Towards a more efficient and easier to use edge computing

ARCTOS LABS

16 June 2023

1 Abstract

Edge computing is currently increasingly applied. Initial use cases tend to be static with respect to which applications should be placed in which edge locations. When additional and more complex use cases emerge, and cloud/telco providers increase their presence with more locations, there needs to be more advanced orchestration capabilities in order to harvest potential performance gains, minimize costs, reduce their CO_2 footprint, or improve resilience by intelligently place application components across the edge-to-cloud continuum.

The article argues that the way forward is to introduce an intent-based concept that centres around the application and its required infrastructure characteristics complemented by cost metrics and a placement engine that can match such intents towards the specific infrastructure that is available.

Arctos Labs provide such a technology.

2 Setting the scene

A lot of expectations are put on the emerging edge computing market. Although real use cases are in the first wave in verticals such as retail or just replacing old on-premises infrastructure with cloud technology. We are still far from the visionary use cases outlined by some proponents, but we can see the market catching on.

Analyst reports like "Edge computing market sizing forecast" from STL Partners [1] indicate growth in the market once more use cases and verticals apply edge computing. Others are more conservative in their short-term forecasts.

There is a wide variety in the definition of edge computing, derived from a large variety of needs. In [6] Linux foundation defines six different layers of compute locations ranging from very centralized data centres to constrained edge locations. This is illustrated in Figure 1. The device edge layers could be devices such as mobile phones, sensors, cameras, or other programmable devices upon which software can be installed. Note that computing on devices may be connected anywhere into a network (public, 5G, or private).

The next location is referred to as on-premises edge, which captures computation within a factory, a retail store, a private data centre of an enterprise, or similar.

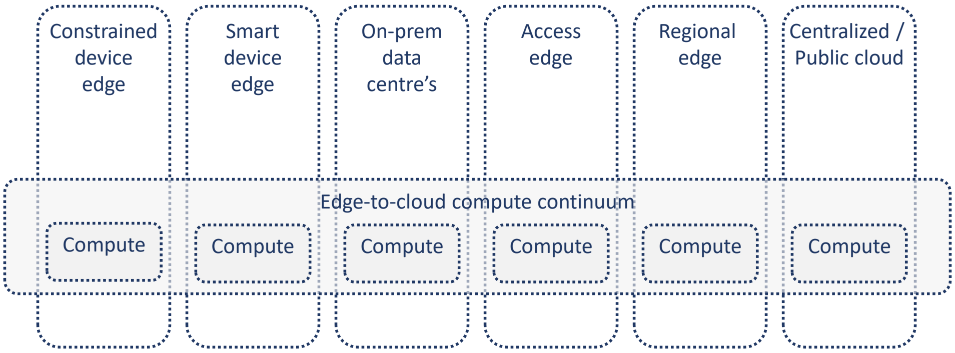

Figure 1: The Edge-to-Cloud continuum

The fourth layer is referred to as the access edge, meaning computing that sits somewhere along the network of a telco operator and is offered as a cloud service, closer to its users.

The fifth location is referred to as regional (or national) edge, with a size that is more like larger data centres enabling pooling gains, but still closer to the users of the data centre. The sixth location, not being referred to as edge, is then centralized locations such as the public cloud.

For the discussion in this article, we need to consider all six locations as outlined above. And we also need to recognize that most IT applications consist of multiple components thus proposing that such components may be distributed along the edge to cloud compute continuum for the overall system to be as efficient as possible. IT applications may also include (use) SaaS components having their specific service access locations.

Application components are further often designed to operate in an environment (or stack) of other software or hardware capabilities. This could be an operating system, a database, data manipulation functions, hardware accelerators, etc. So, these capabilities need also to be considered when optimizing the overall system as well as the application's geographical context.

3 The 2nd and 3rd application problem

Enterprises are increasingly dependent on their IT and the data created and processed by their IT systems. As enterprises operate in a highly competitive environment, it implies that they need to launch new IT capabilities in a very agile manner.

The edge infrastructure (i.e. the fleet of edge nodes) initially deployed for a specific mission will quickly be exposed also to those new applications (2nd and 3rd) and capacity might therefore become scarce.

