End-to-end Energy Efficiency at Service-level in Edge Cloud

Selome Kostentinos Tesfatsion[1]

selome.kostentinos.tesfatsion@ericsson.com

Xuejun Cai[1]

xuejun.cai@ericsson.com

Arif Ahmed[1]

arif.ahmed@ericsson.com

[1]Ericsson Research
Sweden

Abstract

Providing service-level energy efficiency (EE) information could be beneficial in many aspects. It can provide a thorough understanding of the energy footprint of an edge service, enable new service-related value add for the service and support for further optimization of EE. It is generally challenging to provide energy efficiency related information for services in the wide area mobile edge cloud environment from end-to-end perspective, i.e., including efficiency in devices, mobile network and the edge cloud infrastructure. This calls for techniques that support exposure of capabilities of each involved component and modeling for efficient and accurate estimation of EE information. In this work we discuss the challenges involved in providing service-level EE information for edge service covering the mobile network and the edge cloud, here referred to as end-to-end, and our experience in the understanding of it. We also highlight the gap in current capabilities to achieve the intended goal.

1 Introduction

In the last few years, there has been an increasing trend to deploy certain types of applications (e.g., AR/VR, IoT, self-driving cars, gaming, etc.) in the edge cloud in order to reduce the service latency or offload the massive data from

the central cloud or data repository. The edge cloud can be deployed into the mobile network, e.g., 5G, to provide edge services to mobile users connected from a wide area network. Compared to today's huge and centralized Data Centers (DCs), the mobile edge cloud usually consists of many distributed and micro edge sites/DCs. To provide the service to mobile users in the wide area network and meet their performance requirements (e.g., low latency, high data throughput, etc), the edge services could also be distributed in the edge environment. When the client is accessing the edge services, the data traffic associated with the service request could go along the path from the client to the service instance(s) deployed in the edge cloud, which includes the radio network, transport, core network, and the edge cloud infrastructure.

Over the years, energy consumption has become an important concern in the management of systems due to issues associated to high cost and high carbon dioxide emission. For example, for the mobile network, Vertiv estimated in 2019 that the move to 5G would likely increase the total network energy consumption by 150–170% by 2026 [1]. Parallel to this, cloud computing is also expected to require 9% of all electricity supply by 2030 [2]. As a result, a number of efforts have been made to improve efficiency at different management layers. For example, in the cloud a number of approaches have been proposed to optimize efficiency at data center facility, rack, server, and component levels. For the mobile network as well, energy efficiency for the whole network, sub-networks (e.g., core, transport, radio access), single network element/function and equipment has been considered.

There is increasing demand for understanding and managing the energy consumption and efficiency at fine granular service level. For example, some public cloud providers, such as Amazon Web Service (AWS), provide the customers the carbon emission of their services deployed in the cloud. Providing service-related energy information can be beneficial for several reasons: service or infrastructure providers can have a thorough understanding of the energy footprint of the services; service providers may charge depending on the energy consumption (EC) of the service, and hence incentivize for the development of a more EE service; it can be useful to assess and compare mechanisms or solutions with respect to EE of services and to make informed optimization decision to reduce energy usage while fulfilling other requirements.

For edge services, many works have focused on the energy consumed in the edge cloud infrastructure either using the physical server or virtual instances and in some cases the energy consumed in network that connects between edge sites or nodes. However, this energy consumption is only a portion of the total energy consumed by the edge services. Energy is also consumed to handle and process the traffic associated with the services in the different parts of the network, including physical and virtualized network functions in RAN, transport, core network. Especially for those edge services that run in wide area mobile edge, the energy consumption of the different traffic paths in the mobile network could be high. Hence it is important to consider and aggregate the EC and the EE of the edge services from the end-to-end perspective, i.e., including both the mobile network and the edge cloud infrastructure. In this work, we discuss the

challenges and our progress in the understanding of end-to-end (E2E) service-level EE covering the wide area mobile network and edge cloud.

2 E2E Service-level EE

The understanding of energy efficiency requires the ability to understand *how metrics are defined*. Metrics provide a better view of how energy can be optimized. Generally, there is no single energy efficiency metrics that is applicable in all cases. For example, in the mobile network it can be defined using the amount of needed energy per unit of traffic, per unit of connection or per unit of revenue [3]. The performance related metrics in the cloud may consider the amount of work performed, e.g., data volume, number of requests/transactions, and latency. The choice of metrics may vary from case to case, and the selection generally becomes more difficult in cross-domain environment. Based on our understanding, the most common metrics for service-level efficiency are performance-based metrics, such as energy per unit of traffic/throughput, due to ease of comparison.

