**Ph.D. Research Proposal**

**Doctoral Program in “Department Name”**

# Multi-Query Scheduling and Data Retrieval using Customized Frameworks for Secure Grid Servers Environment

**by**

<Name of the Candidate>

<Reg. No of the Candidate>

**<Supervisor Name>**

**<Date of Submission (DD MM 20YY>**

## I. INTRODUCTION / BACKGROUND

Grid computing which is the distributed computing paradigm is an important evolving architecture. In grid computing, security is the major challenging issue to be concentrated [1]. In general, unauthorized user access increases the resource consumption, which is not efficient in grid environment. The grid computing environment is applicable is many areas. Some of them are,

- Healthcare Data Storage
- Big data Analytics
- Smart City Data Storage
- Pollution Data Storage and Retrieval

Three primary research issues in those applications are follows. (1). Authentication schemes often use ineffectual authentication credentials, (2). Fast and efficient scheduling is necessary since it has to handle large-scale users, (3). Lack of semantic representation in keyword search decreases the search accuracy

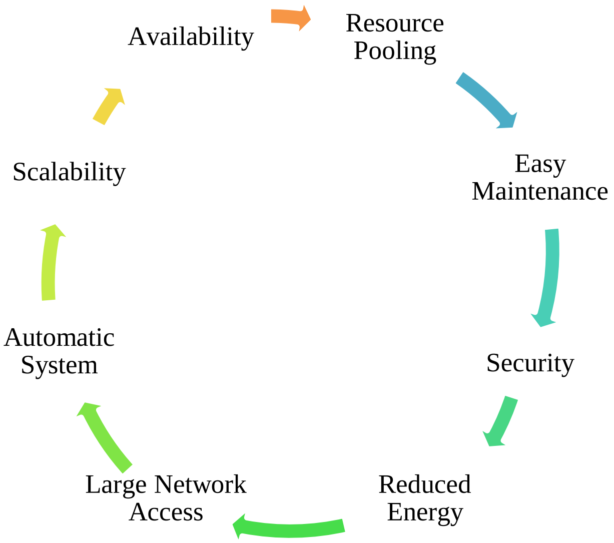


**Figure 1** Characteristic Features of Grid

In order to improve security level, multi-factor authentication scheme is proposed [2]. In multi-factor authentication, ID, password and smartcard are considered. However, it is vulnerable to smartcard loss attack. In addition, a two-factor authentication scheme uses password and smartcard [3]. To ensure security, XOR operations and one-way hash function are involved. The smartcard plays a pivotal role in anonymous authentication scheme also [4]. The smartcard based authentication is effective to server forgery attack. However, smartcard is likely to be tampered which affects the security level in the grid computing. Trust based scheduling is presented with the optimization technique in grid computing [5]. To enable scheduling process, Ant Colony Optimization (ACO) algorithm is presented with Firefly Algorithm (FFA). In this work, the Min-Min algorithm is used for pheromone initialization. In general, ACO algorithm has slow convergence rate which results in non-optimal scheduling.

The ACO algorithm is also used with cuckoo search algorithm to enable optimal scheduling [6]. The cuckoo search algorithm is used to form clusters of resources based on the load value and ACO algorithm performs scheduling process. However, execution of cuckoo search and ACO increases time consumption since both algorithms have higher time consumption. Besides to other techniques, MapReduce which is the big data processing procedure is utilized in grid computing [7]. Integration of MapReduce and grid computing results in minimized time consumption. A fuzzy based security scheme is proposed for grid environment [8]. The fuzzy algorithm computes the trust value of resources based on trust value and security level required. However, this work is unable to validate the users since it only validates the resources. To enable search over the distributed environment (i.e.) grid environment, a Boolean search is enabled [9]. Here, the users are authenticated and authorized by the data owners. Then, the user queries are searched by Boolean search mechanism. However, the Boolean search is ineffective and not suitable for secure search. In addition, authentication by data owner limits the security level. Thus, secure and semantic search over grid environment is still challenging issue.

## 1.1 Research Outline & Scope

This research work mainly focuses on authorized semantic search over grid computing environment. For that, this work covers authentication, query scheduling and retrieval processes.

## 1.2 Research Objectives

The major objective is to minimize the search time and to improve the search accuracy for authorized user queries. The following sub-objectives are follows.

