Placement of Workloads in Edge Compute & Cloud Networks

Mats Eriksson, ArctosLabs Scandinavia AB

1 Abstract

Edge compute promise to deliver cloud computing capabilities for tomorrows applications and exploding amount of data.

Networks with edge compute are complex, and the risk being underutilized and non-profitable is prominent. This is because the delicate balance of proximity to users to achieve performance vs the scale to enable a stable statistical utilization.

This paper introduces Arctos Labs placement technology to enable workload placement optimization that continuously can assess the most optimal distribution of workload components across an edge-to-central cloud continuum in a closed-loop concept.

Placement optimization is a key capability to achieve the needed utilization in a network of distributed edge PoP's. It uses declarative models that specify required service constraints balanced by resource cost models.

This technology can be utilized accompanied by independent SW life-cycle management (orchestrators) to support OTT workloads over multiple service providers, as well as inside a service provider to expose its edge resources.

The technology provides a programmable, open, modular and extendable framework for enterprises and service providers to automate workload placement.

2 Drivers for Edge Compute

Over the last decade, enterprises need for IT & digital communication has accelerated and nowadays are increasingly business critical. An accelerating trend to move workloads to the Cloud accompanied by various co-location alternatives fuel the flexibility available for enterprise IT.

The arrival of edge computing gives additional opportunities for solutions in a world that is increasingly focused on processing of data. Gartner predicts[1] 75% of enterprise data will be processed outside the enterprise network by 2025 (today 10%).

[1] https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders/

Below are some examples of issues that is expected to drive the need for edge compute.

2.1 The issue of Bandwidth

Costs for moving data over long distances will dramatically increase as the amount of data is set to explode due to hungry applications such as Internet of Things (IoT), Big data & Analytics, Artificial Intelligence, Video, VR/AR etc.

Enterprises deploy SD-WAN to combat such networking challenges, but a more holistic networking optimization should also consider the placement of the workload that generate those networking loads. I.e. moving the workload can be assumed to have a great impact on the network load.

The above apply also for networking functionality, such as 5G. Those workloads need to be distributed in an optimal way that fulfils network performance targets without overloading the underlying transport.

2.2 The issue of Latency

The effects of latency may vary a lot, depending on type of application. Some applications do not work at all (e.g. machine control), whilst others may seem to work, although not delivering the intended business value (e.g. trading).

The implication is that workload placement needs to be carefully considered to keep within latency boundaries for those applications. But as scarce edge computing resources should not be unnecessary utilized, applications that are not latency sensitive may be deferred to more central locations.

2.3 The issue of data Privacy & Security

When applying a centralized cloud, the enterprise becomes increasingly dependent upon having a trusted partner for protecting their data. Furthermore, new regulatory aspects apply to many businesses, such as GDPR.

This implies that requirements on enabling control of data placement is emerging rapidly. Such capabilities are needed as a complement to SD-WAN or VPN connections to assure the enterprise is being properly protected.

2.4 The issue of Availability

When compute is moved closer to the edge it may expose an availability aspect with an increasing single point of failure risks. In order to support the enterprise demand for high availability there is a need for a placement function that automatically can select a backup location should the preferred one fail.

3 Achieving Profitability in Edge & Fog Compute

From the above, it becomes clear that there is a range of benefits of moving compute closer to where the data is consumed or generated. A frequently asked question is then: How far?

If we consider the user (edge) to central cloud continuum, we can identify a range of possible locations that are typically interconnected through a transport network of one or several service providers.

As PoP's with a smaller user base (closer) can be assumed to have greater load variation, infrastructure operators need to balance the proximity to users with an acceptable utilization. The ability to deliver an SLA will obviously be dependent on the proximity as well as the probability to have compute available when needed, which then becomes contra dictionary. This issue is further complicated as an edge PoP is likely to have to deliver different types of compute, leading to potential fragmentation. Remote and smaller PoP's will typically imply higher OPEX but the total cost of a network consists also of the cost of moving data across the network.

The above imply that the workload distribution across such a network of edge locations are highly critical to achieve profitability. As applications are increasingly micro services, the individual components need to be carefully placed to achieve high utilization and resulting economics, whilst at the same time deliver the performance needed.