Those new applications will have their own requirements of residing at the edge, thus introducing competition for the available edge resources. In the best of scenarios, the enterprise can shuffle around some application components to create room for new ones. This may lead to, for example, that some applications reside only at every second edge location in the horizontal dimension, utilizing “east-west” connectivity. In other scenarios, the edge infrastructure is extended at access edge or regional edge locations whereby application components become somewhat centralized (or more correctly deployed in an aggregation point, still being defined as the edge).

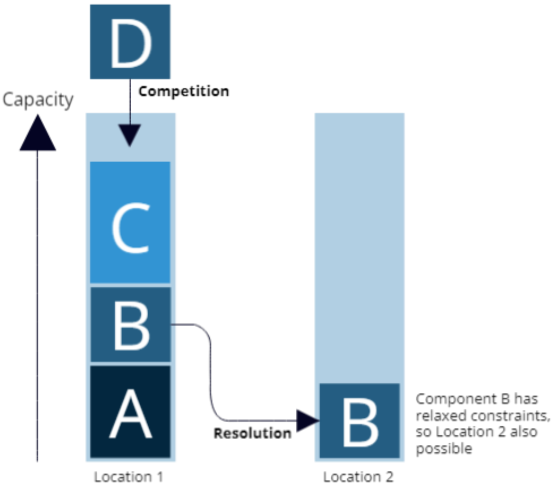
D
Competition
Capacity
C
B
A
Resolution
B
Component B has relaxed constraints, so Location 2 also possible
Location 1
Location 2

Figure 2: Resource competition

Figure 2 illustrates the situation when component D needs to be deployed in Location 1 but the available capacity is not enough. Here the solution would be to move component B to a second-best location (2) where constraints are still fulfilled.

For all these scenarios it is imperative to be able to accurately optimize and control how the placement of the application components is done and what the resulting application characteristics will be. Using labelling or similar techniques will therefore be increasingly difficult, and it is therefore proposed to use a declarative intent concept centred around the needs of the applications and let advanced algorithms resolve the best possible placement for each scenario.

Such algorithms can be applied when planning an application rollout, when planning an infrastructure upgrade as well as being an integral part of application deployment.

Utilizing the above approach, an enterprise can postpone infrastructure investments, and roll out new applications more quickly.

4 Optimizing application characteristics by controlling how they use edge computing

As said, IT applications are composed of components (e.g. smaller applications or micro-services). Some of the components are candidates to be moved further out on the edge-to-cloud continuum, others should reside in a centralized location (e.g. the public cloud) or stay on-premise. The applications (through their components) also have data ingress or egress points that need to be considered. Move processing closer to data ingress or egress points is one of the main ideas of edge computing, but there are often other IT applications and the location of their components that needs to be considered, such as components residing at enterprise on-premises or data egress at the other end of a data pipeline.

It means that the total application performance is subject to the interaction between the set of its components, the data ingress, and egress points, also including platform components or SaaS components. Characteristics such as throughput, elasticity, scalability, and geographical reach with certain latency need to be well understood and can be a complex task. Further, security and similar aspects need to be considered.

How to best place individual application components on a specific edge infrastructure thus is depending on the characteristics of how the edge nodes are interconnected as well as the amount and type of resources available at each location. The edge infrastructure (including the networking) may have characteristics that vary over time, thus leading to the resulting placement optimization needing to adapt accordingly. As the edge nodes may have a very limited resource capacity, competition may force some application components to be deployed in second-best locations. It is not always the case that deploying as far out as possible in the edge-to-cloud continuum is the best as more central locations will have a different economy of scale.

On the other hand, lower utilization on smaller edge locations may impact the cost for produced services. This is because server energy consumption does not scale proportional to the load of the server. The implication is therefore that energy consumed over work done exhibit different values along a server's work span.