- To prevent resources from unauthorized users by eliminating unauthorized users in the grid environment
- To support multi-query scheduling such that retrieval and response time is minimized in the grid environment
- To enable secure search and semantic search over multi-server grid environment

## II. RESEARCH GAPS

### 2.1 Common Problem Statement

In grid server-based big data processing, query search and data retrieval has been applied in many research works. However, searching unauthorized queries over distributed big data increases higher time consumption and also degrades the overall system performance. Besides, scheduling authorized user queries lacks with poor algorithm design and higher time consumption.

### 2.2 Problem Definition

In [1] paper proposes an authentication scheme to authenticate the grid users. The overall grid system is constructed with Grid Users and Resource Broker. To improve security level, Zero-Knowledge Proof (ZKP) based authentication and Intrusion Detection System (IDS) are proposed. Here, the IDS are deployed in the mobile agents to detect the malicious activities. The ZKP algorithm intends to defend against man in the middle attack. Thus, the authentication credentials such as ID and Password are encrypted by RSA algorithm. Besides, enhanced diffie Hellman algorithm is used for key exchange process.

**Problems**

- ZKP based authentication algorithm consumes large time and also involves high complexity.
- The ZKP algorithm only considers ID and PW that are insufficient to validate the grid users. Also, RSA algorithm takes large time for computation even for lower security level. Thus, the authentication procedure is inefficient.
- Involvement of random scheduling is ineffective since it makes large time for user requests. It also leads to schedule the user tasks to untrusted resources.

**Proposed Solutions**

- We present a novel Tri-Factor algorithm that works upon dual biometrics such as finger vein and eye vein which make the system as secure. We present Spongent Function algorithm and in authentication to ensure high-level security with minimum computational time.
- Proposed Spongent algorithm takes lower time consumption since we use lightweight encryption algorithm.
- A novel Refining MapReduce with RDA algorithm which is fast and efficient is proposed.

In [2] paper aims to schedule the grid user tasks by using Map-Reduce procedure. The map-reduce model is optimized with the non-dominated sorting genetic algorithm (NSGA-II) algorithm. This work improves the quality of service (QoS) by optimum scheduling. The objective is formulated as minimization of flow time and maximization of throughput. Initial population is initialized in the mapper phase and optimal scheduling is performed in the reducer phase.

**Problems**

- Consideration of limited metrics and lack of security further affects the overall performance of the grid system. Since, this work performs scheduling for all unauthorized user tasks.
- The NSGA-II algorithm has higher time consumption and the complexity is also high. Thus, user scheduling takes large time. Besides, the convergence is restricted by the involvement of sorting process.

**Proposed Solutions**

- We present RD algorithm to evaluate the fitness value and to schedule the queries to the resources in map-reduce model. It has good convergence and also has lower time consumption.
- We consider all major metrics in scheduling. Presented Spongent Function algorithm eliminates all unauthorized users at initial stage. Thus, resource consumption and time consumption is optimized.

This paper proposes [3] query analysis ontology-based cluster (QAOC) architecture for hadoop big data environment. The overall architecture includes query manager, scheduler and data management. The query manager is responsible to consolidate the user queries which are similar. For scheduling, neuro-fuzzy algorithm is proposed. This algorithm considers query arrival time, query length, and expiry time for scheduling. The ontology is constructed to form clusters of relevant data in data management process. When the user query is initiated then, the searching is carried over the ontology based binary index.

**Problems**

- Although ontology is constructed, here it is used to form clusters. Searching is performed over binary search time which increases retrieval time and has limitations in update and delete. Thus, searching process is inefficient.
- Scheduling by neuro-fuzzy algorithm only considers the query related metrics. But the resource related metrics are also important.

**Proposed Solutions:**

- We present novel ontology based search with Fuzzy with Artificial Fractal Tree structuring. It improves search efficiency and also minimizes searching time.
- Scheduling by Refining MapReduce with RDA considers all major metrics. Involvement of Map-Reduce also minimizes scheduling time.

## III. RESEARCH CONTRIBUTIONS

This research work focuses on secure and semantic search over grid environment. For that we present a Multi-Query Scheduling and Data Retrieval using Customized Frameworks in the customized Map-Reduce enabled Grid Environment. The overall environment includes

- Data Owners (Dos)
- Data Users (DUs)
- Blockchain Nodes

The overwork involves three major phases as, (i) Authentication Phase, (ii) Scheduling Phase, and (iii) Data Retrieval Phase.

