Smart Task Distribution in Combined Fog-Cloud Scenarios

Dissertation
zur
Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)
der Fakultät für Informatik und Elektrotechnik
der Universität Rostock

vorgelegt von
Mohammadreza Pourkiani, geboren am 15.09.1991 in Mashhad, Iran

Rostock, 01.05.2022

Abstract

In order to collect data, most of the IoT-based applications utilize sensors, which are limited in terms of computational and storage capabilities. Therefore, the collected raw data by the IoT sensors must be transmitted to capable servers for processing, storage, and data mining purposes. Fog and Cloud computing are two leading technologies which can provide computation and storage services for IoT-based applications. Cloud provides powerful servers, which are located far from the users and have high latency, while Fog provides servers with limited computational power in the proximity of the users with low latency. As there are different delay-sensitive and delay-tolerable applications in the world of the IoT, utilization of only Fog or Cloud can not be a perfect approach for all of the scenarios.

Moreover, task distribution between the fog and cloud servers is challenging as in combined fog-cloud scenarios, there are various types of servers, which are heterogeneous in terms of hardware, delay, workload, and computational power. This heterogeneity makes the selection process of the most suitable server at each time slice very difficult. In this thesis, our main goal is to find a solution for resource allocation in combined fog-cloud scenarios with regard to the application requirements. For this purpose, we proposed an intelligent task assignment algorithm (MLTD) that runs in the task distributor unit. This algorithm takes the diversity of servers into account, considers the application requirements, and uses machine learning methods for estimating the response times of the fog and cloud servers, in addition to the size of the data that must be communicated over the Internet. By utilizing this method, after receiving the raw data from the sensors, the task distributor unit selects the most suitable server at that time-slice and assigns the received tasks to that server to be processed.

In order to investigate the performance of MLTD, we used that for distributing the tasks of a delay-tolerable application (that solves mathematical questions) and a delay-sensitive application (that provides online healthcare services and utilizes the wireless body sensor networks for data collection to monitor the health status of people who work in environments with high levels of heat stress, such as the steel and iron industries). For distributing the tasks of the discussed applications between the fog and cloud servers, we utilized the Artificial Neural Networks (as the function approximation method) in the task distributor unit. To train the neural networks, we generated different numbers of tasks and ran them on all of the fog and cloud servers. For the training process, we set the response times of servers as the output, and the parameters of the tasks as the input of the neural networks. In the next step, we set the trained neural networks in the broker and therefore made the broker able to select the fastest server for processing a received task from the IoT sensors. We also added more parameters in the training process of the neural networks in

different situations to make our proposed method scalable and usable in different network architectures.

The performance of MLTD has been investigated in different experiments. The achieved results show that this technique can reduce the Internet bandwidth utilization, response time, and resource utilization compared to other proposed methods in the state of the art. The reason is that our proposed technique (unlike the other discussed methods in the state of the art) can predict the response times of available servers, in addition to the future Internet bandwidth utilization at the time of task arrival (before the distribution process). Therefore, MLTD can distribute the tasks with regard to the requirements of applications in terms of response time or bandwidth utilization.

However, we observed that the performance of MLTD entirely depends on the precision of the utilized function approximation methods, which can be affected by using different types of tasks, training methods, and richness of training. Moreover, we also witnessed that our smart task distribution technique performs excellently when the fog and cloud servers provide response times with a difference of more than the error of the utilized function approximation method for predicting the response times.

Abstrakt

Zur Datenerfassung verwenden die meisten IoT-basierten Anwendungen Sensoren, deren Rechen- und Speicherkapazitäten begrenzt sind. Daher müssen die gesammelten Rohdaten von den IoT-Sensoren an fähige Server zur Verarbeitung, Speicherung und zum Data Mining übertragen werden. Fog und Cloud Computing sind zwei führende Technologien, die Berechnungs- und Speicherdienste für IoT-basierte Anwendungen bereitstellen können. Die Cloud bietet leistungsstarke Server, die weit von den Nutzern entfernt sind und eine hohe Latenz aufweisen, während Fog Server begrenzter Rechenleistung in der Nähe der Nutzer mit geringer Latenz bereitstellt. Da es in der Welt des IoT verschiedene verzögerungsempfindliche und verzögerungstolerante Anwendungen gibt, kann die Nutzung von Fog oder Cloud nicht für alle Szenarien ein perfekter Ansatz sein.

