# *A Universal Fairness Evaluation Framework for Resource Allocation in Cloud Computing*

***Prof. Pragati Patil , Chandrashekhar G. Bhagat***

*[1]Project Guide, Department of Computer Engineering, Abha Gaikwad Patil College of Engineering, Nagpur.*
*[2]PG Scholar, Department of Computer Engineering, Abha Gaikwad Patil College of Engineering, Nagpur.*

***Abstract-***

Machine learning is nowadays ubiquitous, providing mechanisms for supporting decision making that leverages big data analytics. However, this recent rise in importance of machine learning also raises societal concerns about the dependability and trustworthiness of systems which depend on such automated predictions. In cloud computing, fairness is one of the most significant indicators to evaluate resource allocation algorithms, which reveals whether each user is allocated as much as that of all other users having the same bottleneck. However, how fair an allocation algorithm is remains an urgent issue. In this paper, we propose Dynamic Evaluation Framework for Fairness (DEFF), a framework to evaluate the fairness of an resource allocation algorithm. In our framework, two sub-models, *Dynamic Demand Model* (DDM) and *Dynamic Node Model* (DNM), are proposed to describe the dynamic characteristics of resource demand and the computing node number under cloud computing environment. Combining Fairness on Dominant Shares and the two sub-models above, we finally obtain DEFF. In our experiment, we adopt several typical resource allocation algorithms to prove the effectiveness on fairness evaluation by using the DEFF framework..

***Keywords: Resource allocation; fairness evaluation; cloud computing etc.***

## 1. Introduction

In cloud computing, computational resources are highly integrated in the “cloud”. Services and applications are provided by virtual machines running over the cloud platform. Hence, computational resources, such as CPU, RAM, bandwidth etc., should be properly scheduled for better service provision. Resource allocation algorithm is widely studied in recent works on shared communication and computing systems. *max-min fairness*[4][6] ensures the allocations of the users with minimal resource demands. In *proportional fairness*[10][14], it attempts to find a balance point in resource allocation among the competing interests. A *fairness* attempts to determine an equilibrium point between allocation fairness and the utilization efficiency of resources. Ref.[17] presents a game theory based approach which introduces a tradeoff between relay fairness and system throughput.

In multi-type resource allocation, ref.[1][3] and ref.[5][11][13] focus on multiple instances of the same resource. Ref.[7] proposes *Dominant Resource Fairness* (DRF) which is designed to ensure the fairness in the allocation of multiple types of resources, such as CPU, RAM and bandwidth etc. [2][8] propose genetic algorithm based approaches to obtain the optimal allocation.

Machine Learning (ML) is nowadays ubiquitous, as most organizations take advantage of it to perform or support decisions within their systems [1], [2]. ML is an area of Artificial Intelligence (AI) in which we use a set of statistical methods and computational algorithms to allow computers to learn from data [3]. ML algorithms can be divided into two main groups: supervised and unsupervised. Supervised learning involves the development of computational models for estimating an output based on previously known inputs and outputs. In unsupervised learning, the models are built based solely on existing inputs but there are no associated outputs that may be used for sake of training.

We may face fairness and transparency issues for both groups of algorithms. It is now commonplace to run ML systems in cloud-based infrastructures, motivated by issues such as elasticity, robustness, and ease of operation [4]. In practice, cloud services are fueling big data analytics, allowing organizations to make better and faster decisions using data that previously were hard or impossible to use [5]. This raises many opportunities in today’s competitive environment, by offering many services using highly scalable technologies on a pay-as-you-go basis. However, it also creates new challenges regarding trust, a paramount concern in critical systems [5]. Regulatory institutions have long focused these properties namely in OECD’s fair information practices [6] and in EU Privacy Directive 95/46/EC [7]. However, such legislation has never received as much emphasis as now. The new EU General Data Protection Regulation (GDPR) [8] shifts the onus to the organizations, demanding them to demonstrate that they are taking the appropriate measures to protect the legal rights of the individuals and their data, requiring privacy-preserving, fair and transparent systems.

## 2. Literature Survey

In this section, we review the related work on fairness and green energy usage.

### *A. Fairness*

Some fair schedulers [9], [10], [11] are proposed to resolve the resource fair allocation problem in the multi-tenant environment. These studies only consider single resource type in the cluster. In order to support fair allocation of multiple resources, new fairness definitions emerge. Dominant Resource Fairness applies the max-min fairness to multiple resource types in Hadoop YARN. Wang et al. [12] extend Dominant Resource Fairness

DOI: 10.55041/IJSREM24024

especially for the heterogeneous environment. These studies resolve the resource fair allocation in different scenarios without considering the tradeoff between performance and fairness. Recently, some studies begin to consider this tradeoff. They theoretically analyze the fairness-efficiency tradeoff for different fairness definitions [13], [14]. Tetris [15] explores the performancefairness tradeoff of Hadoop YARN from system view. Although these studies have observed the tradeoff between performance and fairness, the factors that impact this tradeoff is not explored in detail.