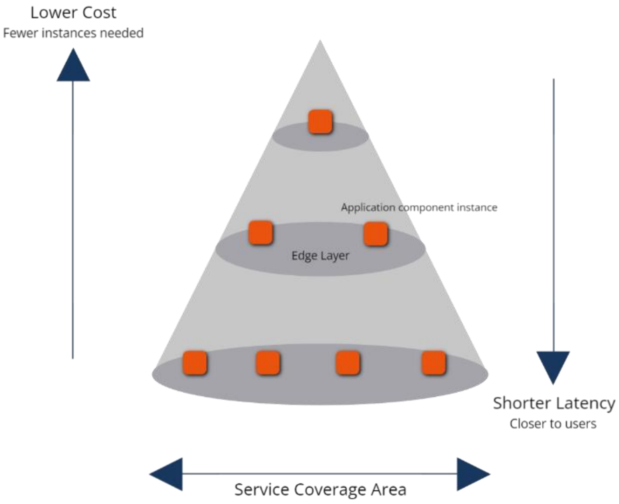

Figure 3: Aspects on Latency and number of components needed to cover a certain area

Figure 3 illustrates the aspect of that when attempting to cover a certain geographic area the number of replicas needed is depending on how far out in the edge to cloud continuum the components are deployed, which in turn is depending on the required latency. Reducing the number of replicas would reduce cost.

It is desirable to avoid linking application components to specific edge infrastructure locations as much as possible so that applications can benefit from edge nodes that are added. This calls for a recurring placement optimization to avoid fragmentation and non-optimal utilization as well as further reducing costs.

It is important to understand the implications of total application performance before deploying an application across the edge to cloud continuum. Utilizing edge to cloud fully will also impact the way that applications are designed. Today we refer to cloud-first or cloud-native, but we need to better define edge-native as a term to capture applications that can scale out over a fleet of edge locations to make use of being closer to where data is produced or used.

5 Energy efficiency aspects

The need for reducing CO_2 footprints will not go away! Global data centres consumed between 0,9 - 1,3% of global electricity usage with an additional 1,1 - 1,4% consumed by data transport [2].

These figures have been approximately flat for some years now although the computing capacity of the total fleet of data centres has increased substantially. I.e. data centres and data transmission networks have become increasingly more efficient. This is often due to their scale of operation as well as contributions from Moore's law. There are however reasons to believe that the effects of Moore's law are declining [5].

With the expected growth of edge computing, the portion of global computation contributed by edge will grow significantly and will quickly constitute a significant portion of the overall computing capacity. Such an evolution thereby also means that the efficiency of edge computing becomes a critical part of the overall computing efficiency and thereby to the electricity consumption.

Observing the energy consumption of a server as a function of its utilization is typically so that 20% of the maximum energy consumption is consumed when the server is idle, and then increases upwards when utilization increases to its maximum consumption at full utilization (simplification). Servers configured in performance mode expose a flatter relation between power consumption and utilization meaning the amount of power consumed at idle is significantly higher for performance mode configurations. The conclusion here is that it is beneficial to try to utilize servers as much as possible or turn them off.

But there are other factors of the total efficiency that need to be considered, relating to what is normally referred to as PUE (power usage efficiency). PUE captures the portion of electricity that goes to IT equipment (servers) compared to the portion that goes to the power supply and cooling. Larger data centres are designed to operate at a sweet spot whereby the power and cooling portion is minimized. For edge data centres with smaller capacity, such as on-premises, access or regional, there will be a larger variety in the instantaneous PUE, due to constraints in how cooling can be accomplished, ambient temperature, utilization, etc. This leads to fans needing to start or even compressors which quickly deteriorate the PUE and such non-IT could consume as much as the IT equipment, whilst at the sweet spot it may be as low as 10-20% also for edge data centres.

To minimize the overall energy consumption, the above implies that application components should be moved into or out of edge data centres to make them operate at the sweet spot. Such optimization could have a significant impact on the consumption of electricity and thus improve the operational cost.

In a joint study from RISE and Arctos Labs it is shown that up to 50% of cloud hardware could be saved by utilizing edge locations properly and under the assumptions of the model of the study [7].

6 The aspect of moving data around

In 2006, British mathematician Clive Humby coined the phrase, “Data is the new oil”. Many enterprises are increasingly producing data as part of their operations and data is often transported somewhere, and often to a public cloud or central data centre.

