Literature Survey

On

MEC Server placement and Task Scheduling

Abstract

A literature survey was conducted on the project topic MEC server placement and task scheduling. Initially, we cover all the prerequisites for the topic i.e. 5g, IoT, some basic Machine learning algorithms, cloud computing, and 3GPP briefly. A literature review of various research papers relates to the past research on the same and similar topics were conducted. The paper referred to were 1. IEEE document 9064145 Joint User Association and VNF Placement for Latency Sensitive Applications in 5G Networks
IEEE document 8939566 Low-Cost MEC Server Placement and Association in 5G Networks
IEEE document 8318685 Mobile Edge Computing Resources Optimization: A Geo-Clustering Approach
Researchgate document 334534213 Edge computing server placement with capacitated location-allocation. Various algorithms were mentioned LowMEP, MILP, PACK in the research papers that were comprised of search and graph algorithms like block-coordinate descent algorithm for placement and problems like capacitor clustering problems for reducing latencies.

Prerequisites

PART 1

5g and Iot

5g (5th generation) refers to the wireless tech generation which is characterized by very high data transmission rates. The frequency range lies between 3-300GHz.
5g and IoT technology is more than just a new generation of wireless technology. 5G network is mainly used to meet the needs of three major types of services in the future:

- Enhanced mobile broadband services for approximately 2.5 billion users worldwide, such as high-definition video, telemedicine, telecommuting.
- Massive Internet of Things which can reach challenging locations
- The need for critical communication scenarios that meet ultra-high reliability and ultra-low latency.

Internet of things is a growing network of internet-connected physical devices processing the ability to collect and share massive volumes of information/data.

Cloud Computing

- Cloud Computing refers to the on-demand delivery over the Internet with charges on a usage basis, Provide Online data storage, configuration, and accessing of online applications.
- Clouds can provide services via public/private networks where the service hosting system is at a remote location.
- Few Advantages of Cloud Computing: On-demand self-service, Cost-effective, Applications as utilities over the Internet,etc
- Cloud Computing -> Deployment Model and Service Model. Deployment Model includes:

1. Public Cloud 2.Private Cloud 3.Hybrid Cloud

- Service Model includes:
  IAAS - Infrastructure as a Service
  PAAS - Platform as a Service
  SAAS - Software as a Service
- A few cloud service companies are: Amazon’s AWS, Microsoft Azure, Google Cloud

MEC and 3GPP

- MEC is a Computing platform that could run applications inside wireless base stations to provide services to mobile users. The basic idea is to migrate the cloud computing platform from the inside of the mobile core network to the edge of the mobile access network to realize the flexible use of computing and storage resources.
- 5G-based MEC services can fully realize the advantages of low latency and high bandwidth in the Internet of Things, video, medical, retail, and other industries
- MEC also helps in reducing Network Congestion.

- The 3rd Generation Project Partnership is the organization that creates and maintains the technical standards for global mobile communication technologies, including GSM, GPRS, EDGE, UMTS, HSPA, LTE, and future 5G technologies.
- 3GPP is divided into three Technical Specification Groups (TSG), each one covering a broad aspect of mobile radio networks. The TSGs in turn consist of several Working Groups (WG). The working groups all have their own technical focus within the area of their TSG.

Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
Basic Machine Learning Algorithms:

- **Linear Regression**: Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, rather than trying to classify them into categories.

- **Logistic Regression:** Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

- **Decision Trees:** Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

continued...

- Naive Bayes: Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks.

- kNN (k- Nearest Neighbors): KNN is a model that classifies data points based on the points that are most similar to it. It uses test data to make an “educated guess” on what an unclassified point should be classified as.

There are also some search based and graph-based algorithms that can be used but the study is being done by the team to become more proficient in their logic and usage

Research Papers

PART 2

Low Cost MEC Server Placement and Association in 5G Networks

Aim: The placement of MEC server and association algorithm that promises an end-to-end latency with a minimum number of servers to be placed.

Algorithm: LowMEP

Working: An algorithm is implemented to determine the location of the MEC server as well as the serving MEC server for each radio access network (RAN) that satisfies a certain delay budget i.e. latency(2ms - 50ms) with the minimum number of MEC servers to be placed. To that end, we formulate the problem as a capacitated clustering problem, which finds the minimum number of clusters to cover all elements as well as the association between each cluster and the elements for given constraints.

Conclusion: In 5G networks, MEC servers need to be located as much close as possible to the UEs with distributed UPFs(user plane function) for delay-sensitive services such as URLLC. It is important to find cost-efficient MEC server placement within the conditions that carriers must set up to support URLLC. In this paper, we propose a practical algorithm, called LowMEP, which finds a minimum number of MEC servers considering both delay and workload budget in 5G networks. Finally, we performed experiments to evaluate LowMEP with a real dataset using different machine learning algorithms provided by Shanghai Telecom. The result shows that LowMEP performs MEC server placement with fewer MEC servers than other algorithms.

Mobile Edge Computing Resources Optimization: a Geo-clustering Approach

Aim: MEC server systems dealing with key issues which include dimensioning such systems in terms of server size, server number, and server operation area to meet MEC goals which can be tackled by a mixed-integer linear program and a graph-based algorithm taking into account a maximum MEC server Capacity.

Working: 1) In a MEC deployment MEC servers are positioned in the infrastructure close to the edge of the network they are small-scale datacenters with low to moderate resources
collocated with the base stations, access points, and/or placed in the access/aggregation network.
2) MEC geo clustering and Mixed-integer linear program were first formally described but as the large-scale dimension such as large server size, or more number of servers of MEC systems and mobile communications makes simulation-based approaches almost irrelevant, a graph-based algorithm is used.
3) Based on the spatial distribution of the communications, finding a MEC partition (the dividing of an area into MEC clusters with each cluster having a single MEC server) that favors the application representation at the edge instead of at the core network.