### (i) Authentication Phase

In first phase, the DOs and DUs are authenticated by blockchain technology. For authentication, this research proposes Tri-Factor Authentication algorithm. Initially, all DUs and DOs are registered their ID, Password (PW) and Dual Biometrics such as Eye Vein and Finger Vein at blockchain. Based on Dual Biometrics, a bio-key is generated for all DOs and DUs. For bio-key generation, present Spongent Function algorithm. In authentication, the bio-key and random bits of biometrics are validated. In this phase, the unauthorized requests are eliminated.

### (ii) Scheduling Phase

The authorized DU requests are scheduled in order to access distributed servers. For query scheduling, we present novel Refined MapReduce procedure. In this algorithm, the user queries are mapped to optimal grid resources. To evaluate the resources for user queries, we enable Red Deer Algorithm phase. The RDA evaluates the resources based on multiple criteria such as Trust Level, Resource Score (Available Bandwidth and Queries), and Time Score (Response Time and Execution Time). GIS provides this information for scheduling.

### (iii) Data Retrieval Phase

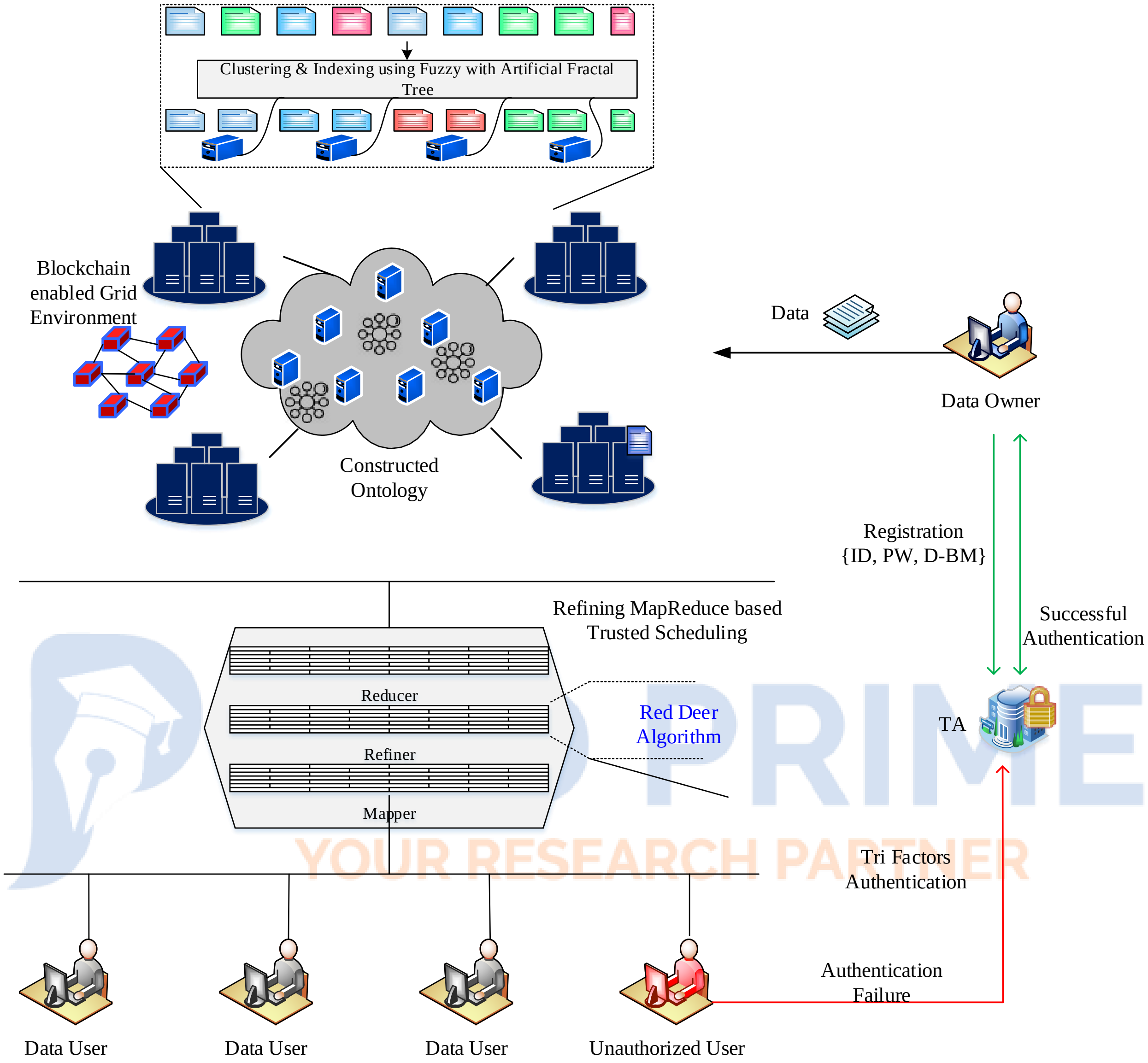
The data from authorized DOs is stored in the grids. Then for stored data, Fuzzy based Artificial Fractal Tree is constructed. This enables cluster based indexing. Besides, ontology is constructed to enable semantic search for user queries. Every transaction is added to the block nodes and it is connected in a chain basis. All grid resources search semantically for all user queries. At last, the grid resources return the optimum results for the user queries.

