set of component services that make formula (3) achieve the maximum aggregate value according to the reliability values of component services in the current state. However, each static optimization process will end within a limited time due to the limitation of the scenario described in this research, whether or not the optimal value is obtained. On the other hand, EDA is easy to converge to a local optimum as other evolutionary algorithms. However, its probability can be reduced by increasing population diversity. Population diversity means that when the EDA falls into a local optimum, some individuals are still kept (these individuals may not have the highest fitness value in this iteration), and these individuals may jump out of the local optimal area and help to search the global optimal direction area.

Using multipopulation strategies can effectively increase population diversity of EDA. Branke et al. [31] proposed that the whole search space can be divided into several regions, and different populations should be arranged in different regions. However, they use distance to divide the search region, which leads to a large amount of calculation and is not suitable for the real-time situation. We adopt three populations (i.e., three probability models) to implement the static optimization process. The search region is divided by the individual fitness value to reduce the amount of calculation and make full use of the characteristics of the individuals with high fitness values to make our mechanism quickly converge toward the optimal value, while reducing the probability of converging to the local optimum.

The static optimization process generates NP individuals according to the initial probability model to form the initial population. It is divided into three subpopulations consisting of superior, general, and poor solutions, which are represented by the symbols $\mathrm{pop}_b(g)$, $\mathrm{pop}_m(g)$, and $\mathrm{pop}_w(g)$ (the numbers in parentheses represent the index of generation population), respectively. The three probability models are updated as $P_b^i(g)$, $P_m^i(g)$, and $P_w^i(g)$ (i is the i th QoS attribute of the component service) based on the individual characteristics of the three subpopulations. A new generation of population is sampled according to the three updated probability models. It is redivided into three subpopulations consisting of superior, general, and poor solutions, which are used to reupdate three probability models. Then the new models generate new population, and the mechanism iterates in this way until the termination condition is met.

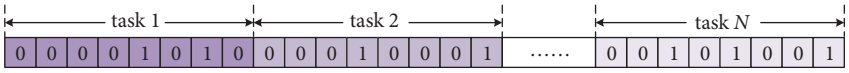


Figure 3: A chromosome of composite scheme.

The initial probability model needs to be constructed in order to generate the first population. The optimization problem in this research is a discrete problem, and each component service is independent of each other. Three probability models are used to increase population diversity, so we first build three discrete probability models: $P_b^i(0) = (p_b^{i,1}(0), p_b^{i,2}(0), \ldots, p_b^{i,M}(0))$, $P_m^i(0) = (p_m^{i,1}(0), p_m^{i,2}(0), \ldots, p_m^{i,M}(0))$, $P_w^i(0) = (p_w^{i,1}(0), p_w^{i,2}(0), \ldots, p_w^{i,M}(0))$, $(1 \le i \le N)$, where M is the index of the candidate component service. $P_b^i(0)$, $P_m^i(0)$, and $P_w^i(0)$ are M-dimension vectors and represent the probability that each candidate component service in the service class will be selected.

For example, we do not know which candidate component service in the service class can form the optimal composite service, so we assume that they are selected with equal probability $1/M$. Therefore, we can construct the initial probability model as $P_b^i(0) = P_m^i(0) = P_w^i(0) = (1/M, 1/M, \cdots 1/M)(1 \le i \le N)$. The probability of selecting a candidate component service in its service class in the superior probability model converges to 1 through the iterative process of the EDA, and the probability of selecting other component services converges to 0. General and poor probabilistic models are utilized to increase population diversity and avoid missing optimal solutions.

During the gth iteration of the multipopulation EDA, new individuals are sampled according to the three probability models updated during the $g-1$th iteration, i.e., $P_b^i(g-1)$, $P_m^i(g-1)$, and $P_w^i(g-1)$, and three subpopulations are obtained: $\text{pop}_b'(g)$, $\text{pop}_m'(g)$, and $\text{pop}_w'(g)$. The fitness values of the individuals in each subpopulation are calculated, and then all of individuals are sorted according to the values. According to the proportion, they are redivided into three subpopulations, i.e., the superior, general, and poor subpopulations $\text{pop}_b(g)$, $\text{pop}_m(g)$, and $\text{pop}_w(g)$. The gth generation subpopulations are obtained. Using their individual characteristics, update the respective probability models to get the probability models of the gth generation: $P_b^i(g), P_m^i(g), P_w^i(g)$.

The following is the update process of the three probability models in the gth iteration.

3.2.1. Updating the Probability Model of Superior Solutions. After calculating the fitness values of each individual in the subpopulation $\text{pop}_b(g)$, they are sorted in descending order, and the top $B(B \le NP)$ individuals are selected to form the group $\text{better}(g)$. Then update $P_b^i(g-1)$ by statistics, smoothing, and adding forgetting factor to get $P_b^i(g)$.

(i) Statistics. Count the values of B individuals in $\text{pop}_b(g)$. If the candidate component service $s_{i,j}$ in the service class S_i is adopted $s(0 \le s \le B)$ times, then record it as $p_b^{',i,j}(g) = s/B$.

(ii) Smoothing. For the above statistical process, sometimes the value of $p_b^{',i,j}(g)$ is large, and sometimes it is 0, so the statistical probability value needs to be smoothed. Based on the principle that the QoS values are relatively close, the probability should not be much different, and the statistical results are smoothed in our research. The model $p_b^{',i,j}(g)$ is updated by formula $p_b^{'',i,j}(g) = \sum_{h=1}^{M} e^{(-|j-h|)} p_b^{',i,j}(g)$. Then $p_b^{'',i,j}(g)$ is normalized to get $p_b^{'',i,j}(g)$.

(iii) Adding forgetting factor. Forgetting factor θ is added to eliminate the excessive influence of the $g-1$th generation on the evolutionary direction of the offspring and avoid the algorithm falling into local optimum. After adding the forgetting factor $p_b^{i,j}(g) = \theta p_b^{i,j}(g-1) + (1-\theta) p_b^{'',i,j}(g)$, the gth generation optimal solution probability model is obtained after the above updating process, i.e., $P_b^i(g) = (p_b^{i,1}(g), p_b^{i,2}(g), \ldots, p_b^{i,M}(g))(1 \le i \le N)$.

3.2.2. Updating the Probability Model of Poor Solutions. The updating process is similar to that of superior solutions for the poor solutions probability model. The difference is that individuals are sorted in ascending order after calculating the fitness value of individuals in subpopulation $\text{pop}_w(g)$, and the top W ones are selected to form group $\text{worse}(g)$. Then, use the same process as the superior solution to calculate the gth generation probability model $P_w^i(g) = (p_w^{i,1}(g), p_w^{i,2}(g), \ldots, p_w^{i,M}(g))(1 \le i \le N)$.

