Eco-Design and Optimization of the Edge Cloud

Rickard Brännvall
RISE Research Institutes of Sweden AB
Luleå, Sweden

Tina Stark
RISE Research Institutes of Sweden AB
Luleå, Sweden

Jonas Gustafsson
RISE Research Institutes of Sweden AB
Luleå, Sweden

Mats Eriksson
Arctos Labs Scandinavia AB
Luleå, Sweden

Jon Summers
RISE Research Institutes of Sweden AB
Luleå, Sweden

***Abstract*—Edge computing is proposed to support new digital services such as live video processing, 5G, artificial intelligence, virtual reality, robotics, IoT, and direct monitoring and control of cyber-physical systems, as these applications are assumed to place higher requirements on the digital infrastructure in terms of latency, privacy, and capacity for high-performance computation. There is however a risk that more heterogeneous configurations dimensioned for very high utilization scenarios will leave hardware idle during normal use. This short note reports on numerical investigations of where to place computation workload in an edge-cloud network topology considering the trade-off between the location-specific cost of computation and data communication.**

I. INTRODUCTION

The provisioning of edge capacity introduces cost considerations that need to be accounted for in the design of a distributed edge-cloud infrastructure, that is, a network that combines traditional large-scale centralized cloud computing with a multitude of distributed edge data center (DC) nodes. The computation cost can vary between different locations in such a network, for example, because of economies of scale aspects that may make smaller DCs less efficient [1]. The worldwide electricity consumption of DCs is significant, which is why efficiency matters when considering carbon footprints. Data communication further increases the energy consumption to a degree comparable to the computation itself [2].

Cost of computation: A standard measure of DC energy efficiency is PUE, which is the ratio of total energy consumed in the DC over the energy consumed by the IT hardware for productive computational work. According to [3], the average PUE in the world for 2021 is 1.57, although data centers located in the Nordic countries show lower PUE due to colder ambient conditions and free cooling. Server hardware can draw somewhere between 25-60% of its maximum power when idle, which is why the utilization of the individual server should be kept high to operate efficiently. Energy consumed by DC as a whole also does not scale linearly with its workload as the cooling system introduces additional overheads and rapidly growing (superlinear) energy costs with higher utilization. These are only two reasons why the (instantaneous) PUE at any time can be different from the (longer-term) average.

This work was supported through the Datacenter Innovation Region (DIR) project at Luleå University of Technology and the Celtic Next project AI-NET Aniara, nr C2019/3-2.

Cost of communication: The cost of communication, that is, to move data from the device to the DC, shows a decreasing trend according to recent estimates [4]. In terms of hard numbers, [5] estimates the energy consumption to be 0.08 kWh/GB as an average across the network. A few years later, [6] estimates the marginal energy cost at a much smaller value of 6.7e-5 kWh/GB. The former (and higher) estimate is calculated as an average of the total cost, including the communication infrastructure. It thus considers the cost of a potential expansion of the infrastructure that a substantial increase in total communication might warrant if the limit of data transfer capacity is met. The latter (and much lower) estimate represents the marginal cost of communicating an additional unit given a fixed data communication infrastructure.

In this work, an edge DC is considered to be a small DC furthest out in the network, close to the end-user devices designed to meet some unique constraints, such as latency or privacy. For example, when the application would not fulfill its purpose if running in a remote cloud too far from the device for fast enough data communication or where privacy/confidentiality considerations mandate data processing within some physical boundaries. For flexible workloads, i.e., loads absent of such hard constraints, it may be more optimal to select a particular location over another. Therefore, we investigate the question:

Where is the most resource-efficient location to place a specific workload at a particular time in a dynamic edge-to-cloud continuum?

Resource efficiency is intended to be read in the broader meaning that factors in energy efficiency and environmental footprint. It relates to the two aspects 1) the optimal placement of a task such that the selected assets complete the most computational work for the least energy at any given time, and 2) the investment decision and design of an edge-cloud data center infrastructure. The first aspect concerns the marginal cost of producing an additional service in a given edge-cloud infrastructure. In contrast, the second relates to the overall cost of building and operating such infrastructure.

II. RESULTS

Here we summarize results from numerical experiments with two models also reported with more detail in [7]. The intention is to illustrate how cost savings can be obtained by moving data and computational tasks to different locations in the edge-to-cloud continuum. An optimization perspective on edge compute is considered, assuming that premises are already built for other reasons, such as the deployment of 5G, with edge data center hardware provisioned to meet latency, privacy, and computational capacity requirements. We then assume additional flexible workloads that originate in the same locales, but that can be deployed with perfect discretion either at the edge or at a centralized cloud data center.

A short presentation of each model and a summary of our results from its use in simulation follows.