## Performance Evaluation

Finally, we will evaluate the proposed work in terms of the following performance metrics,

- Response Time
- Search Accuracy
- Retrieval Time
- Authentication Time
- Latency
- Cluster Purity (%)

SYSTEM ARCHITECTURE


Clustering & Indexing using Fuzzy with Artificial Fractal Tree
Blockchain enabled Grid Environment
Constructed Ontology
Data
Data Owner
Registration {ID, PW, D-BM}
Successful Authentication
Refining MapReduce based Trusted Scheduling
Reducer
Refiner
Mapper
Red Deer Algorithm
TA
Tri Factors Authentication
Authentication Failure
Data User
Data User
Data User
Unauthorized User

## IV. RESEARCH NOVELTIES

- Dual bio-metrics (Eye Vein and Finger Vein) based authentication ensure high level security
- Refining Map-Reduce based scheduling minimizes time consumption
- Involvement of RDA algorithm optimize the scheduling process
- Semantic ontology based searching improves accuracy

V. PREVIOUS WORKS & LIMITATIONS

**Reference 1**

**Title:** Adaptive security architectural model for protecting identity federation in service oriented computing

**Concept**

This paper presents a secure architectural model for distributed cloud and grid computing. In general, identity theft, identity management, authentication, data theft, and trust computation are major security concerns. As of now, security assessment markup language (SAML), open authentication and open identity are presented as security solutions. However, these security solution are insufficient for assuring security in distributed computing. Thus, this paper presents a secure architecture for distributed computing environment.

**Reference 2**

**Title:** Understanding security failures of multi-factor authentication schemes for multi-server environments

**Concept**

In this paper, a multi-factor authentication scheme is analysed and proposed. The main of this work is to analyse the multi-factor authentication in multi-server environment. In majority of multi-factor authentication schemes use smartcard, ID and passwords. The main pitfalls of the multi-factor authentication scheme are smart-card loss attack and there is no forward secrecy is achieved. This paper highlights that the smartcard and password alone is insufficient to authenticate the users in multi-server environment.

**Reference 3**

**Title:** Revisiting Anonymous Two-Factor Authentication Schemes for IoT-Enabled Devices in Cloud Computing Environments

**Concept**

This paper revisits the two-factor authentication scheme for cloud users in multi-server environments. The two-factor authentication scheme uses smartcard and password to validate the users. In that, multiple operations such as bitwise XOR operation, string concatenation and one-way hash function. Majorly the two-factor authentication scheme is constructed upon the strong assumption that smartcard cannot be tampered.

**Limitation**

- In practical, there is high possibility for smartcard tampering. Thus, the authentication scheme is inefficient.

**Reference 4**

**Title:** An improved anonymous authentication scheme for distributed mobile cloud computing services

**Concept**

This paper proposes an improved anonymous authentication scheme is proposed to authenticate users in distributed cloud computing. This paper builds upon smartcard based authentication procedure. Initially, a smartcard generator selects two cyclic groups over addition and multiplication. To embed the secret keys in smartcard, a one way hash function is used. The authors highlighted that the proposed scheme is efficient for server forgery attack.