3.2.3. Updating the Probability Model of General Solutions. If we use the same process as the superior or poor solutions to update the general solutions probabilistic model, it will converge to the superior solutions or the poor solutions. The model will fail to represent the characteristics of general solutions, thus losing the population diversity. Therefore, this study adopts the following strategies to update the general solutions probability model $P_m^i(g-1)$. First select some individuals with poor fitness values from the superior subpopulation $\text{pop}_b(g)$, then select some individuals with superior fitness values from the poor subpopulation $\text{pop}_w(g)$, and finally form the general subpopulation $\text{middle}(g)$ with the individuals in the general subpopulation $\text{pop}_m(g)$. The following formula is adopted to update $P_m^i(g-1)$ according to the characteristics of $\text{middle}(g)$ and the gth superior solutions probability model $P_b^i(g)$.

$$p_m^{i,j}(g) = \begin{cases} 0.5 + \alpha\left(p_b^{i,j}(g) - 0.5\right); & p_b^{i,j}(g) > 0.5, \\ 0.5 - \alpha\left(0.5 - p_b^{i,j}(g)\right); & p_b^{i,j}(g) < 0.5, \\ 0.5; & p_b^{i,j}(g) = 0.5, \end{cases} \tag{5}$$

where $0 \le i \le N$, $0 \le j \le M$, α is used to control the closeness of $P_m^i(g)$ and $P_b^i(g)$, and the larger the value is, the closer it is. It can be seen from formula (5) that the general solution probability model $P_m^i(g)$ is always between the superior solution one $P_b^i(g)$ and the central probability vector (0.5, 0.5, . . ., 0.5), which can describe the characteristics of the general solution and maintain the diversity.

Our research uses three subpopulations to establish three probability models and achieve the division of search space. Different subpopulations have different effects and optimization goals. The superior solution subpopulation gathers individuals with higher fitness value from three subpopulations and eventually converges to the optimal solution. The poor solution subpopulation will eventually converge to the worst solution. But its purpose is to find some solutions with better fitness value in the region with poor fitness and add them in the superior solution subpopulation to the next iteration process. It is used to increase population diversity and prevent missing optimal solutions. The general solution subpopulation helps the superior solution subpopulation to quickly find individuals with better fitness and adds them to the superior solutions subpopulation to speed up the optimization process.

3.3. Dynamic Optimization Process. If the deployment location of the component service changes, only the intermediate optimization results, not the optimal solution, can be obtained through the static optimization process. The optimization process needs to be continued to provide the optimal composite service to users who will pass this location in the future. However, the static optimization process can only end within a limited time for real-time requirements. If the optimization process is restarted every time, it is time consuming and it is difficult to obtain the optimal solution. Therefore, we can save the intermediate optimization results obtained by the static optimization process to continue the optimization process.

However, the subsequent optimization process will fall into local optimum if the number of intermediate solutions we employed is just one. Naturally, we try to use multiple intermediate solutions to improve the optimization process. The optimized intermediate solutions cannot be saved indefinitely due to the limitation of storage space. Otherwise, a complex storage space management scheme is also required to reuse the excessive optimized intermediate solutions.

According to above analysis, the steps of CSP-EC proposed in our study are as follows after the component service deployment of a certain location changes.

Step 1: The intermediate optimization results of the m users who first passed through the location are saved. These m users can perform optimization process in parallel and end them within a limited time (see Algorithm 1 for implementation). m users use the optimized intermediate solution obtained by themselves to alleviate traffic jam.

In Algorithm 1, the time complexity of lines 3–7 is all $O(1)$. The line 8 uses the quicksort with lower time complexity, so the average time complexity is $O((n_1 + n_2 + n_3)\log(n_1 + n_2 + n_3))$. The lines 9–15 are a WHILE loop and ends within a limited time, so the number of loop is limited. Therefore, the time complexity of the internal statements determines the time complexity of WHILE loop. The time complexity of line 10 is $O(1)$. In line 11, the time complexity of statistical process is $O(n_1^2)$, and the time complexity of smoothing and adding forgetting factor is all $O(n_1)$. Therefore, the time complexity of line 11 is $O(n_1^2)$. Similar to line 11, the time complexity of line 12 is $O(n_2^2)$, and the line 13 is $O(n_3^2)$. The time complexity of line 14 is $O((n_1 + n_2 + n_3)\log(n_1 + n_2 + n_3))$. Therefore, the time complexity of WHILE loop is $O(n_1^2 + n_2^2 + n_3^2 + (n_1 + n_2 + n_3)\log(n_1 + n_2 + n_3))$. Based on the above analysis, the time complexity of Algorithm 1 is $O(n_1^2 + n_2^2 + n_3^2 + (n_1 + n_2 + n_3)\log(n_1 + n_2 + n_3))$.

Step 2: After saving the intermediate optimization results of the first m users, CSP-EC will reuse them to continue the optimization process until they find the optimal solution (see Algorithm 3 for implementation), and later users can use the optimal composite service to achieve the alleviate traffic jam. However, these optimized intermediate solutions can also be divided into better and worse according to their fitness values, and the better ones should be reused with a greater probability to speed up the optimization process. It is not appropriate to reuse them according to the uniform distribution rule, and the probability of being selected should be proportional to its fitness. Therefore, we use the roulette method to choose to reuse intermediate optimization results [32]. The steps can be summarized as follows:

(1) Calculate the probability that m optimized intermediate solutions are selected. The calculation method adopts the following formula:

$$P(x_i) = \frac{f(x_i)}{\sum_{j=1}^{m} f(x_j)}. \tag{6}$$

$P(x_i)$ represents the probability that the ith optimized intermediate solutions are selected. $f(x_i)$ is the fitness value of ith optimized intermediate solutions. Therefore, $\sum_{i=1}^{m} P(x_i) = 1$.

(2) Calculate the cumulative probability of m optimized intermediate solutions.

$$p_i = \sum_{j=1}^{i} P(x_j). \tag{7}$$

(3) Construct the interval in which m intermediate solutions are selected.

$$s_i = \begin{cases} (0, p_i]; & i = 1 \\ (p_{i-1}, p_i]; & 1 < i < m. \\ (p_i, 1]; & i = m \end{cases} \tag{8}$$

(1) **Input**: the QoS of component service after the deployment location changes
(2) **Output**: the optimized intermediate solution better obtained in limited time
(3) $P_b^i(0) = (1/M, 1/M, \cdots 1/M)(1 \le i \le N)$;
(4) $P_m^i(0) = (1/M, 1/M, \cdots 1/M)(1 \le i \le N)$;
(5) $P_w^i(0) = (1/M, 1/M, \cdots 1/M)(1 \le i \le N)$; \\ 0th generation probability model of three sub-populations are initialized.
(6) $g = 0$; \\The variable g is initialized to zero.
(7) The number of the superior sub-population is n_1. The number of general sub-population is n_2. The number of poor sub-population is n_3;
(8) Three sub-populations are obtained according to their models sampling, respectively. The fitness value of three sub-population individuals are calculated, and sorted together by descending order. The top n_1 individuals are selected to form the superior solutions sub-population $\text{pop}_b(1)$. The middle n_2 individuals are selected to form the general solutions sub-population $\text{pop}_m(1)$. The last n_3 individuals are selected to form the poor solutions sub-population $\text{pop}_w(1)$;
(9) **while** limited time **do**
(10) $g = g + 1$;
(11) For $\text{pop}_b(g)$, select the top $B(B \le n_1)$ individuals form the group better(g), and then update $P_b^i(g)$ by the rules of statistics, smoothing and adding forgetting factor;
(12) For $\text{pop}_w(g)$, select the last $W(W \le n_3)$ individuals form the group worse(g), and then update $P_w^i(g)$ by the rules of statistics, smoothing and adding forgetting factor;
(13) First select some individuals with poor fitness values from the superior sub-population $\text{pop}_b(g)$, then select some individuals with superior fitness values from the poor sub-population $\text{pop}_w(g)$, finally form the general sub-population middle(g) with the individuals in the general sub-population $\text{pop}_m(g)$. Update $P_m^i(g)$ according to formula (5);
(14) Three new sub-populations are obtained according to their updated models sampling, respectively. The fitness value of three sub-population individuals are calculated, and sorted together by descending order. The top n_1 individuals are selected to form the g +1th superior solutions sub-population $\text{pop}_b(g+1)$. The middle n_2 individuals are selected to form the g +1th general solutions sub-population $\text{pop}_m(g+1)$. The last n_3 individuals are selected to form the g +1th poor solutions sub-population $\text{pop}_w(g+1)$;
(15) **end while**
(16) Save the optimized intermediate solution better in memory space;

Algorithm 1: The algorithm for the first m users to seek optimized intermediate solution.

