J. Parallel Distrib. Comput. 107 (2017) 76–86

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

journal homepage: www.elsevier.com/locate/jpdc

Non-cooperative power and latency aware load balancing in distributed data centers

Rakesh Tripathi [a], S. Vignesh [a], Venkatesh Tamarapalli [a,*], Anthony T. Chronopoulos [b], Hajar Siar [c]

[a] Department of CSE, IIT Guwahati, Assam, India
[b] Department of Computer Science, University of Texas at San Antonio, USA
[c] Faculty of Electrical and Computer Engineering, Semnan University, Semnan, Iran

HIGHLIGHTS

- Proposed game theory-based novel load balancing scheme for distributed datacenters.
- Scheme provides latency fairness across the users.
- Algorithm minimizes the operating cost for data center operators.
- Algorithm achieves optimal result with very low complexity.

ARTICLE INFO

Article history:
Received 25 September 2016
Received in revised form
8 March 2017
Accepted 17 April 2017
Available online 27 April 2017

Keywords:
Distributed data centers
Game theory
Front-end proxy servers
Optimization of combined energy and latency cost

ABSTRACT

In this paper we propose an algorithm for load balancing in distributed data centers based on game theory. We model the load balancing problem as a non-cooperative game among the front-end proxy servers. We model the operating cost associated with a data center as a weighted linear combination of the energy cost and the latency cost. We propose a non-cooperative load balancing game with the objective of minimizing the operating cost and obtain the structure of Nash equilibrium. Based on this structure, a distributed load balancing algorithm is designed. We compare the performance of the proposed algorithm with the existing approaches. Numerical results demonstrate that the solution achieved by the proposed algorithm approximates the global optimal solution in terms of the cost and it also ensures fairness among the users.

1. Introduction

In the recent past, there has been a tremendous growth in the demand for Internet-scale services like Web search, video streaming, and online gaming. Almost all of them are implemented on geographically distributed data centers. A geo-distributed data center is a collection of small geographically distributed data centers which provides high reliability and performance. In an operational data center there are several front-end proxy servers that map the client requests to the appropriate data centers. These proxy servers use different objectives like maximizing the utilization, minimizing the latency or the operating cost to map the requests. Since the client requests are handled in a distributed manner by the front-end nodes, there is a need for an efficient, distributed algorithm to determine the mapping strategy, also termed load balancing strategy.

In general, a load balancing strategy can be classified as static, semi-static or dynamic. In the static approach [12], all the information necessary for the decision making is available before the execution of the algorithm and it remains constant during the execution. In the semi-static approach [10], the required information is available at the beginning of each time step or a well defined point. For example, the load on the system and the cost of serving the same are available with reasonable accuracy just before the time slot. In the dynamic approach [16] the information is not known till the point of execution and it might change during the course of execution.

* Corresponding author.
E-mail addresses: t.rakesh@iitg.ernet.in (R. Tripathi), s.vignesh@iitg.ernet.in (S. Vignesh), t.venkat@iitg.ernet.in (V. Tamarapalli), Anthony.Chronopoulos@utsa.edu (A.T. Chronopoulos), h_siar@semnan.ac.ir (H. Siar).

http://dx.doi.org/10.1016/j.jpdc.2017.04.006
0743-7315/

Load balancing algorithms can also be classified as centralized or decentralized. In a centralized approach, one node in the system collects the information necessary to decide the strategy for load balancing. In the decentralized approach, multiple nodes participate to decide the load balancing strategy, either cooperatively or independently. Decentralized approach (cooperative or otherwise) is resilient and scalable compared to the centralized one, particularly for large-scale distributed data centers. In the cooperative approach, all the nodes form a coalition to decide the optimal solution, which improves the performance [27]. In the non-cooperative approach, each node maximizes/minimizes the utility independently, but eventually reach an equilibrium [21]. Non-cooperative game solutions to the load balancing problem are computationally efficient and can be implemented in a distributed manner across all the participating players (in our case, the front-end proxies) [35,7].

The operational cost of a data center is influenced by factors such as electricity prices, server/data center failure, green energy availability, and client demand. Therefore, load balancing is challenging as the designed strategy must consider the spatio-temporal variation in these factors to minimize the operating cost. Most of the literature on load balancing in distributed data centers [25,26] considered minimizing the operating cost, which is profitable for the operators, but has ignored the users' perspective. Users consuming the same resources may pay the same price, but experience variable delays. For business continuity, ensuring fairness in service latency across the requests from different clients is also important. The load balancing algorithms in the literature designed to provide fairness, did not consider the energy cost. Therefore, we consider the linear combination of operating cost (or energy cost) and revenue loss due to latency (including the network and queueing delays) as the objective function.