**Limitation**

- However, the smartcard tampering issue is not resolved in this paper. If the smartcard lost or tampered, the user authentication fails.

**Reference 5**

**Title:** Trust based resource selection with optimization technique

**Concept**

This paper proposes a trust based resource selection by optimization technique in grid computing. In grid computing, scheduling is an important process. Here, the main aim is to minimize the makespan in grid environment. For trust based scheduling, ant colony optimization (ACO) algorithm is presented. In ACO, the pheromones update is performed by Min-Min algorithm. The results are then initialized as fireflies in firefly algorithm (FFO). Then, the FFA algorithm obtains the optimal scheduling. Trust value and makepan parameters are considered as objective functions.

**Limitation**

- In general, ACO algorithm has slow convergence rate which is ineffectual in scheduling.

**Reference 6**

**Title:** On scheduling transaction in grid computing using cuckoo search-ant colony optimization considering load

**Concept**

This paper presents load based scheduling in grid computing environment. The aim of this work is to balance the load among grid resources. At first, the cuckoo search algorithm forms clusters of resources. Then, optimal scheduling is performed by ant colony optimization (ACO) algorithm. The cuckoo search algorithm uses load as the metric to form clusters. The ACO algorithm considers makespan, miss ratio and throughput for performing optimal scheduling.

**Limitation**

- Both ACO and cuckoo search has slow convergence rate which increases time consumption for scheduling.

**Reference 7**

**Title:** An improved integrated Grid and MapReduce-Hadoop architecture for spatial data: Hilbert TGS R-Tree–based IGSIM

**Concept**

This paper presents an Integrated Grid and Spatially Indexed MapReduce (IGSIM) to enable fast search. For fast search, R-tree and R*-tree spatial indexes are constructed in the MapReduce environment. Based on them Hilbert TGS R-Tree index is constructed to enable parallel processing. This paper highlights that the involvement of MapReduce boosts up the parallel processing and minimize the time consumption.

**Limitation**

- Though time consumption is minimized, the search accuracy is low due to the absence of semantic ability to search.

**Reference 8**

**Title:** Fuzzy-Based Integration of Security and Trust in Distributed Computing

**Concept**

This paper aims to improve security of distributed computing environment such as grid computing. In precise, this work designs a trusted grid (T-grid) computational model for secure computations. For security purpose, fuzzy logic is proposed. The fuzzy algorithm finds the security level needed to be employed in the grid computing. The fuzzy algorithm considers final trust value and security level as input and provides secured final trust value as the final output. This final trust value is considered as reputation value for the future purpose.

**Limitation**

- However, trust based computation alone is not sufficient to assure desired security level in the grid environment since value evaluates the reputation of resources but the users are not validated in this work.

**Reference 9**

**Title:** Enabling Encrypted Boolean Queries in Geographically Distributed Databases

**Concept**

In this paper, secure multi-search methodology is presented. In this work, the data is stored in distributed manner across multiple servers. Here the data owner encrypts the data with searchable index construction. Then, the data and index is stored in the distributed data bases. When a user search is enabled, Boolean search is performed for data retrieval. Here the data users are authenticated based on non-interactive authorization scheme. In order to improve retrieval time, the user queries are searched in parallel.

**Limitations**

- Boolean search is performed only based on the query terms (i.e.) it doesn't consider semantic representations of the query which degrades the efficiency.
- The authorization is performed between owners and users which results in unauthorized access since the data owner is not authenticated.

**Reference 10**

**Title** – Performance Comparison of Heuristic Algorithms for Task Scheduling in IaaS Cloud Computing Environment

**Concept** –Various heuristic methods have been proposed for the optimal scheduling tasks in cloud computing environment. However, selecting the appropriate algorithm for solving the task assignment problem of a particular nature is complex because the methods are developed under different assumptions. The scheduling algorithms are First Come First Serve (FCFS), Minimum Completion Time (MCT), Minimum Execution Time (MET), Max-Min, Sufferage, and Min-Min are the Heuristic algorithms which are used for the task scheduling in cloud computing

**Reference 11**

**Title** – Normal Cloud Model based Algorithm for Multi-Attribute Trusted Cloud Service Selection

**Concept –**

This paper proposes cloud model for security sensitive users to select the trusted cloud services. Firstly, the trust evaluation mechanisms among different entities in human society are fitted and the multi-granularity selection standard of trust levels based on Gaussian cloud transformation is constructed. Then, the calculation model of user preferences based on the cloud analytic hierarchy process is developed. Finally, the trusted cloud service selection algorithm based on two-step fuzzy comprehensive evaluation is proposed and select the cloud services for users.