(4) Calculate the selected optimized intermediate solutions. Generate a uniformly distributed random number r in [0, 1]. The selection object is determined according to the interval to which r falls.

The implementation of the optimized intermediate solution selection algorithm is as Algorithm 2.

In Algorithm 2, when the generated random number is between $(\text{sum}, \text{sum} + P(x_i)]$, i is considered to be selected, so the probability that i is selected is $P(x_i)$. The time complexity of Algorithm 2 is $O(m)$.

The algorithm for users who need to reuse optimized intermediate solution is as follows.

In Algorithm 3, the time complexity of line 3 is $O(m)$, and that for lines 4–5 is $O(1)$. The time complexity of line 6 is $O(n_1^2 + n_2^2 + n_3^2 + (n_1 + n_2 + n_3)\log(n_1 + n_2 + n_3))$. Therefore, the time complexity of Algorithm 3 is $O(n_1^2 + n_2^2 + n_3^2 + (n_1 + n_2 + n_3)\log(n_1 + n_2 + n_3))$.

4. Experiments

In this section, extensive experiments are conducted based on the real-world dataset to illustrate the effectiveness of the proposed mechanism. Specifically, we compare the proposed mechanism with several mechanisms regarding the quality of the optimized intermediate solution and the optimal solution, the stability of the optimal solution, and the computation time of the optimal solution.

4.1. Experiment Setup. A QoS dataset for component services is required to perform experiments. We adopt the QWS2 (https://www.uoguelph.ca/qmahmoud/qws/), a Web services QoS dataset, which includes 2507 real Web services, and each of them includes 9 QoS attributes. We only focus on two QoS attributes, i.e., reliability and response time. The EC environment is constructed according to the distribution of base stations in Shanghai Telecom dataset (https://sguangwang.com/TelecomDataset.html). Shanghai is divided into several 1 km × 1 km areas as shown in Figure 4. The adopted optimal composite service calculation mechanism for different location is the same after component service deployment changes. Therefore, we only select one area in Figure 4 to run our CSP-EC and verify its effectiveness. The experiments were conducted on a PC with Intel(R) Core (TM) i5-4210U 2.39 GHz CPU, 8.0 GB of RAM, Windows 8.1 64 bit, and Matlab R2013a.

We generated our composite services by inserting tasks and control structures. We generated a number of candidate services with different QoS attributes for each task based on the QWS2 dataset. As far as we know, there are no similar scenarios in other studies, so we cannot find an approach to directly compare with our mechanism. However, GA and Particle Swarm Optimization (PSO) are also evolutionary algorithms, and they are generally adopted by the research of composite service [33, 34]. Therefore, we use GA and PSO as our comparative approaches. In addition, we use the standard EDA as another comparative approach.

(1) **Input**: $P(x_i)(1 \le i \le m)$.
(2) **Output**: Intermediate solutions selected to be reused
(3) sum = 0;
(4) r = **RANDOM**(0, 1); \\r is a random number from 0 to 1
(5) **for** i = 1 to m **do**
(6) sum = sum + $P(x_i)$;
(7) IF $r \le$ sum return i;
(8) **end for**;

Algorithm 2: The optimized intermediate solution selection algorithm.

(1) **Input**: the QoS of component service after the deployment location changes
(2) **Output**: the optimized intermediate solution better obtained in limited time
(3) Run Algorithm 2 to select reused optimized intermediate solutions for user;
(4) $g = 0$; \\The variable g is initialized to zero
(5) Read three probability models of sub-population from memory space, and obtain $P_b^i(0), P_m^i(0), P_w^i(0)$;
(6) The remaining steps are the same as lines 8–16 of Algorithm 1;

Algorithm 3: The algorithm for users who need to reuse optimized intermediate solution.

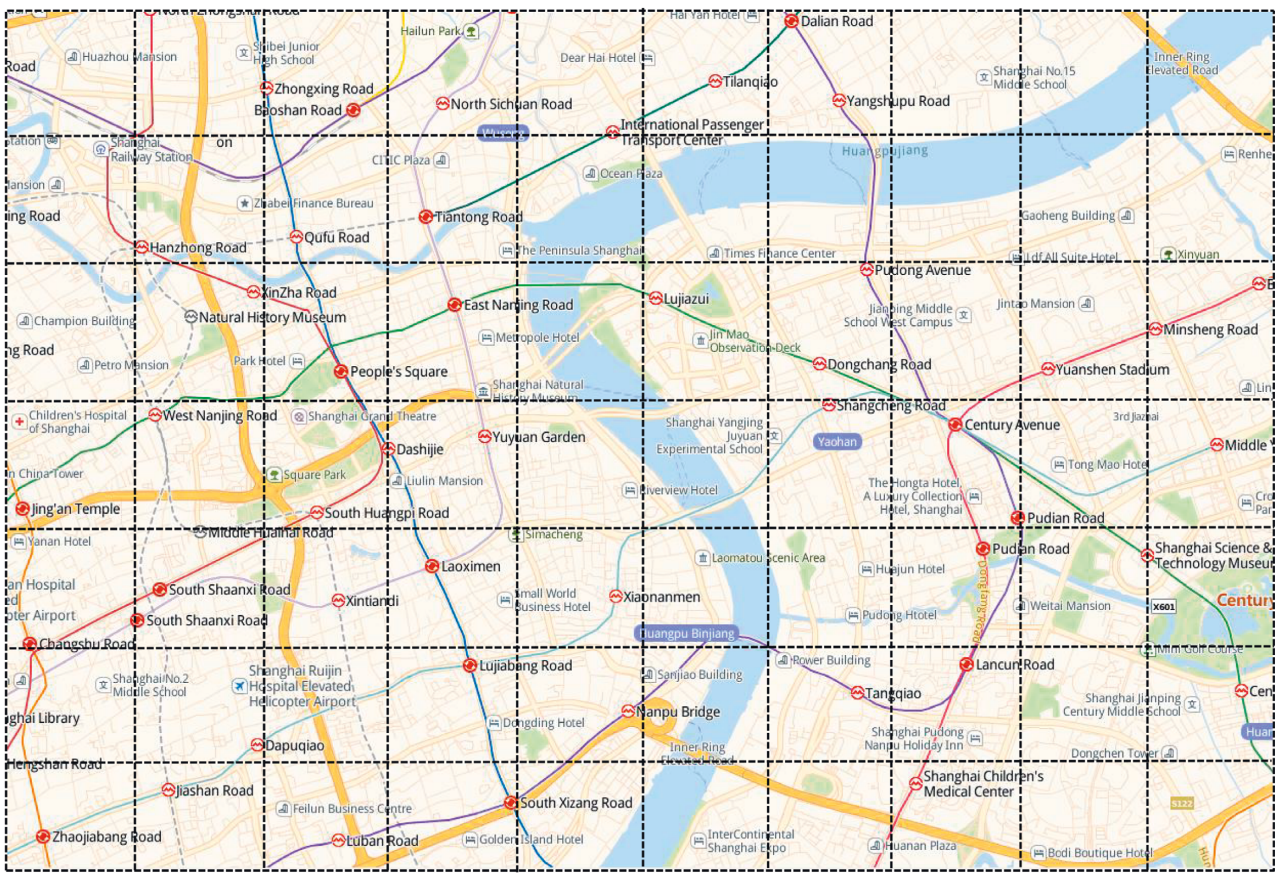


Figure 4: Area division of partial Shanghai urban area.