It becomes quite clear that such placement optimization needs to evolve from a manual activity towards a fully automated.

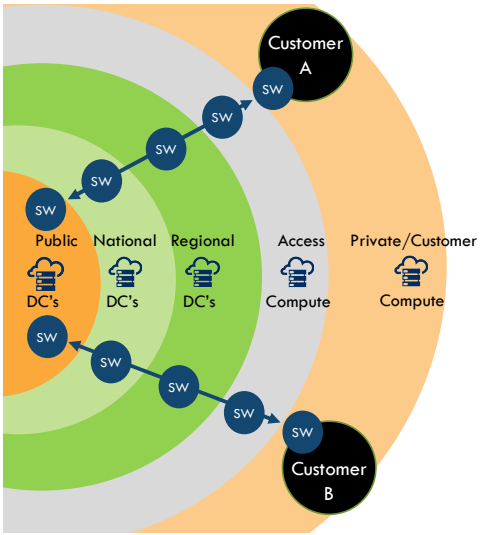

Figure 1: The Edge - Centralized Cloud Continuum

4 Introducing an Open & Model-driven Cost-constraint Placement Optimization technology

To address the need for edge compute automation, Arctos Labs have developed a placement optimization technology.

This placement optimization will operate in a context as illustrated below:

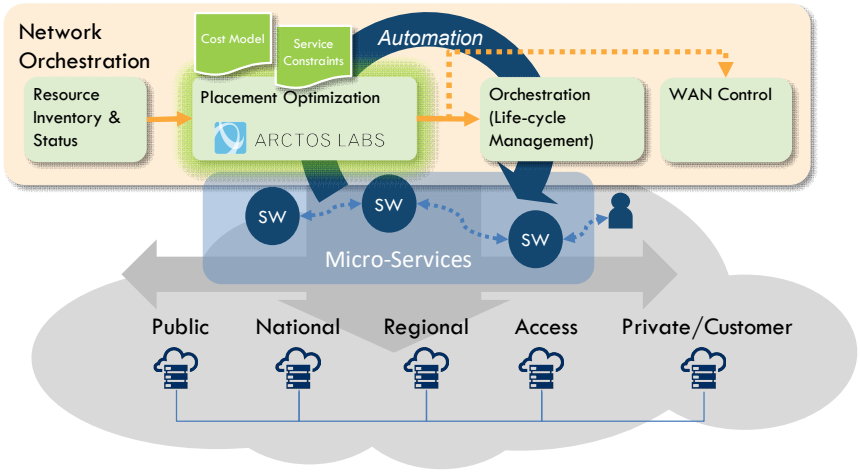

Figure 2: Overview of the Arctos Labs optimization technology

The technology can continuously optimize placement of individual components of a service over a central cloud to edge continuum. In order to do so, the placement interacts with a life-cycle management (orchestration) capability to execute the placement result. This way, the placement can be combined with different orchestrators. This makes the technology suitable both for enterprises that want to utilize edge compute as well as for service providers that need to expose an edge compute service.

This technology adds an important component into an overall automation solution that is complemented with other components.

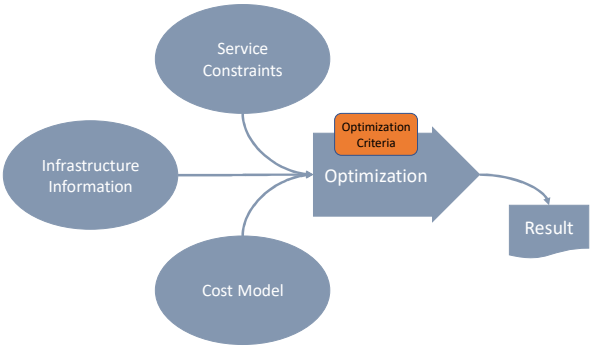

Figure 3: Overview of the optimization concept

It is based on the following model concepts:

4.1 Service specification model (constraints)

This means that any composed SW service subject to placement optimization will be accompanied by a model that describes requirements (constraints) on the underlying infrastructure for the composed SW service to operate satisfactorily.

This model is used by the placement optimization engine to exclude potential placement alternatives that do not meet the constraints specified.