**Reference 12**

**Title -** Blockchain for Secure EHRs Sharing of Mobile Cloud based E-health Systems

**Concept -**

Blockchain is decentralized technology that mainly designed for employing security between cloud storage and data users or owners. This paper addressed the problem of how to reliably share HER (Electronic Health Records) among mobile users while guaranteeing high security levels in cloud environment. Performance of the proposed HER is investigated and analyzed in Amazon Web Services

**Limitations**

- Conventional blockchain technology consists of several drawbacks. So we modify the structure of blockchain and also use some other algorithms instead of using built-in algorithms.

**Reference 13**

**Title -** A Secure Technique for Unstructured Big Data using Clustering Method

**Concept –**

Today big data enabled cloud environment for data storage and retrieval is a growing research topic. In this paper secure technique is proposed for big data encryption, compression and clustering over cloud environment. The flow of this work is following: (1). Get the big text file as input, (2). Implement SDES algorithm on input file for encryption, (3). Apply Huffman compression technique of encrypted data, (4). Error control technique is applied on compressed file for error correction and (5). Apply clustering on error controlled data.

**Limitations**

- We use MapReduce paradigm for security purpose and also use some other lightweight cryptography algorithms for encryption and decryption operations.

**Reference 14**

**Title -** Enhancing Cloud-Based IoT Security through Trustworthy Cloud Service: An Integration of Security and Reputation Approach

**Concept –**

In this paper, authors have proposed trust assessment framework for improving security and reputation of cloud when deliver services to users. It enables to compute the trust for cloud services in order to ensure the security of cloud based IoT. Set of cloud specific metrics are considered to examine the cloud services security. In addition to it, feedbacks ratings are proposed to quality of cloud service are exploited in the reputation-based trust assessment method in order to evaluate the reputation of a cloud service.

**Reference 15**

**Title -** Task scheduling scheme based on sharing mechanism and swarm intelligence optimization algorithm in cloud computing

**Concept –**

In this paper, the author proposes a hybrid intelligent optimization algorithm of fusion sharing mechanism was proposed to realize dynamic scheduling of cloud tasks. First, the virtual machine scheduling is encoded as bees, ants and genetic individuals. Then, using artificial bee colony (ABC), ant colony optimization (ACO) and genetic algorithm (GA), the optimal solution is found in each neighborhood. Finally, by a mechanism of sharing, three algorithms regularly exchange their solutions and obtain the optimal solution as the current optimal solution for the next iteration process, in order to accelerate the algorithm convergence and enhance the accuracy of convergence.

**Limitations**

- Ant colony optimization suffers from trade-off between makespan and load
- Algorithmic complexity as it requires more than one iterations for task scheduling

**Reference 16**

**Title -** A multi-model estimation of distribution algorithm for energy efficient scheduling under cloud computing system

**Concept –**

In this paper, the author proposes a multi-model estimation of distribution (mEDA) algorithm to determine both task processing permutation and voltage supply levels (VSLs). How to manage the applications under computing systems such as a cloud computing system in a more efficient way is a focus problem. The primary performance goal is to reduce the execution time (makespan) of the application. As the need to cloud computing grows, the environmental influence of data centers attracts much attention. This paper aims at the scheduling of the precedence-constrained parallel application to minimize time and energy consumption efficiently.

**Limitations**

- Tasks are scheduled based on arrival time. So, when a large task arrives first it will be executed for a long time
- Waiting time of the tasks is increased

**Reference 17**

**Title -** Bi-objective decision support system for task-scheduling based on genetic algorithm in cloud computing

**Concept –**

In this paper, the author proposes a bi-objective decision support system for task scheduling based on genetic algorithm in cloud computing. This paper addresses the task-scheduling in cloud computing. This problem is known to be NP-hard due to its combinatorial aspect. The main role of the proposed model is to estimate the time needed to run a set of tasks in cloud and in turn reduces the processing cost. A genetic approach for modeling and optimizing a task-scheduling problem in cloud computing is proposed. The experimental results demonstrate that the proposed solution successfully competes with previous task-scheduling algorithms.