If the task has n candidate component services, the parameter settings of GA were set as follows: the population size is 0.25 n, the crossover rate is 0.7, and the mutation rate is 0.3. The parameter settings of PSO were set as follows: the number of initial particles is 20, and the learning factor is 2. The standard EDA parameters were set as follows: the

population size is 0.25 n, and the ratio of selected dominant individuals is 0.5. Our CSP-EC parameters were set as follows: the population size is 0.25 n, the superior solutions subpopulation size is $0.6 \times 0.25n$, the general and poor solution subpopulation size is $0.2 \times 0.25n$, the forgetting factor θ is 0.5, the closeness parameter of the general and superior solution subpopulation is 0.5, and the number of saved optimized intermediate solution m is 0.25 n.

The average vehicle speed of first-tier cities in China in 2016 was 24.7 km/h according to data of the "2016 Smart Travel Big Data Report released" by Didi (https://www.199it.com/archives/556606.html); i.e., it was about 7 m/s. The roads in the cities have different length. We set the time period for each user to run the optimization process to 2 seconds to have enough time to lane change and traffic diversion. Because the moment when users pass the same location is subject to Poisson distribution, the moment that each vehicle starts the optimization process is also subject to Poisson distribution.

We designed three types of experiments to study the effect of different problem size and response time constraints on the performance of all the approach. (1) Composite services include 5 tasks, and time constraint is $5 * 200 = 1000$ ms. The number of candidate component services for each task is changed from 100 to 1000 in accordance with the rule of 100 increments each time. This type of experiment is called Experiment A, which studies the impact on the performance of the approach after the problem size increases with the increase in the number of candidate component services. (2) Each task has 100 candidate component services, and the task number of composite service is changed from 5 to 50 in accordance with the rule of 5 increments each time. The time constraint is the task number times 200 ms. This type of experiment is called Experiment B, which studies the impact on the performance of the approach after the problem size increases with the increase in the number of tasks. (3) Composite services include 5 tasks, and each task has 500 candidate component services. The response time of composite service is changed from 600 ms to 1500 ms in accordance with the rule of 100 ms increments each time. This type of experiment is called Experiment C, which studies the impact on the performance of the approach for different response time constraint. Each type of experiments was run 60 times, and the average value was used as the final experiment result.

4.2. Quality Evaluation of Optimized Intermediate Solutions. When the deployment of component service changes in a certain location, optimization intermediate solutions of the first 0.25 n users need to be saved for the future users reuse. Their solutions are obtained by running the optimal process within two seconds, so the solutions obtained after the first two seconds are the worst after the deployment change. Therefore, we first compared the quality of optimized intermediate solutions obtained after the first two seconds. The quality can be calculated by their fitness functions (i.e., formula (3)). The higher the fitness value, the better the solution, and vice versa.

We calculated optimized intermediate solutions by our mechanism and the comparative approaches according to the setting of three types of experiments in Section 4.1. The results are shown in Figure 5 and Table 2 (the values of Experiment B are too small to draw in Matlab, so we have tabulated the experimental results). As can be seen from the figures and table, for three kinds of experiments, the fitness values of optimized intermediate solution obtained by our method and EDA are significantly higher than those obtained by GA and PSO, and the values obtained by our method are slightly higher than EDA.

The reason why our method can obtain the best solution is that we adopt improved EDA. The idea of EDA originates from GA, but it uses a different evolutionary model compared with GA. GA simulates biological evolution at the micro level. Its evolutionary process is the selecting operations and restructuring operations of a mass of genes for population chromosome, and the optimal solution can be found step by step by combining more good chromosomes. However, due to the existence of crossover and mutation operations, it is easy to destroy the structure of the selected and restructured better chromosomes, thus prolonging the optimization process. PSO does not include crossover and mutation operations, thus shortening the optimization process. Therefore, optimized intermediate solutions obtained by PSO are slightly better than GA.

EDA simulates biological evolution at the macro level. It can make statistical analysis for the external biological performance characteristics and update the probability model of evolution, so it can control the search direction of the algorithm globally. Moreover, EDA does not include crossover and mutation operations, so it can control the optimal process at the macro level and quickly converge to the optimal solution. Therefore, fitness values obtained by EDA and our method are much higher than that of GA and PSO. In addition, our method adopts the mechanism of multipopulation evolution and forgetting factor to improve the standard EDA, so the fitness values of our method are higher than that of standard EDA.

4.3. Quality Evaluation of Optimal Solutions. A key indicator of judging the pros and cons of a method is to study the quality of the final solution. This section compares and analyzes the optimal solution obtained by the four methods. The results are shown in Figure 6 and Table 3 (the values of Experiment B are also too small to draw in Matlab, so we have tabulated the experimental results). As can be seen from the figures and table, the fitness values of the optimal solution obtained by EDA and our method are similar to optimized intermediate solution, which are significantly higher than those obtained by GA and PSO, and the values obtained by our method are higher than EDA.

The quality of the optimal solutions obtained by our method is the best compared with the comparison methods. The reasons are analyzed as follows: EDA, PSO, and GA are all evolutionary algorithms, and they can easily fall into a local optimum and thus cannot converge to the optimal

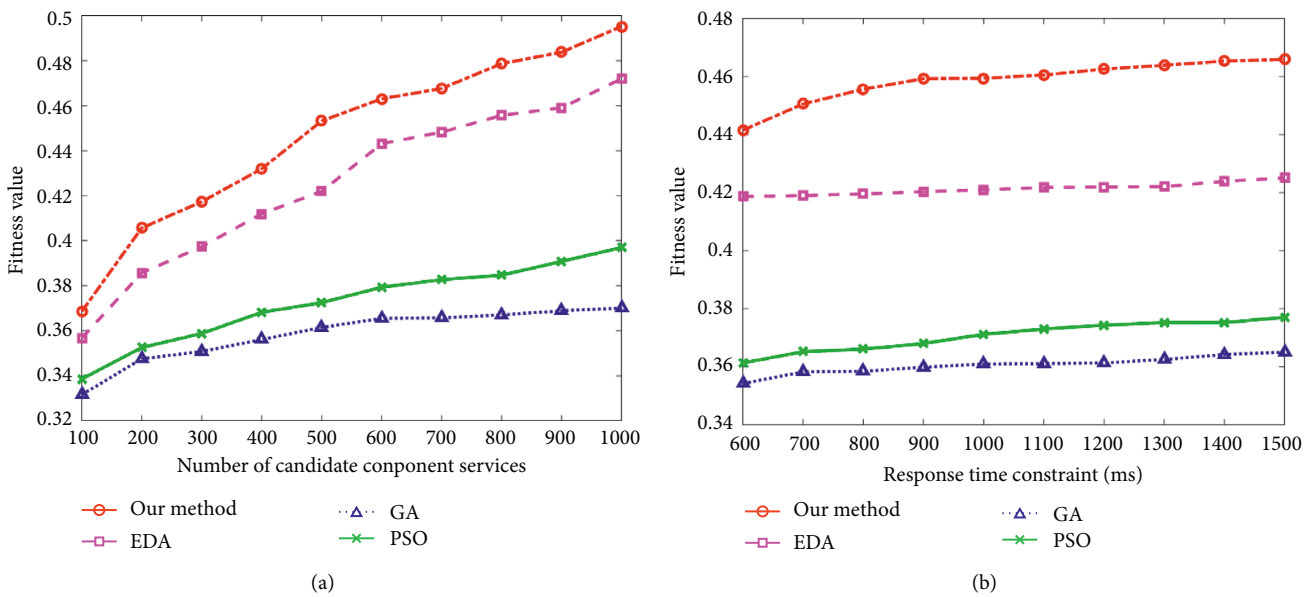


Figure 5: Experimental results for fitness values of optimized intermediate solution. (a) The results of Experiment A. (b) The results of Experiment C.