The model is an artefact that is either defined at design time by the application service programmer, or at on-board time by a service provider as a result of integration activities. The model captures two aspects.

1. **The topology of the Service**

   This part of the model defines the SW service break-down into Service components, as well as the links for data interaction between those.

   This model can for example be a Network Service Descriptor in NFV scenarios, or a TOSCA model in OTT cloud scenarios.

2. **The constraints of the Service**

   This part defines in an abstract and declarative way the needed characteristics of the underlying infrastructure. This part includes compute specifications (type and amount), storage as well as constraints on the links interconnecting the service components.

4.2 A Cost model

One aspect of placement is the cost related to the alternatives. This model captures a "cost" of using resources, so that a cost aspect can be calculated as part of the placement process.

Cost in this context is not equal to price. Cost is a factor by which we can capture that certain resources are preferred over others in each point in time.

Below is a couple of examples on how the cost model may be applied.

- **Create a bias for certain resources**

  If the provider wants to avoid using certain resources (if possible), in favour of others they can be given different costs. The cost difference may be based on resource types, resource location or current resource utilization.

  This cost calculation can be subject to complex, dynamic algorithms

- **Resource bundles**

  The resource amount is often bundled and having a cost associated to each bundle. An example is that a 100 Mb/s link may not be 10 times more expensive than a 10 Mb/s. Furthermore, there might not even be a 32 Mb/s link type available.

  The above implies that there can be specific resource bundles (types) to choose from and each of them can be given a specific cost.

- **Real pricing & 3rd party resource providers**

  As placement can include public clouds or communication links from 3rd parties, the optimization needs to capture costs from such 3rd party providers. In these cases, the cost for such resources equals their pricing.

Therefore, the cost could either be a fixed price list as agreed with the provider, or a cost retrieved over an API in real time.

4.3 An infrastructure model (information)

The placement optimization needs to have access to information on the underlying infrastructure network over which it shall consider placement.

The first very basic information needed is the available set of PoP's (Points of presence), i.e. the various compute locations possible. For each location (or type of location), there needs to be information available that defines the type of resources available in the location. This information also points to the cost model.

Furthermore, there needs also to be information available on communication link attributes that interconnects the PoP's. This includes for example latency, jitter etc. Communication attributes also include links from Customer Premise Equipment (CPE) to the PoP's.

In order to enable continuous placement optimization, the infrastructure information needs to be continuously updated on a regular basis to reflect the status of the infrastructure.

4.4 The optimization criteria & functionality

Based on the information models listed above, the placement optimization engine will conduct an optimization according to a specific criterion.

The process can be simply described as:

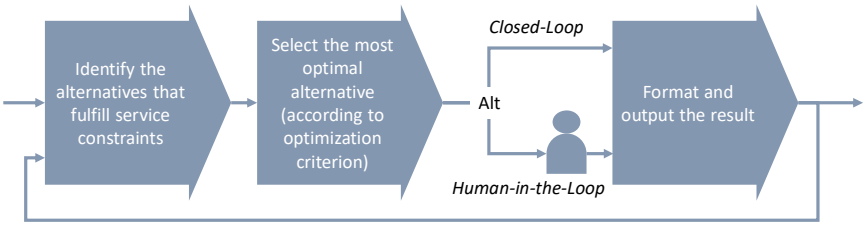

Figure 4: Overview of the optimization process

The most straightforward example of this would be to optimize for lowest possible cost. It basically means that out of the possible placements that fulfil the constraints, the alternative with lowest cost is selected. There could also be other optimization criterions, such as lowest possible latency (omitting the cost factor). Note also that cost can be a constraint, so that only alternatives below a certain cost is considered.

4.5 Optimization criterion

The most important aspect of the optimization criterion is that it needs to be business specific. This means the optimization criterion needs to be "programmable" in the sense of that it needs to be defined from the business priorities of the service provider enterprise.

4.6 Human-in-the-Loop

One important aspect of the optimization loop is to be able to decide whether optimization should operate as a closed-loop or as a human-in-the-loop.

Putting the human in the loop is important to be able to increase trust, so that any output from optimization can be reviewed prior to actions being taken. Such a capability is also needed for what-if and simulation scenarios.