In a distributed data center, the user requests are served by the front-end proxy servers independent of each other. Each proxy server prefers to get its requests served by the data center first to minimize the service delay. In order to model this selfish nature in distributed load balancing, we use the non-cooperative game theory approach. We propose a game-theoretic distributed load balancing algorithm, that is executed across a finite number of front-end proxy servers. The objective of the game is to minimize the sum of the energy cost and the revenue loss due to delayed service. The proposed approach reduces the cost compared to an approach that only minimizes the operating cost as in [40].

In summary, the main contributions of this work are as follows.

- For the first time, we model the load balancing in distributed data centers as a non-cooperative game among the front-end proxies. We consider the spatio-temporal variation in the electricity price, the offered load, and the availability in the model. We prove that the Nash equilibrium is the solution of this game, which is guaranteed to exist since the proposed objective function is continuous, convex and increasing [21,31]. We characterize the Nash equilibrium and propose a distributed algorithm for computing the same.
- We evaluate the performance of the non-cooperative game theoretic algorithm (abbreviated as NCG) along with the existing ones, such as the proportional scheme and the global optimal scheme, using real-world data. The proposed NCG algorithm shows better fairness (in service latency) at a comparable cost.

The rest of this paper is organized as follows. Section 2 presents the literature on game theoretic approaches related to our work. In Section 3 we present the architecture and the cost model used in the formulation. We present the non-cooperative game model and derive the structure of Nash equilibrium in Section 4. A distributed algorithm to solve the game and its analysis is given in Section 5. In Section 6, we present numerical results that demonstrate the performance of our algorithm against the optimal approach. Section 7 concludes the paper and proofs of various theorems are presented in the Appendix.

2. Related work

In this section we review the literature on cost-aware load balancing and other game theory-based approaches in distributed systems, which are relevant to our work.

(a) Electricity price-aware load balancing:

The problem of load balancing requests across different data centers leveraging electricity price variability has been addressed in [26,25], where the requests are routed to data centers operating with cheaper electricity. The authors of [40], considered availability of green energy sources, along with the electricity prices while choosing the data center. All these works used an optimization framework that is solved centrally. The work in [38] proposed a decentralized algorithm for load balancing considering the server availability, but did not optimize the latency or the operating cost at the data center. The work in [15] proposed three distributed algorithms for load balancing based on Gauss–Seidel, gradient projection and gradient descent methods. These methods are all known to be computationally expensive and may not be effective for dynamic load balancing.

(b) Cooperative game theoretic approaches:

Numerous efforts have been made for load balancing and resource management using cooperative game theory for example in communication network [28], distributed systems (DS) [8,23], and grid computing [34]. In [28], the authors presented an excellent summary on coalition games and their use in wireless communication networks. They listed the applications of game theory for distributed resource allocation, congestion control, power control, and spectrum sharing in cognitive radio networks. The work in [8] modeled the static load balancing in a single class DS with heterogeneous computers, as a cooperative game among the nodes (being assigned user jobs). It is shown that the Nash bargaining solution (NBS) gives an optimal solution and it is also fair. A similar approach was used in [23], where the communication cost for job transfer between the nodes was also considered. The authors of [34] addressed the problem of job allocation in a grid environment. They presented a distributed algorithm based on the structure of the NBS to minimize the average job completion time. In [39], the authors proposed a data center selection framework to ensure latency fairness across clients using the NBS, and presented algorithms based on dual decomposition and sub gradient methods. All these works have studied the structure of NBS for load balancing in architectures not similar to the ones used in geo-distributed data centers and they also used different objectives. In this paper, we use non-cooperative game theory to account for the selfish nature in decentralized load balancing and due to its computational efficiency.

(c) Non-cooperative game theoretic approaches:

The authors of [14] proposed a price bidding strategy for multiple users competing for servers in a cloud environment, where non-cooperative game is used with the objective of maximizing the net profit while being time efficient. The authors of [17] proposed an auction-based online mechanism for VM provisioning in cloud environment. The main drawback of using these approaches is the slow response time, as the bidder has to wait for auction clearing. In this paper, we address the problem of online load balancing in Internet data centers, where the requests are small in size to be served with minimum latency.

The authors of [7] addressed the problem of load balancing in a DS consisting of n computers (or nodes) shared by m client regions (classes). They proposed a distributed algorithm for load balancing in distributed system using non-cooperative