Table 2: The fitness values of optimized intermediate solution for Experiment B method; the number of tasks.

Method	The number of task									
	5	10	15	20	25	30	35	40	45	50
Our method	0.3624	0.1306	0.0413	0.0137	0.0389	0.0014	0.0004	0.0001	0.00003	0.00001
EDA	0.3493	0.1061	0.0346	0.0081	0.0056	0.0009	0.0002	0.00008	0.00001	0.000006
PSO	0.3384	0.0877	0.0276	0.0051	0.0013	0.0003	0.00005	0.00002	0.000008	0.0000009
GA	0.3293	0.0790	0.0178	0.0036	0.0007	0.0001	0.00003	0.000006	0.000001	0.0000002

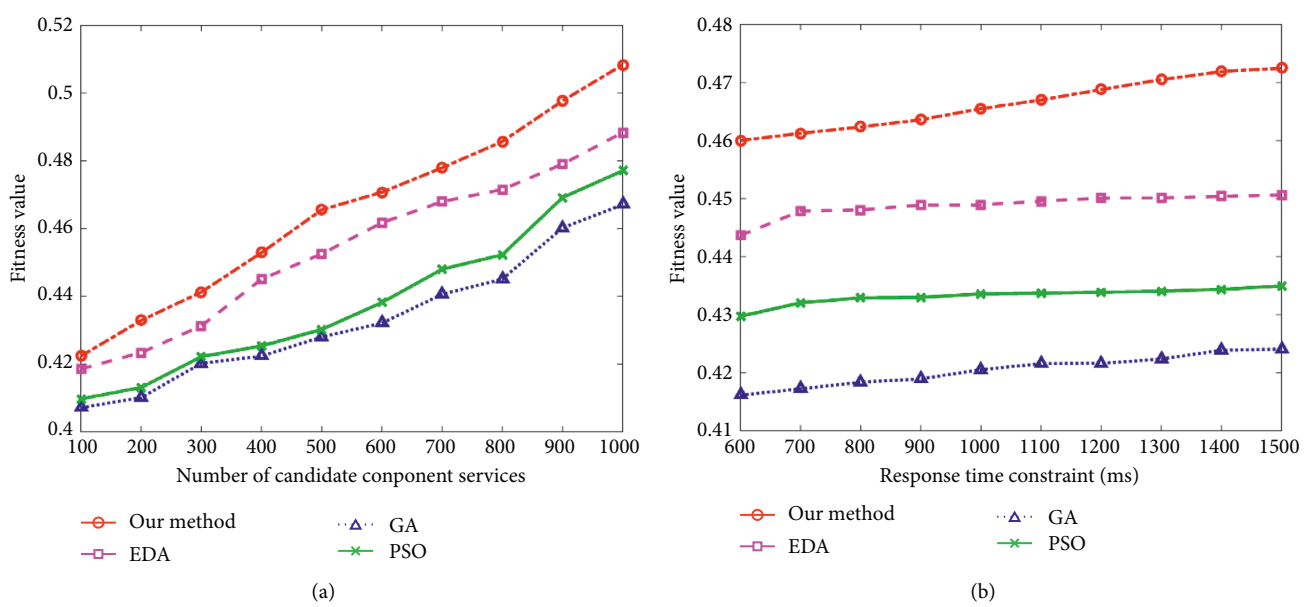


Figure 6: Experimental results for fitness values of optimal solution. (a) The results of Experiment A. (b) The results of Experiment C.

solution. Evolutionary algorithm selects the individuals of the next generation based on the fitness value. The individuals with lower fitness value are often discarded, but perhaps the best solution can be found by exploring in the direction of this discard individual. Therefore, it is necessary to reduce the probability of the evolutionary algorithm

Table 3: The fitness values of optimal solution for Experiment B method; the number of tasks.

Method	The number of task									
	5	10	15	20	25	30	35	40	45	50
Our method	0.4071	0.1713	0.0587	0.0197	0.0559	0.0024	0.0008	0.0002	0.00007	0.00002
EDA	0.3930	0.1513	0.0463	0.0108	0.0316	0.0019	0.0006	0.00009	0.00004	0.000009
PSO	0.3901	0.1229	0.0242	0.0054	0.0028	0.0006	0.00009	0.00002	0.000008	0.000002
GA	0.3826	0.1030	0.0234	0.0050	0.0010	0.0002	0.00005	0.00002	0.000002	0.0000002

falling into a local optimum by increasing the population diversity. Our method adopts two kinds of strategies to increase the population diversity, one through multipopulation evolution and the other through adding a forgetting factor. Multipopulation evolution keeps individuals with lower fitness value to form poor solutions subpopulation and continues to explore the direction of them. If individuals with higher fitness value appear, they are added to the superior subpopulation to avoid missing the optimal solution. The next generation population can be prevented from being affected by the current generation too much by adding a forgetting factor, thereby reducing the probability of falling into a local optimum. Therefore, the quality of optimal solutions obtained by our method is the best.

4.4. Stability Evaluation of Optimal Solutions. In addition to comparing the pros and cons of the final solution, it is necessary to study the stability of multiple solutions to measure the pros and cons of a method. The stability can be measured by dispersion. The smaller the value, the higher the solution stability, and vice versa. We adopt standard deviation to measure the dispersion of the solutions. The dispersion formula is as follows:

$$\text{Dispersion} = \sqrt{\frac{\sum_{i=1}^{n}\left(f_i - \overline{f}\right)^2}{n}}, \tag{9}$$

n is number of solutions, f_i is the fitness value of the solution obtained at the i th time, and $\overline{f}$ is the average value of n solutions.

According to the three types of experiments in Section 4.1, each of them was performed 60 times. Therefore, we studied the dispersion of the 60 solutions in this section, and the results are illustrated in Figure 7. The dispersion of the solution obtained by our method is the smallest compared with three comparison methods for the three types of experiments.

Our method divides the solution space by three subpopulations. It fully explores all directions in which the optimal solution may appear, and all directions are saved by probability models. That is to say, our method keeps the solution diversity, so that the obtained solution stably converges to near the optimal solution, and the probability of falling into a local optimum is very low. It can also be seen from Figure 7(b) that the gap between the dispersion of our method and the comparison method ones is getting larger and larger with increasing the number of tasks. This is because when the problem scale is small, the probability of the comparison method falling into a local optimum is still relatively small. However, with the problem scale increase, the probability becomes larger and larger, which causes the instability of the solution. Therefore, our method is more scalable.