There are obviously several aspects of such data transport. The first is about the energy consumption of sending data around. Sending less data will consume less electricity obviously.

Secondly, it is about the cost for an enterprise of moving data. The cost of bandwidth from network operators (ISPs) is one element, the cost of ingress, or more importantly egress of cloud providers is another.

These costs may well justify somewhat more expensive computing further out in the edge-to-cloud continuum depending on whether the application is compute-intense or data-intense. An analysis of this aspect is done in [3] and [4].

Thirdly, we must consider the potential flooding of the data transmission networks should all data created travel all the way to central locations. Being aware of the expected data tsunami as the consequence of “the data is the new oil” it becomes common sense to process data before sending the resulting output onward, when possible.

There are also other aspects related to the movement of data, such as privacy or regulations (like GDPR).

7 The ability to host “compute for good” applications on your edge locations

Edge locations are typically much smaller compared to mega data centres. When edge location capacity becomes smaller, it leads to a larger variety in the utilization due to a lack of pooling gain effects, i.e. those locations may be operating at a below-optimal utilization from time to time.

It means that there is compute capacity available that will come at a small marginal increase in energy consumption as explained in previous chapters. In case enterprises do not have applications that can be deployed to such edge locations, the enterprise could consider using the surplus edge location capacity for what is referred to as "compute for good". Compute for good refers to workloads that for example perform cancer research computations or other tasks that generally improve our life here on the planet. Earlier examples of these applications were using screen savers.

Such an approach contributes to running edge data centres at the sweet spot. In order to make the most out of the fragments of computing that can be made available there is a need for advanced placement that takes the interaction between the components of the application into consideration to achieve efficient fleets of such fragments so that more advanced applications could utilize this surplus capacity.

8 The need to evolve towards a true cattle concept for fleets of edge locations with intelligent placement

It is natural to start off by using labelling to categorize your fleet of edge locations. But for reasons outlined in this paper, we believe there will be a need for a more advanced concept when usage of edge computing becomes more complex. It is hard to foresee which labels are needed beforehand, and cumbersome to add them afterward should there be a need to distinguish one location from another that is otherwise identical.

The needed concept should be based on a declarative intent model of what application components require in terms of characteristics and a placement engine that can match such intents towards the specific infrastructure that is available. The placement engine can tap into live infrastructure information that is perhaps further enhanced by applying data analytics, filtering, or even prediction before feeding into such a placement engine.

This concept also makes it possible to implement various optimization schemes, such as cost optimization, energy optimization, or performance optimization as outlined by this article.

The concept also gives improved resilience as a side-effect as it can resolve the optimal fall-back solution in case of failures thereby avoiding the error-prone activity of defining and implementing every disaster recovery scheme.

Such a placement engine will nicely integrate with infrastructure monitoring solutions that provide insights into the infrastructure and its utilization and thereby closing the loop and enabling full automation of life-cycle management. If prediction is applied in the monitoring and data analytics stack, it also enables acting in a proactive manner to trends in the load of the edge infrastructure.

The Arctos Labs ECO (Edge Cloud Placement Optimization) provides such a concept.

9 References

[1] - Edge computing market sizing forecast, STL Partners, https://stlpartners.com/research/edge-computing-market-sizing-forecast/

[2] - Data Centres and Data Transmission Networks, IEA, https://www.iea.org/reports/data-centres-and-data-transmission-networks

[3] – Cost modelling of edge compute, Mats Eriksson, https://www.researchgate.net/publication/344151306_Cost_modelling_of_edge_compute

[4] – Eco-Design and Optimization of the Edge Cloud https://www.arctoslabs.com/white-papers

[5] - Q&A: Neil Thompson on computing power and innovation, MIT News

[6] - Sharpening the Edge: Overview of the LF Edge Taxonomy and Framework https://www.lfedge.org/wp-content/uploads/2020/07/LFedge_Whitepaper.pdf

[7] – Cost Optimization by energy aware workload placement for the edge cloud continuum https://ri.diva-portal.org/smash/get/diva2:1751120/FULLTEXT01.pdf