Darüber hinaus ist die Aufgabenverteilung zwischen den Fog- und Cloud-Servern eine Herausforderung, da es in kombinierten Fog-Cloud-Szenarien verschiedene Arten von Servern gibt, die in Bezug auf Hardware, Verzögerung, Arbeitslast und Rechenleistung heterogen sind. Diese Heterogenität macht die Auswahl des am besten geeigneten Servers in jeder Zeitscheibe sehr schwierig. Das Hauptziel in dieser Arbeit ist es, eine Lösung für die Ressourcenzuweisung in kombinierten Fog-Cloud-Szenarien unter Berücksichtigung der Anwendungsanforderungen zu finden. Zu diesem Zweck haben wir einen intelligenten Aufgabenzuweisungsalgorithmus (MLTD) vorgeschlagen, der in der Aufgabenverteilungseinheit läuft. Dieser Algorithmus berücksichtigt die Vielfalt der Server, die Anforderungen der Anwendung und nutzt Methoden des maschinellen Lernens, um die Antwortzeiten der Fog- und Cloud-Server sowie die Größe der über das Internet zu übermittelnden Daten zu schätzen. Mit dieser Methode wählt die Aufgabenverteilereinheit nach dem Empfang der Rohdaten von den Sensoren den zu diesem Zeitpunkt am besten geeigneten Server aus und weist die empfangenen Aufgaben diesem Server zur Bearbeitung zu.

Um die Leistung von MLTD zu untersuchen, haben wir es für die Verteilung der Aufgaben einer verzögerungstoleranten Anwendung (die mathematische Fragen löst) und einer verzögerungsempfindlichen Anwendung (die Online-Gesundheitsdienste anbietet und die Wireless-Body-Sensornetzwerke für die Datenerfassung nutzt, um den Gesundheitszustand von Menschen zu überwachen, die in Umgebungen mit hohem Hitzestress arbeiten, z. B. in der Stahl- und Eisenindustrie) eingesetzt. Um die Aufgaben der besprochenen Anwendungen zwischen den Fog und Cloud-Servern zu verteilen, haben wir Artificial Neural Networks (als Methode zur Funktionsannäherung) in der Aufgabenverteilungseinheit eingesetzt. Um die Neural Networks zu trainieren, haben wir eine unterschiedliche Anzahl von Aufgaben generiert und sie auf allen Fog- und Cloud-Servern ausgeführt. Für den Trainingsprozess haben wir die Antwortzeiten der Server als Ausgabe und die

Parameter der Aufgaben als Eingabe der neuronalen Netze festgelegt. Im nächsten Schritt setzten wir die trainierten neuronalen Netze in den Broker ein, so dass dieser in der Lage war, den schnellsten Server für die Verarbeitung einer empfangenen Aufgabe von den IoT-Sensoren auszuwählen. Wir fügten auch weitere Parameter in den Trainingsprozess der neuronalen Netze in verschiedenen Situationen ein, um unsere vorgeschlagene Methode skalierbar und in verschiedenen Netzwerkarchitekturen verwendbar zu machen.

Die Leistung von MLTD wurde in verschiedenen Experimenten untersucht. Die erzielten Ergebnisse zeigen, dass diese Technik die Internet-Bandbreitennutzung, die Antwortzeit und die Ressourcennutzung im Vergleich zu anderen vorgeschlagenen Methoden in früheren Arbeiten reduzieren kann. Der Grund dafür ist, dass die von uns vorgeschlagene Technik (im Gegensatz zu den anderen vorgeschlagenen Methoden) die Antwortzeiten der verfügbaren Server sowie die künftige Auslastung der Internetbandbreite zum Zeitpunkt der Ankunft der Aufgabe (vor dem Verteilungsprozess) vorhersagen kann. Daher kann MLTD die Aufgaben im Hinblick auf die Anforderungen der Anwendungen in Bezug auf die Antwortzeit oder die Bandbreitennutzung verteilen.

Wir haben jedoch festgestellt, dass die Leistung von MLTD vollständig von der Genauigkeit der verwendeten Funktionsapproximationsmethoden abhängt, die durch die Verwendung verschiedener Aufgabentypen, Trainingsmethoden und die Reichhaltigkeit des Trainings beeinflusst werden kann. Darüber hinaus konnten wir feststellen, dass unsere intelligente Aufgabenverteilungstechnik hervorragend funktioniert, wenn die Antwortzeiten von Fog- und Cloud-Servern um mehr als den Fehler der für die Vorhersage der Antwortzeiten verwendeten Funktionsapproximationsmethode abweichen.