4.5. Evaluation on the Running Time. We compared the running time of our method with that of the comparative methods to further verify the efficiency of our method. The shorter running time, the better the method. The experimental results are illustrated in Figure 8. It can be seen from the figure that running time of our method is the shortest for three types of experiments. When the number of tasks or candidate component services is fewer (i.e., the problems size is smaller), we can observe that running time of our method is not much different from the compared methods through Figures 8(a) and 8(b). The reason for this is that our method adopts multipopulation mechanism, and it takes time to divide population. When the problem size is small, the running time of three comparison methods is relatively short. The time cost of dividing population particularly affects the overall running time of our method, but in general it is still shorter than the compared methods. However, as the size of problem increases, the time cost of dividing population becomes negligible, and the advantage of our method becomes increasingly apparent. Therefore, our method is more suitable to be applied to scenarios with a large-scale problem; that is, its scalability is better.

It can be observed from Figure 8(c) that the running time is longer when the response time constraint is shorter. The running time is gradually reduced with the response time constraint increase, and it stabilizes after the constraint is greater than 1000 ms. The reason for this phenomenon is that the optimal problem in this study has a response time constraint. During updating the population, all methods need to discard the individuals if they do not meet the constraint. Only when all the individuals in the population meet the constraint, the population can be successfully updated. The probability of generating unsatisfied constraint individual will be high when the response time constraint is lower. To generate the satisfied individual it is needed to regenerate one, so the running time of the method is increased. As the response time constraint increases, the probability of generating the unsatisfied individual is slowly decreasing, and the running time of the method decreases accordingly. When the constraint is greater than 1000 ms, the probability of the generating unsatisfied individual tends to stabilize, and the running time of the method also tends to stabilize. Therefore, the response time constraint is set as the number of tasks times 200 ms when designing Experiments A and B.

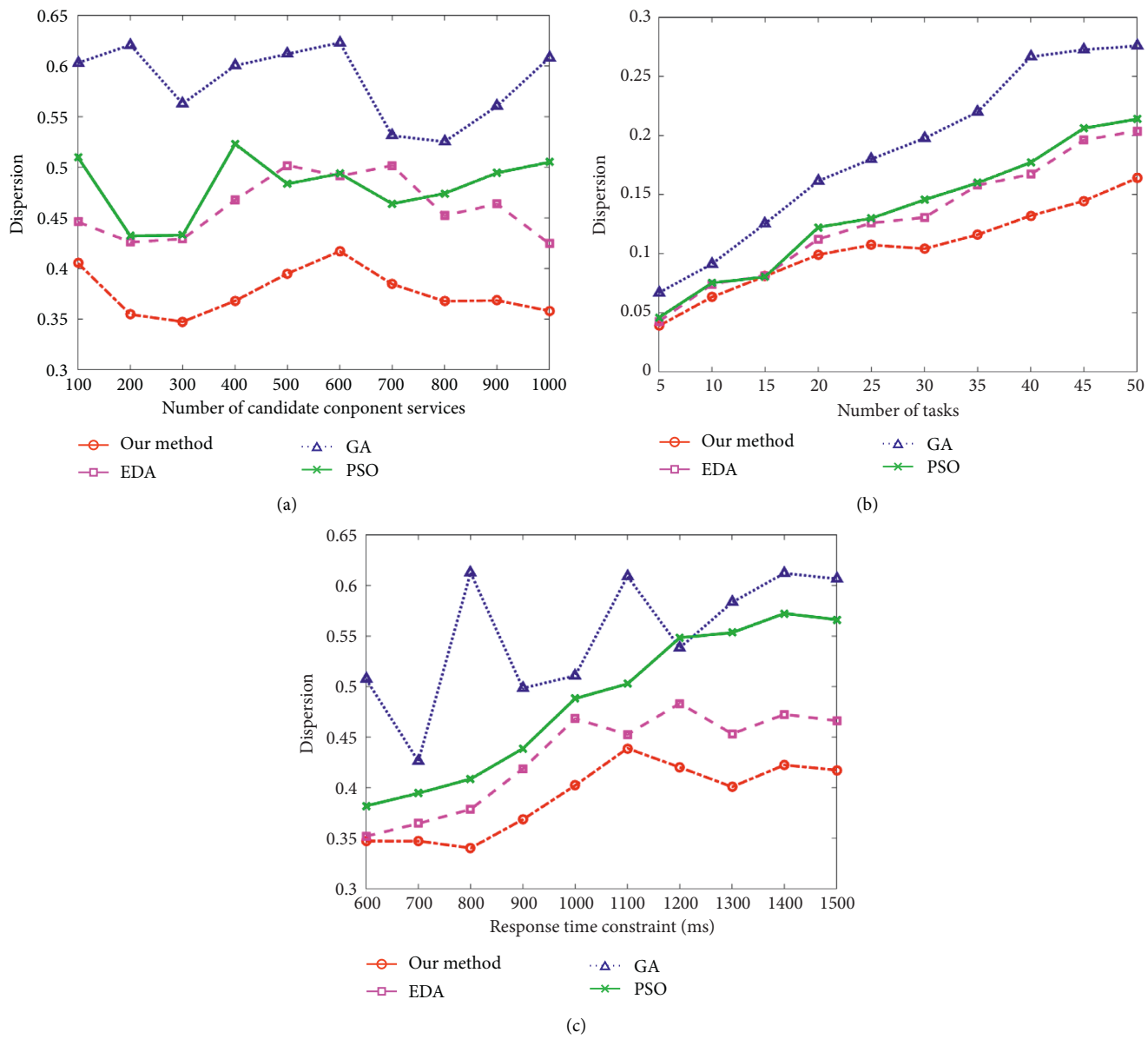


Figure 7: Experimental results for optimal solution dispersion. (a) The results of Experiment A. (b) The results of Experiment B. (c) The results of Experiment C.

Based on the above analysis, the running time of Experiments A and B also increases with the scale of the problem increase, so they show an upward trend on the figure. Meanwhile, with the response time constraint increase, the running time of Experiment C decreases. The running time tends to stabilize after the response time reaches 1000 ms. Therefore, it first decreases in the figure and then no longer changes significantly.

5. Related Works

Our research is a composite service provision problem in EC, and it is a dynamic multiobjective optimization problem with location awareness, high real-time requirement, and high reliability requirement. Therefore, the main similar literature is analyzed in this section.

Imed et al. introduce a formal model of the Web service configuration and its correctness requirements that permit ensuring the correct Web service execution from functional and transactional points of view. However, the proposed method is only suitable for solving the optimal composition service when the QoS attribute values of the component services are unchanged. Once the QoS value changes, the obtained combined service may not be the optimal combined service. Deng et al. [35] also study the optimal composition service problem with the unchanged QoS attribute values. This literature could form service compositions that not only satisfied both the time constraints and QoS constraints in a mobile service composition, but also ensured the composition to be executed successfully to the greatest extent in the uncertain mobile environment.