Continuation...

4) The MILP is evaluated and the clustering algorithm using a dataset of mobile communications in a city provided by a mobile operator. It was shown that the clustering takes into account the spatial distribution of the communications and enables to largely offload of the core.
5) Then, the clustering algorithm was evaluated on larger problem sizes and outline the benefits of MEC even for very small MEC server sizes. The obtained MEC clusters have well-balanced loads and enable to keep a large portion of the traffic at the edge.
6) Finally, the MEC partition was evaluated over a week of communications and shows that it largely supports temporal dynamics.

Conclusion: 1) As operators are transforming their network architectures and are looking for deploying computation resources close to the users to improve QoE (Quality of Experience), it is necessary to adequately dimension MEC systems.
2)In this paper, the problem was formulated as a mixed-integer linear program and presented a graph-based algorithm that enables finding a partition of MEC areas that consolidates traffic at the
edge, in MEC servers.
3)It was evaluated using a real-world dataset from a mobile operator. The evaluation results, beyond quantifying the benefits of the MEC approach, show that the core can be largely offloaded. They also show that the algorithm provides MEC areas that are well balanced in terms of load and close to optimal.

Edge Computing server placement with capacitated location allocation

Aim: The aim is to improve application latencies and reduce data transfer load in the opportunistic Internet of Things systems. Minimizing the latencies that are approximated through different distance measures, (2) minimizing the deployment costs of the servers while limiting the maximum latency, (3) optimizing the trade-off between latency and deployment cost, or (4) maximizing the user connections, i.e. coverage, within the clusters.

Algorithm: PACK: It minimizes the distances between servers and their associated access points while taking into account capacity constraints for load balancing and enabling workload sharing between servers. It also includes practical concerns like prioritized locations and reliability concerns.

Working: 1. The solution of two major problems i.e. first, the placement of edge servers, and second, the allocation of computing capacity to those servers to suit their expected workload is optimized using the algorithm.
2. The algorithm works in a very effective manner to reduce latencies between access points and server as well as a balanced load between the edge servers.
3.The algorithm allows prioritizing server locations and provide reserve computational capacity as a reliability measure for edge application execution.

Continuation...

4.A NP-hard complex optimization problem that seeks the best compromise between distance and balance.
5.A block-coordinate descent algorithm is used for the placement of the server.

Conclusion: Solving the edge server placement problem provides the foundation for further study of other edge resource provisioning problems, such as online load balancing. In this paper, we presented and evaluated a novel edge server placement algorithm, PACK, that considers an extensive set of functional and non-functional parameters to find an optimal server placement under a set of algorithmic and pragmatic constraints.

Joint User Association and VNF Placement for Latency Sensitive Applications in 5G Networks

Aim: The core idea behind MEC servers (multi-access edge computing) is to meet the rising demand for IoT devices and an unprecedented amount of generated data. The idea behind MEC is to move data, virtualization, and processing capabilities from central data centers to the edge of the network. In this paper, we study the problem of joint user association, VNF placement, and resource allocation, employing the mixed-integer linear programming (MILP) technique to minimize (i) the service provisioning cost, (ii) the number of VNF instances in the network, and (iii) the transport bandwidth consumption.

Working: It is considered three different VNF placement possibilities including (i) the MEC host collocated with gNB that the user is associated with, (ii) cloud data centers, and (iii) reusing the VNF instance that has been already placed on a MEC host collocated with adjacent gNBs. We develop a MILP model with three objectives to minimize (i) the service provisioning cost, (ii) the number of VNF instances, and (iii) the transport network utilization. An efficient user association mechanism leads to better resource utilization, load balancing, and energy consumption. It is assumed that each user requests a service with a certain bit rate and delay tolerance. Upon receiving the service request from the user, the network provider shall make a decision on how to embed the request to the network such as to make sure that the user service requirements are satisfied, while the network resources are used in the most efficient way.

Conclusion: Results demonstrated the outperformance of the MILP-cost algorithm comparing to the other algorithms in terms of CPU utilization that is due to the importance of CPU cost parameter in the objective function. Future aim to tackle the scalability issue of the proposed methods by proposing a heuristic algorithm, which can reach a near-optimal solution in a considerably shorter time scale.

References and Bibliography

PART 3

References and Bibliography

- **IEEE Papers**

1). Low Cost MEC Server Placement and Association in 5G Networks
https://ieeexplore.ieee.org/document/8939566
2). Joint User Association and VNF Placement for Latency Sensitive Applications in 5G Networks
https://ieeexplore.ieee.org/abstract/document/9064145
3). Edge computing server placement with capacitated location allocation
https://www.researchgate.net/publication/334534213_Edge_computing_server_placement_with_capacitated_location_allocation
4). Mobile Edge Computing Resources Optimization: a Geo-clustering Approach
https://hal.archives-ouvertes.fr/hal-02065474/document

- **Courses and other resources referred**

1). Machine Learning A-Z™: Hands-On Python & R In Data Science
https://www.udemy.com/course/machinelearning/
2).IoT (Internet of Things) Wireless & Cloud Computing Emerging Technology
https://www.coursera.org/learn/iot-wireless-cloud-computing
3).lgorithms Building
https://www.coursera.org/learn/algorithmic-toolbox
4).Edge Computing (IBM Cloud)
https://www.youtube.com/watch?v=cEOUeItHDdo
5).3gpp basic info https://itectec.com/spec/6-5-support-of-3gpp-application-layer-architecture-for-enabling-edge-computing

Thank You