As mentioned in the previous section, the edge services can be distributed and be accessed by the end users connected from a wide area network. The E2E EE would consider the efficiency in the edge sites for hosting and running all service replicates and the efficiency in the mobile network for accessing the service and carrying traffic. These would require the ability of *how to efficiently collect and calculate the E2E EE of the edge services*. In the mobile network, the service could be accessed from multiple UEs and the traffic from the end users could go through different parts of the network. Therefore, there will be energy consumed to handle/process the traffic in the different parts of the network. One way to extract the energy related information in the mobile network could be to consider the traffic paths from the end users, i.e., UEs, towards the replicates in the edge sites. Traffic paths can be defined based on the serving area of the service replicates. That is, when an end user is accessing the service replicates running in an edge site from any place within the service coverage area, the performance would meet the minimum requirement of the service based on, for example, optimized deployments of service replicates. The traffic paths in these coverage area could then be considered to calculate EE. For the E2E EE, multiple options may exist to collect and aggregate EE information. Some options could be: 1)Per-path: EE of a service is determined based on the aggregation (e.g. average, weighted average) of the EEs of the individual paths in the whole coverage area. 2) All-paths: the EE is estimated based on the overall energy efficiency of all the paths for the service replicates. One alternative is that the EE can be estimated based on the E2E EE covering both the mobile network and edge cloud of all the paths while in another alternative is that the EE can be calculated by aggregating the EE of all the paths in the mobile network and EE of all the replicates in the edge site(s). 3) Per-coverage area: the EE can be calculated by aggregating the EE of the paths in all the coverage areas. The choice of options may have different implications. For example, in the per-

path option and a more accurate all-paths option, path information may not be available if the paths are operated by multiple connectivity service providers. In the per-coverage area case a path belonging to more than one coverage area will be considered more than once in the calculation due to overlapping between coverage areas. This may lead to overestimation of the overall EE of the service.

Another aspect in the *EE measurement is the challenge in extracting service-level information in the mobile network.* EC and performance information in the network functions and devices (e.g., routers, switches) may not be available at service level. In addition, it might be challenging to estimate the EC of some NF attributed to the service, ie., in how to determine the proportion of the EC of a NF that is used by the service. The network devices in the data plane (e.g., base station, user plane function, etc) and transport devices consume energy for the data traffic directly. However, control plane functions (e.g., the AMF, SMF, etc) have no direction relation to the data traffic of the service and it is not easy to measure or estimate the energy consumption caused by a service. In general, the control plane could have much less contribution to the energy consumption of services when compared to the data plane. However, their usage may need to be taken into account for a more accurate estimation.

3 Conclusion

In this work, an E2E service-level EE for a wide area mobile edge cloud has been considered. Challenges related to the selection of EE metrics, collection and consolidation of EE related information, and service-level EE information in the mobile network have been discussed together with the progress of our work in this direction. As part of our conclusion, we see that it is challenging to model the E2E EE using the current capabilities provided mainly by the mobile network. In moving forward, it would be important to extend current capability of the network to provide service level energy efficiency information, have new capabilities and exposure logics in place both in the mobile network and the edge cloud, exposing the required information preferably in a standardized manner, and selection of appropriate aggregation method based on the given scenarios.

References

[1] "MWC19: Vertiv and 451 research survey reveals more than 90 percent of operators fear increasing energy costs for 5G and edge," accessed 2022. [Online]. Available: https://www.vertiv.com/en-emea/about/news-and-insights/news-releases/mwc19-vertiv-and-451-research-survey-reveals-more-than-90-percent-of-operators-fear-increasing-energy-costs-for-5g-and-edge/

[2] "Inefficient clouds that devour electricity – new research collaboration between Umea University and Ericsson (umu.se)," accessed 2022.

[Online]. Available: https://www.umu.se/en/news/ericsson-and-umea-university-_11323660/

[3] "Going green: benchmarking the energy efficiency of mobile," accessed 2022. [Online]. Available: https://data.gsmaintelligence.com/api-web/v2/research-file-download?id=60621137&file=300621-Going-Green-efficiency-mobile.pdf