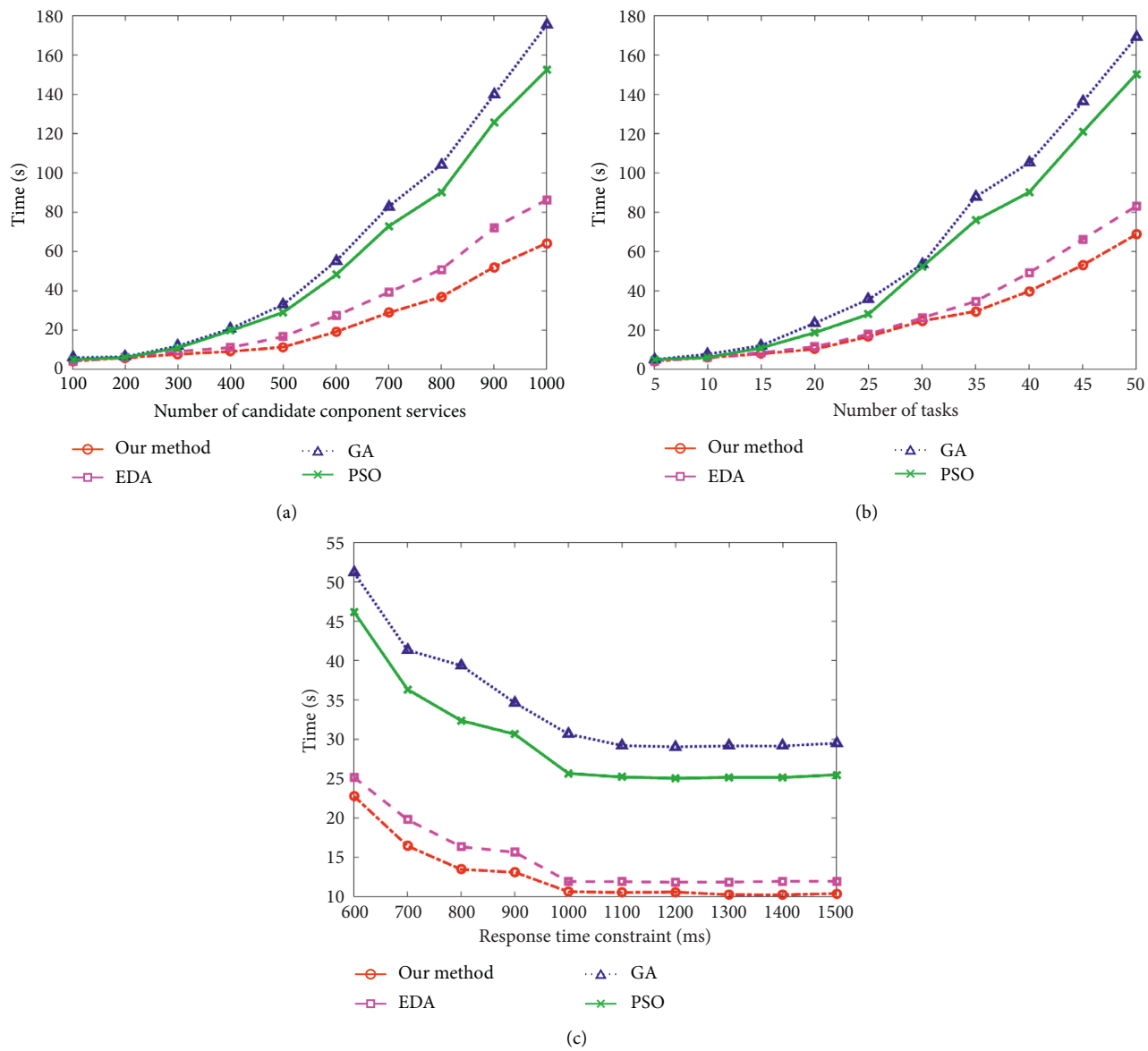


Figure 8: Experimental results comparison for optimal solution running time. (a) The results of Experiment A. (b) The results of Experiment B. (c) The results of Experiment C.

Lv et al. [36] proposed an event driven continuous query approach, i.e., QSynth, to intelligently cope with different types of dynamic services and thus enable evolution of service composition. Moreover, they generalized a new graph problem: dynamic single-source optimal Directed Acyclic Graphs problem. It was argued that API recommendation for framework evolution, as another typical application, could also be modeled and addressed efficiently using the proposed new graph approach. However, QSynth adopted a more complicated data structure, and processing this kind of data structure required a large amount of calculation. It also needed to store intermediate results. When the problem size is large, the required storage space will be much.

Yang and Hu [37] considered that service providers often exaggerate the QoS attributes of the component services they provide in the dynamic Internet environment. Runtime adaptations of composite service execution plan become very important to recover from Web service failure or improve the overall QoS attribute of composite service. Therefore, they proposed a novel empirical approach to accelerate QoS-aware runtime adaptation. Based on historical records, they use Support Vector Machines to capture the relationship between candidate services and adaptation scenarios which is used at runtime to predict the probabilities that candidate services will be used for upcoming adaptation scenarios. Then candidate services are pruned based on these probabilities estimates to reduce the search space. However, this literature only considered the possibility of alternative schemes meeting global QoS requirements and ignored the diversity of alternative schemes. Diversity can effectively change the search direction, so

avoiding falling into a local optimum to find the global optimal solution.

The ability to manage service changes and exceptions during composite service execution is a vital requirement for the dynamic and volatility environment. Barakat et al. [38] presented a novel adaptive execution approach, which efficiently handled service changes occurring at execution time. The adaptation was performed as soon as possible and in parallel with the execution process, thus reducing interruption time, increasing the chance of a successful recovery, and producing the most optimal solution according to the current environment state. However, the method proposed in this literature relies heavily on the QoS distribution of component services and some assumptions, so the scalability is poor.

6. Conclusion and Future Work

CSP-EC is proposed in this paper. It is triggered when the component service deployment changes at a certain location and ended when the optimal solution is obtained. CSP-EC is a gradual optimal process and adopts multipopulation EDA. It calculates optimized intermediate solutions within a limited time of the first m users after the deployment changes and saves them for reuse by later users. The roulette selection mechanism is used to select the reused optimized intermediate solutions, so the reused probability of the solutions with higher fitness value is increased, and the optimization process is accelerated. The experimental results show that the proposed mechanism can achieve a higher quality, a better stability, and a shorter execution time compared with other approaches.

Although the proposed mechanism has certain advantages, it also has some disadvantages. (1) CSP-EC is not proved to be better than the comparative approach through strict mathematics, but it is only verified from the experimental results. (2) CSP-EC is implemented only in the simulation experiment environment. How it is deployed and behaves in a real environment requires further verification. We will address the above shortcomings and further improve the effectiveness of CSP-EC in future work.

Data Availability

Data are available at https://www.uoguelph.ca/qmahmoud/qws/ and https://sguangwang.com/TelecomDataset.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant nos. 61902112, 62072159, and 61872149 and in part by the Science and Technology Foundation Project of Henan Province under Grant no. 222102210011.

References

[1] W. Shi, H. Sun, J. Cao, Q. Zhang, and W. Liu, "Edge computing-an emerging computing model for the internet of everything era," *Journal of Computer Research and Development*, vol. 54, no. 5, p. 907, 2017.

[2] L. Hao and L.-M. Zhou, "Evaluation index of school sports resources based on artificial intelligence and edge computing," *Mobile Information Systems*, vol. 2022, pp. 1–9, Article ID 9925930, 2022.

[3] P. Yuan and R. Huang, "Integrating the device-to-device communication technology into edge computing: a case study," *Peer-to-Peer Networking and Applications*, vol. 14, no. 2, pp. 599–608, 2021.

[4] N. Hu, X. Cen, F. Luan, L. Sun, and C. Wu, "A novel video transmission optimization mechanism based on reinforcement learning and edge computing," *Mobile Information Systems*, vol. 2021, pp. 2021–2110, Article ID 6258200, 2021.

[5] P. Yuan, Y. Cai, Y. Liu, J. Zhang, Y. Wang, and X. Zhao, "Prorec: a unified content caching and replacement framework for mobile edge computing," *Wireless Networks*, vol. 26, no. 4, pp. 2929–2941, 2020.

[6] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing-a key technology towards 5g," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, 2015.

[7] P. Yuan, Y. Cai, X. Huang, S. Tang, and X. Zhao, "Collaboration improves the capacity of mobile edge computing," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10610–10619, 2019.