**Limitations**

- Though decision making is fast, it is not stable as only two parameters are considered for task scheduling
- Scheduling is not effective

BIBLIOGRAPHY

Mohamed, M.I., Hassan, M.F., Safdar, S., & Saleem, M.Q. (2019). Adaptive security architectural model for protecting identity federation in service oriented computing. Journal of King Saud University - Computer and Information Sciences.

Wang, D., Zhang, X., Zhang, Z., & Wang, P. (2020). Understanding security failures of multi-factor authentication schemes for multi-server environments. Comput. Secur., 88.

Wang, P., Li, B., Shi, H., Shen, Y., & Wang, D. (2019). Revisiting Anonymous Two-Factor Authentication Schemes for IoT-Enabled Devices in Cloud Computing Environments. Security and Communication Networks, 2019, 2516963:1-2516963:13.

Chaudhry, S.A., Kim, I.L., Rho, S., Farash, M.S., & Shon, T. (2019). An improved anonymous authentication scheme for distributed mobile cloud computing services. Cluster Computing, 22, 1595-1609.

Kumar, E.S., & Vengatesan, K. (2018). Trust based resource selection with optimization technique. Cluster Computing, 22, 207-213.

Mahato, D.P., Sandhu, J.K., Singh, N.P., & Kaushal, V. (2019). On scheduling transaction in grid computing using cuckoo search-ant colony optimization considering load. Cluster Computing, 1-22.

Singh, H., & Bawa, S. (2019). An improved integrated Grid and MapReduce-Hadoop architecture for spatial data: Hilbert TGS R-Tree-based IGSIM. Concurr. Comput. Pract. Exp., 31.

Kumar, P.S., & Ramachandram, S. (2019). Fuzzy-Based Integration of Security and Trust in Distributed Computing. Advances in Intelligent Systems and Computing.

Yuan, X., Yuan, X., Zhang, Y.H., Li, B., & Wang, C. (2020). Enabling Encrypted Boolean Queries in Geographically Distributed Databases. IEEE Transactions on Parallel and Distributed Systems, 31, 634-646.

Ennahbaoui, M., & Idrissi, H. (2018). Zero-Knowledge Authentication and Intrusion Detection System for Grid Computing Security.

Devarajan, R., Prakash, M., & Suresh, J. (2019). Computational grid scheduling architecture using MapReduce model-based non-dominated sorting genetic algorithm. Soft Computing, 1-13.

Pradeep, D., & Sundar, C. (2020). QAOC: Novel query analysis and ontology-based clustering for data management in Hadoop. Future Generation Computer Systems, 108, 849-860.

Syed Hamid Hussain Madni., Muhammad Shafie Abd Latiff1, Mohammed Abdullahi., Shafii Muhammad Abdulhamid., Mohammed Joda Usman., (2017). Performance Comparison of Heuristic Algorithms for Task Scheduling in IaaS Cloud Computing Environment, PLOS One, PP. 1-26

Yang, Y., Liu, R., Chen, Y., Li, T., & Tang, Y. (2018). Normal Cloud Model-Based Algorithm for Multi-Attribute Trusted Cloud Service Selection. IEEE Access, 6, 37644–37652.

Nguyen, D. C., Pathirana, P. N., Ding, M., & Seneviratne, A. (2019). Blockchain for Secure EHRs Sharing of Mobile Cloud based E-health Systems. IEEE Access, 1–1.

Nafis, M. T., & Biswas, R. (2019). A secure technique for unstructured big data using clustering method. International Journal of Information Technology.

Li, X., Wang, Q., Lan, X., Chen, X., Zhang, N., & Chen, D. (2019). Enhancing Cloud-Based IoT Security through Trustworthy Cloud Service: An Integration of Security and Reputation Approach. IEEE Access, 1–1.

FU Xiao. Task Scheduling Scheme Based on Sharing Mechanism and Swarm Intelligence Optimization Algorithm in Cloud Computing[J].Computer Science, 2018, 45(6A): 290-294.

Wu, C., & Wang, L. (2018). *A multi-model estimation of distribution algorithm for energy efficient scheduling under cloud computing system. Journal of Parallel and Distributed Computing, 117, 63–72.*