5 Use cases and scenarios

Edge applications can generally be classified into three different categories:

1. **Data pipelining**
   These are use cases where a large amount of data needs to be processed and condensed into information prior to further application functionality. Examples of these are: IoT, Video surveillance, video caching.

   The placement optimization can in these categories find the places for the data processing elements to reduce data transport costs and performance optimal.

2. **Device off-loading**
   This category of use cases is those where the compute capacity of the user device is either limited or battery powered, and additional compute can be delivered by the network. Examples are AR/VR, Artificial Intelligence, IoT sensor data processing.

   The placement optimization can in these categories find suitable compute locations that can off-load devices in a way that meet required latency.

3. **Local clustering**
   These use cases connect a set of devices into a cluster. The cluster is typically defined from required communications characteristics so that the devices can coordinate its work tasks. Examples are V2X, gaming and drone control.

   The placement optimization can in these categories make sure that devices connected to the cluster are connected within the intended specification constraint, including any needed compute host.

The placement technology can enable the following features in an overall system:

- **Multi-provider cloud broking** – ability to select the most suitable service provider for placement of workloads

- **Hybrid- and Multi-cloud** – the capability to optimize where workloads are placed to minimize overall costs and/or achieve required performance constraints

- **Dynamic workload placement for cost optimization** – the ability to dispatch sporadic workloads onto suitable PoP's based on cost optimization

- **Battery optimization of wireless devices** – capability to optimize between radio transmissions energy consumption and compute consumption and place workloads to achieve best possible battery performance

- **Compute arbitrage** – the placement technology can collect non-utilized compute at edge sites and expose them for additional services and achieve higher utilization

- **Self-healing** – the ability to select alternative places for workloads in case of failures on links or PoP's

6 Proof-of-Concept

Arctos Labs have together with Telenor, Windriver and Netrounds developed a PoC integrated with ETSI Open source MANO. This PoC have been appointed as an official OSM PoC [2]

The PoC utilizes four PoP's each having it's dedicated OpenStack and Netrounds vProbe. The vProbes continuously monitor inter-PoP links through active testing and makes such information available through the Netrounds control center. This information alongside other inventory data is consumed by the placement engine for placement.

The optimization result is subject to service constraint models, cost models and optimization criteria selection with its output being forwarded to the ETSI MANO which executes the SW life-cycle management.

The PoC can demonstrate how different NFV network services becomes distributed across the four PoP's included.

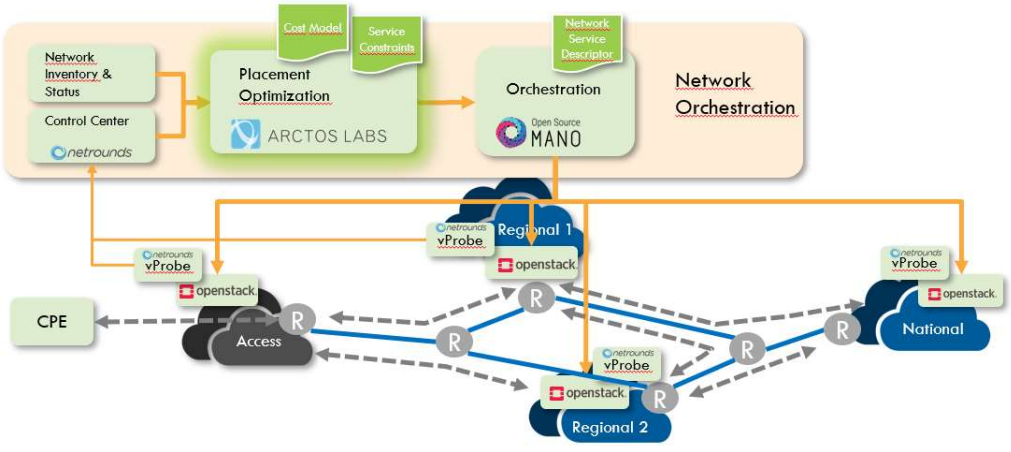

Figure 5: Overview of placement PoC

[2] https://osm.etsi.org/wikipub/index.php/OSM_PoC_5_-_Placement_of_Workloads_in_Distributed_Cloud_Networks