[8] T. Lei, S. Wang, J. Li, and F. Yang, "A cooperative route choice approach via virtual vehicle in iov," *Vehicular Communications*, vol. 9, pp. 281–287, 2017.

[9] P. Yuan and C. Wang, "Oppo: an optimal copy allocation scheme in mobile opportunistic networks," *Peer-to-Peer Networking and Applications*, vol. 11, no. 1, pp. 102–109, 2018.

[10] H. Gao, W. Huang, and Y. Duan, "The cloud-edge-based dynamic reconfiguration to service workflow for mobile ecommerce environments: a qos prediction perspective," *ACM Transactions on Internet Technology*, vol. 21, no. 1, pp. 1–23, 2021.

[11] S. Kumar, R. Bahsoon, T. Chen, and R. Buyya, "Identifying and estimating technical debt for service composition in saas cloud," in *Proceedings of the 2019 IEEE International Conference on Web Services (ICWS)*, pp. 121–125, IEEE, Milan, Italy, July 2019.

[12] S. Wang, T. Lei, L. Zhang, C.-H. Hsu, and F. Yang, "Offloading mobile data traffic for qos-aware service provision in vehicular cyber-physical systems," *Future Generation Computer Systems*, vol. 61, pp. 118–127, 2016.

[13] X. Zhao, P. Yuan, Y. Chen, and P. Chen, "Femtocaching assisted multi-source d2d content delivery in cellular networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, p. 125, 2017.

[14] A. Zhou, S. Wang, X. Ma, and S. S. Yau, "Towards service composition aware virtual machine migration approach in the cloud," *IEEE Transactions on Services Computing*, vol. 13, no. 4, pp. 735–744, 2020.

[15] S. Wang, Y. Ma, B. Cheng, F. Yang, and R. N. Chang, "Multidimensional qos prediction for service recommendations," *IEEE Transactions on Services Computing*, vol. 12, no. 1, pp. 47–57, 2019.

[16] P. Yuan, M. Song, and X. Zhao, "Effective and efficient collection of control messages for opportunistic routing algorithms," *Journal of Network and Computer Applications*, vol. 98, pp. 125–130, 2017.

[17] X. Zhang, Z. Li, C. Lai, and J. Zhang, "Joint edge server placement and service placement in mobile edge computing," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11261–11274, 2022.

[18] P. Yuan, H. Wu, X. Zhao, and Z. Dong, "Percolation-theoretic bounds on the cache size of nodes in mobile opportunistic networks," *Scientific Reports*, vol. 7, no. 1, p. 5662, 2017.

[19] S. Wang, Ao. Zhou, M. Yang, L. Sun, C.-H. Hsu, and F. Yang, "Service composition in cyber-physical-social systems," *IEEE Transactions on Emerging Topics in Computing*, vol. 8, no. 1, pp. 82–91, 2020.

[20] S. Wang, Ao. Zhou, F. Yang, and R. N. Chang, "Towards network-aware service composition in the cloud," *IEEE Transactions on Cloud Computing*, vol. 8, no. 4, pp. 1122–1134, 2020.

[21] P. Yuan, L. Fan, P. Liu, and S. Tang, "Recent progress in routing protocols of mobile opportunistic networks: a clear taxonomy, analysis and evaluation," *Journal of Network and Computer Applications*, vol. 62, pp. 163–170, 2016.

[22] Z. Xue, L.-P. Zhao, M. Zhang, and B.-X. Sun, "Three-way decisions based on multi-granulation support intuitionistic fuzzy probabilistic rough sets," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 4, pp. 5013–5031, 2020.

[23] P. Yuan, P. Liu, and S. Tang, "Rim: relative-importance based data forwarding in people-centric networks," *Journal of Network and Computer Applications*, vol. 62, pp. 100–111, 2016.

[24] Z. Xue, M. Zhang, Y.-xiang. Li, Li-ping. Zhao, and B.-xin. Sun, "Double-quantitative generalized multi-granulation set-pair dominance rough sets in incomplete ordered information system," *Symmetry*, vol. 12, no. 1, p. 133, 2020.

[25] S. Peng, H. Wang, and Yu Qi, "Estimation of distribution with restricted Boltzmann machine for adaptive service composition," in *Proceedings of the 2017 IEEE International Conference on Web Services (ICWS)*, pp. 114–121, IEEE, Honolulu, HI, USA, June 2017.

[26] M. S. Hossain, M. Moniruzzaman, G. Muhammad, A. Ghoneim, and A. Alamri, "Big data-driven service composition using parallel clustered particle swarm optimization in mobile environment," *IEEE Transactions on Services Computing*, vol. 9, no. 5, pp. 806–817, 2016.

[27] Y. Song, L. Liu, H. Ma, and A. V. Vasilakos, "A biology-based algorithm to minimal exposure problem of wireless sensor networks," *IEEE Transactions on Network and Service Management*, vol. 11, no. 3, pp. 417–430, 2014.

[28] L. Liu, Y. Song, H. Zhang, H. Ma, and A. V. Vasilakos, "Physarum optimization: a biology-inspired algorithm for the steiner tree problem in networks," *IEEE Transactions on Computers*, vol. 64, no. 3, pp. 818–831, 2015.

[29] F. Tao, D. Zhao, Y. Hu, and Z. Zhou, "Resource service composition and its optimal-selection based on particle swarm optimization in manufacturing grid system," *IEEE Transactions on Industrial Informatics*, vol. 4, no. 4, pp. 315–327, 2008.

[30] H. Mühlenbein and G. Paaß, "From recombination of genes to the estimation of distributions i. binary parameters," in *International Conference on Parallel Problem Solving from Nature*, pp. 178–187, Springer, 1996.

[31] J. Branke, T. Kaußler, C. Smidt, and H. Schmeck, "A multi-population approach to dynamic optimization problems," in *Evolutionary Design and Manufacture*, pp. 299–307, Springer, 2000.

[32] A. Lipowski and D. Lipowska, "Roulette-wheel selection via stochastic acceptance," *Physica A: Statistical Mechanics and Its Applications*, vol. 391, no. 6, pp. 2193–2196, 2012.

[33] P. Asghari, A. M. Rahmani, and H. H. S. Javadi, "Privacy-aware cloud service composition based on qos optimization in internet of things," *Journal of Ambient Intelligence and Humanized Computing*, 2020.

[34] M. Hosseini Shirvani, "Bi-objective web service composition problem in multi-cloud environment: a bi-objective time-varying particle swarm optimisation algorithm," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 33, no. 2, pp. 179–202, 2021.

[35] S. Deng, L. Huang, H. Wu, and Z. Wu, "Constraints-driven service composition in mobile cloud computing," in *Proceedings of the 2016 IEEE International Conference on Web Services (ICWS)*, pp. 228–235, IEEE, San Francisco, CA, USA, July 2016.

[36] C. Lv, W. Jiang, S. Hu, J. Wang, G. Lu, and Z. Liu, "Efficient dynamic evolution of service composition," *IEEE Transactions on Services Computing*, vol. 11, no. 4, pp. 630–643, 2018.

[37] M. Yang and X. Hu, "Svm-based efficient qos-aware runtime adaptation for service oriented systems," in *Proceedings of the 2016 IEEE International Conference on Web Services (ICWS)*, pp. 396–403, IEEE, San Francisco, CA, USA, July 2016.

[38] L. Barakat, S. Miles, and M. Luck, "Adaptive composition in dynamic service environments," *Future Generation Computer Systems*, vol. 80, pp. 215–228, 2018.