***B. Renewable energy-aware computing***

There have been some research on green-aware scheduling systems. Some studies maximize the usage of renewable energy by delaying the execution of batch jobs [16], [17]. Goiri et al. have conducted a series of studies and develop a green data center prototype to manage both deferrable and non-deferrable workloads at the presence of renewable energy [18]. Chen et al. [19] propose ReinDB that integrates renewable energy supply into database systems. EU has

founded a project called DC4Cities which proposes a technical and business related solution to optimize the usage of renewable energy in smart cities. Some attention has been paid to leverage battery to utilize the renewable energy efficiently. Few of the previous studies have paid attention to the energy efficiency of the workload, particularly in the MapReduce/Hadoop cluster.

- *Shuo Wang et. al. 2018,* In this paper, they design NXT-Freedom, a bandwidth guarantees enforcement framework that divides network capacity based on per-VDC fairness while achieving work-conservation. To ensure per-VDC fair allocation, a hierarchical max-min fairness algorithm is proposed. To be applicable to non-congestion-free network core and to be scalable, NXT-Freedom decouples computing per- VDC allocation from enforcing the allocation. Through evaluation of a prototype, we show that NXT-Freedom achieves per-VDC performance isolation, and can be rapidly adapted to flow variation in cloud datacenter. In cloud datacenter, it should be rational to enforce fair allocation on network resources among VDCs (virtual datacenters) in terms of multi-tenant model. Traditionally, cloud networks are shared in a best-effort manner, making it hard to reason about how network resources are allocated. Prior works concentrate on either providing minimum bandwidth guarantee or achieving work-conserving based on the VM-to-VM flow policy or per-source policy, or both. However, fair allocation on redundant bandwidth among VDCs is ignored.

- *Fengpan Zhao et. al. 2020,* Recently, privacy has become a major social issue since personal data is collected and analyzed from different IoT devices. To prevent the disclosure of private information, and violation of data protection rules, Federated Learning caters to this need. It provides a way to train a global model without exposing raw personal data. One of the most popular tools in this paradigm is Federated Averaging, where a few selected devices are forwarded to a global model, and the gradients thus obtained are averaged at the server. However, this aggregated global model faces the problem of dissimilar performance over clients due to unbalanced data and non-Independent and identically distributed data. In this paper, we proposed a novel framework called clusterbased Federated Averaging to achieve a fair global model by organizing the devices into groups and selecting clients from each group equally. In this way, the accuracy of the minority group could be improved significantly at the low expense of the majority group. To follow the federated learning's instinct of privacy protection, we adapt the training weights as the features to divide the users ensure the clients' training data does not leave their devices. We applied our framework on three popular datasets in machine learning: MNIST, Fashion MNIST, and Cifar-10. The experiments demonstrated that our framework could train a fair shared model effectively and efficiently.

- *Zhaojie Niu and Bingsheng He 2021,* In this paper they conduct detailed studies on the factors which impact the tradeoff between different factors. Based on the observations in our study, we propose workloadaware, energy-efficient and green-aware optimizations and implement them into Hadoop YARN. Particularly, in this thesis proposal, we propose to explore the following research problems. First, we explore the tradeoff between fairness and performance, and improve the performance of the state-ofthe- art approach by up to 225%. Second, we consider the energy efficiency, renewable energy supply as well as battery usage and reduce the brown energy consumption of existing systems by more than 25%. Third, we will explore the relationship between fairness and energy consumption, and eventually we will develop multi-objective optimizations for performance, fairness and energy consumption.

- *Nuno Antunes et. al. 2020,* Machine learning is nowadays ubiquitous, providing mechanisms for supporting decision making that leverages big data analytics. However, this recent rise in importance of machine learning also raises societal concerns about the dependability and trustworthiness of systems which depend on such automated predictions. Within this context, the new general data protection regulation (GDPR) demands that organizations take the appropriate measures to protect individuals' data, and use it in a privacy-preserving, fair and transparent fashion. In this paper we present how fairness and transparency are supported in the ATMOSPHERE ecosystem for trustworthy clouds. For this, we present the scope of fairness and transparency concerns in the project and then discuss the techniques that are being developed to address each of these concerns. Furthermore, we discuss how fairness and transparency are used with other quality attributes to characterize the trustworthiness of cloud systems.