Model 1: We consider three types of workloads: 1) Web Server, 2) Video Compression, and 3) Image Manipulation, each taken from the SPEC benchmarks [8] with different requirements for computation and data communication. The energy cost is scaled up by the average PUE of the data center that carries the workload, where centralized DCs are assumed to be more efficient than edge DCs. The communication cost adds up for each level traversed in a tree-like edge cloud topology. Both compute and communication are marginal costs and do not depend on the total utilization of the hardware.

Result 1: The model was run for the three different workloads on each layer first with a fixed PUE of 1.57 and then with a variable PUE that depends on the assumed data center size and efficiency. Each PUE assumption was tested with both the higher and the lower communication costs. At the higher communication cost, the model showed substantial energy savings (10% – 65%) from placing communication-heavy Video compression and Web server workloads at the edge DC node instead of at the cloud DC. These results were similar for both PUE assumptions, although the savings were smaller if the edge DC node had worse PUE than the cloud. There were no significant savings when we instead assumed the lower communication cost.

Model 2: This model considers the aggregated cost for all workloads on DC nodes in the network. Instead of using a simple PUE factor for each node independent of how much workload it carries, energy efficiency now depends on utilization; an edge DC node now has an operation sweet spot, where the overall DC efficiency is at its best when the computational load is at around 60 – 80% utilization. Experiments with Model 2 aim to illustrate how computational load can be redistributed to the edge to bring its hardware closer to the sweet spot by comparing local and global optimization strategies over a fleet of hundreds of edge DC nodes and an (elastic) cloud DC.

Result 2: The local and global optimization strategies that place more of the flexible workload at the edge show

(a) 4 – 6% of operational energy savings, and
(b) 50% lower average hardware requirements,

compared to the base scenario of placing all flexible workloads at the central cloud DC. For (a), it is a net saving across both cloud and edge nodes due to better utilization of servers now running closer to the operation sweet spot. The hardware savings (b) are possible in the cloud DC since less of its computational resources are needed. For our simulated example of 100 edge data center nodes with 24 servers per node, cloud capacity requirements drop from 840 to 360 servers.

III. Discussion

The exercises with Model 1 illustrated the benefit of placing data-hungry workloads at the edge closer to the end user devices to avoid unnecessary communication with a central DC and showed savings of up to 60% under the higher communication cost assumption. These results were quite robust for different PUE. However, the effects almost disappear when we use the lower communication cost estimate based on the marginal cost of adding an extra data packet on a communication network that is already up and running. A limitation of Model 1 is that it ignores the cost of idle servers. Model 2 addresses this by factoring in the instantaneous utilization of hardware. The simulations show moderate energy savings (4–6%) by placing more workloads at the edge to bring it closer to its operation sweet spot. Furthermore, the reduced use of the central cloud DCs could enable substantial savings of its hardware, pointing in the direction of a future leaner cloud.

The savings on energy bills and hardware investment need to be put into dollar terms to be comparable. It would require assumptions on amortization models, hardware procurement, the (regional) price of electricity, and local environmental circumstances that affect PUE, which is beyond the scope of this short note. If also LCA costs of the equipment are accounted for, the benefits of utilizing edge equipment more optimally should be even more apparent. A complete analysis would translate all energy and LCA costs into an equivalent measure (for example, CO_2 equivalents). However, a comparison will depend very much on modeling details, like assumptions about the energy mix for the region where the hardware is produced and the region where the equipment is operated. We, therefore, leave such analysis for future work.

A premise of this study is that some workloads can be moved freely around in the cloud-edge continuum with perfect discretion, which is, of course, only possible in certain business models for the underlying digital services and ownership of hardware resources. If market solutions do not emerge naturally, one may consider how policy and other incentives can be changed to support this development.

Conclusions: The results from the exercises with the two models point to the conclusion that distributing the workloads over edge data center nodes is motivated by sustainability considerations – we saw savings both in terms of electrical energy and hardware equipment – that would translate into reduced GHG emissions and a lower life-cycle footprint.

References

[1] "United states data center energy usage report," Tech. Rep., 2016.
[2] G. Kamiya *et al.*, "Data centres and data transmission networks," Tech. Rep., 2022, international Energy Agency, Paris.
[3] D. Bizo, R. Ascierto, A. Lawrence, and J. Davis, "Uptime institute global data center survey 2021," Tech. Rep., 2021, uptime Institute.
[4] J. Aslan *et al.*, "Untangling the estimates: Electricity intensity of data transmission," *Journal of Industrial Ecology*, vol. 22, 2017.
[5] J. Malmodin *et al.*, "LCA of data transmission and ip core networks," in *2012 Electronics Goes Green 2012+*, 2012, pp. 1–6.
[6] J. Malmodin, "The power consumption of mobile and fixed network data services," in *Proceedings of Electronic Goes Green 2020*, 2020.
[7] R. Brännvall *et al.*, "Cost optimization by energy aware workload placement for the edge cloud continuum," Luleå, Tech. Rep., oct 2022.
[8] "Standard Performance Evaluation Corporation, https://spec.org/."