- *Carlee Joe-Wong and Soumya Sen , 2018* As more businesses use the cloud for their computing needs, datacenter operators are increasingly pressed to perform effective and fair allocation in this multi-resource, multi-tenant setting. The presence of multiple resources allows an operator to offer different types of pricing strategies (e.g., bundled vs. unbundled) that can have different effects on its revenue. Pricing also affects the demand and resource allocation decisions across clients who typically require different ratios of each resource (e.g., CPUs, memory, bandwidth) to process their jobs, which results in a complex trade-off between fairness and revenue maximization. We develop an analytical framework to investigate the fairness and revenue tradeoffs that arise in a datacenter's multi-resource setting and the impact of different pricing plans on the operator's objective. We derive analytical bounds on the operator's fairness-revenue tradeoff and compare tradeoff points for different pricing strategies on a data trace taken from a Google cluster.

## 3. Problem Definition

Abundant allocation algorithms are proposed for resource allocation in cloud computing, how to evaluate the fairness of an allocation approach is less studied. Fairness evaluation model for single- type resource allocation algorithm. DRF based unified framework, named *Fairness on Domi- nant Shares* (FDS), for fairness evaluation, in which the efficiency of resource utilization is also considered. In FDS, two key factors are introduced, β and λ. β indicates the fairness type and λ emphasizes the resource utilization (efficiency). However, in cloud computing, the resource demands of the computing nodes (virtual machines) can vary at different task phases.

We define a task phase as a period in which a node is executing one computing task. For example, when the platform is solving equations in different sizes concurrently with computing nodes, the node number can be different according to the size and complexity of the equations. Moreover, the nodes which finish tasks will be terminated, and occupied resources can be released, whereas new nodes will be created for new tasks, and new resource allocations begins. Hence, the resource demand and the node number can change in different period under cloud environment. Both of these dynamic features in cloud are not adequately considered in existing research works.

To address the two issues, we propose a Dynamic Evaluation Framework for Fairness (DEFF) in resource allocation. Our model contains two sub-models, Dynamic Demand Model (DDM) and Dynamic Node Model (DNM). The previous depicts resource demand of the nodes in each task phase, whereas the later gives a description to the variation of node number. With combination of DDM and DNM, we obtain our evaluation model DEFF, which can better adapt the cloud environment.

## 4. Objectives

- To explore the relationship between fairness and energy consumption
- To develop multi-objective optimizations for performance, fairness and energy consumption.
- To conduct detailed studies on the tradeoff and propose bi-criteria optimization algorithms to address the tradeoff between different factors,
- To adopt several typical resource allocation algorithms to prove the effectiveness on fairness evaluation by using the DEFF framework.

## 5. Existing system

- We may face fairness and transparency issues for both groups of algorithms. It is now commonplace to run ML systems in cloud-based infrastructures, motivated by issues such as elasticity, robustness, and ease of operation.
- In practice, cloud services are fueling big data analytics, allowing organizations to make better and faster decisions using data that previously were hard or impossible to use.
- This raises many opportunities in today's competitive environment, by offering many services using highly scalable technologies on a pay-as-you-go basis.
- However, it also creates new challenges regarding trust, a paramount concern in critical systems.

## 6. Proposed system

- From the context of the project we introduce an initial set of techniques that are being developed not just to support but also to monitor and assess fairness and transparency in the context of ML applications and systems. Finally, we present concrete examples of practical application of these techniques in Lemonade, and how they integrate with other components.

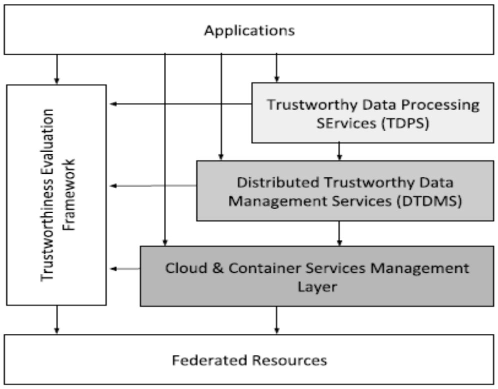
Applications
Trustworthiness Evaluation Framework
Trustworthy Data Processing SErvices (TDPS)
Distributed Trustworthy Data Management Services (DTDMS)
Cloud & Container Services Management Layer
Federated Resources

Fig. 1. The model of the ATMOSPHERE project.

- The application itself or the execution framework for adjusting parameters to increase trust or to react to runtime failures in federated infrastructures, up to the limits on resource allocation that a user may have set - avoiding infinite consumption of resources. Fairness and transparency are mainly monitored at the layer of the data processing service.

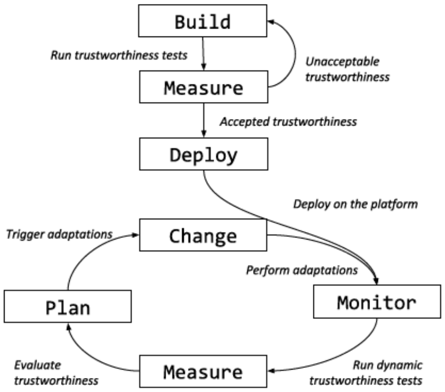

Fig. 2. The lifecycle of ATMOSPHERE applications.

## 7. Conclusion

In this work, a framework for evaluating two cloud pricing strategies–bundled and resource pricing–in terms of their resulting fairness and revenue. We first characterize client demand for resources as a function of the prices offered under these different pricing plans. After showing some analytical bounds on the tradeoff between fairness and revenue, we compare achieved fairness and revenue under the two pricing plans. We finally use data taken from a Google cluster to numerically evaluate the impact of resource capacity and volume discounts on the operator's fairness-revenue tradeoff.

## 8. Future Work

Future extensions of this work will consider an additional pricing scheme: differentiated pricing, in which the operator can choose a per-job price for each client independent of the client's resource requirements. We do not consider such a pricing plan here since bundled and resource pricing are more practically relevant; in practice clients are generally not charged different per-job prices. One could also extend our work to take into account job completion deadlines, which impose an additional constraint on the resources allocated at any given time. We also plan to consider tradeoffs between revenue, fairness, and operational efficiency, e.g., through examining the total amount of leftover resources.

## References

[1] BARUAH S K, GEHRKE J, PLAXTON C G. Fast scheduling of periodic tasks on multiple resources[ C]. IPPS. IEEE Computer Society, 1995,280-288.

[2] CAMPEGIANI P. A Genetic Algorithm to Solve the Virtual Machines Resources Allocation Problem in Multi-tier Distributed Systems[C]. VPACT'09, 2009.

[3] BARUAH S K, COHEN N K, PLAXTON C G, and Donald A. Varvel. Proportionate Progress: A Notion of Fairness in Resource Allocation[J]. Algorithmica, 1996, 15(6): 600–625.

[4] BERTSEKAS D, GALLAGER R. Data Networks[M]. Prentice Hall, 1992.

[5] BLANQUER J M, ÖZDEN B. Fair Queuing for Aggregated Multiple Links[C]. SIGCOMM, 2001: 189–197.

[6] CHARNY A, CLARK D, JAIN R. Congestion Control with Explicit Rate Indication[C]. International Conference on Communications, 1995(3): 1954-1963.

[7] GHODSI A, ZAHARIA M, HINDMAN B, *et al*. Dominant Resource Fairness: Fair Allocation of Multiple Resource Types[J]. In Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation, 2011, 24–24.

[8] GU J H, HU J H, ZHAO T H, *et al.* A New Resource Scheduling Strategy Based on Genetic Algorithm in Cloud Computing Environment[J]. Journal of Computers, 2012(7): 42–52.

[9] WONG C J, SEN S, LAN T, *et al*. Multi-resource Allocation: Fairness-Efficiency Tradeoffs in A Unifying Framework[C]. In INFOCOM, IEEE, 2012: 1206–1214.

[10] KELLY F P. Charging and Rate Control for Elastic Traffic[J]. European Transaction on Telecommunications, 1997(8): 33-37.

[11] KLEINBERG J M, RABANI Y, and TARDOS É. Fairness in Routing and Load Balancing[J]. Journal of Computer System Sciences, 2001, 63(1):2–20.

[12] LAN T, KAO D, CHIANG M, *et al*. An Axiomatic Theory of Fairness in Network Resource Allocation[ C]. In INFOCOM, IEEE, 2010, 1343–1351.

[13] LIU Y and KNIGHTLY E W. Opportunistic Fair Scheduling over Multiple Wireless Channels[C]. INFOCOM, 2003.

[14] MASSOULIÉ L and ROBERTS J. Bandwidth Sharing: Objectives and Algorithms[C]. INFOCOM, 1999, 1395–1403.

[15] TAN L, PUGH A C and YIN M. Rate-based Congestion Control in ATM Switching Networks Using A Recursive Digital Filter[J]. Control Engineering Practice, 2003, 11(10): 1171–1181.

[16] TIAN J F, YUAN P, and LU Y Z. Security for Resource Allocation Based on Trust and Reputation in Computational Economy Model for Grid[C]. Frontier of Computer Science and Technology 2009, IEEE, 2009: 339–345.

[17] TENG Y L, HUANG T, LIU Y Y, *et al*. Cooperative Game Approach for Scheduling in Two-Virtual- Antenna Cellular

Networks with Relay Stations Fairness Consideration[J]. China Communications, 2013, 10(2):56–70.
[18] ZHU D, MOSSÉ D, and MELHEM R G. Multiple-Resource Periodic Scheduling Problem: How Much Fairness is Necessary?[C] RTSS, IEEE Computer Society, 2003: 142–151.
[19] ZUKERMAN M, TAN L S, WANG H W, *et al*. Efficiency-Fairness Tradeoff in Telecommunications Networks[C]. IEEE Communications Letters, 2005: 643–645.