# Service Level Agreements (SLAs) for an Edge controller

## Introduction

Service Level Agreement (SLA) is a contract, between the service provider and the customer. It specifies the key service objectives, metrics, acceptable quality levels and compensation.

With the rise of Internet of Things (IoT), billions of new heterogeneous devices are connected to the internet. With the critical requirements on low-latency, high bandwidth which the apps have it is important to have these operations processed in a very short time. The idea of edge computing is developed to offload compute intensive operations to the edge of network, near to the user to be processed in a short time. An edge is a small datacenter located in short physical distance, therefor the latency for the data latency is lower. An edge is a simple data center and they could potentially fail and strategies are developed to handle these potential failures without disrupting a running application.

In this SLAs document, the terms which represents the above specifications which are agreed by the service provider and the customer is defined. In the first section the parties which are defined who is responsible for the service. Under the periodic review section it is defined how often this agreements will be reviewed. The third section, Service Definition describes the service in detail. The next sections of this document defines the SLAs, how these agreements are managed, exclusions, responsibilities, measurements and penalties and right of termination.

## Parties

This SLA is between the following parties:

The client

The service provider

## Periodic Review

This document will be affective from the 16th November, 2016 and will run for a period of 2 weeks.

The SLAs may be reviewed at any point, by mutual agreement. In the event of a change of the clients IT system the SLAs may be reviewed.

## Service Definition

An edge controller which efficiently distribute the workload among the edges. The expected workloads, failures which are to occur and load balancing migration of VMs are taken into consideration.

The edge computing infrastructure consists of VMs, PMs and Micro-datacenters (edges). Each edge includes maximum 10 PMs but the number of running PMs differs due to the work load. The edges are placed in a grid and one edge is placed at a position in the grid.



Figure 1: Edge in the grid

The requests for VMs are generated randomly by uniformly distributing the creation time and duration. There are some failures expected such as hardware failure, VM failure, application failure and edge failure. Failure tolerance and management will be discussed in the next section of this document. Migration of VMs are performed as the user moves to make sure the app is still in a small physical distance to the user without disrupting the existing connection. When migrating an image of VMs physical memory is pushed across the network and the dirtied images will be re-sent till the dirtied pages falls below a certain threshold.

## Service Agreement

#### VM characteristics

The agreed characteristics of the VMs will be as follows

|  |  |
| --- | --- |
| CPU | 4 Core 3GHz |
| Memory | 4GB |
| Network bandwidth | 54MBit/s |
| Running time |  |

#### PM characteristics

The agreed characteristics of the PMs will be as follows

|  |  |
| --- | --- |
| CPU | 4 Core 3GHz |
| Memory | 8GB |
| Network workload |  |
| Energy utilization |  |

#### Micro-datacenter (edge)

The agreed characteristics of the edges will be as follows

|  |  |
| --- | --- |
| Number of edges | 1000 |
| Number of PMs per edge | 10 |
| Number of edges placed at a position | 1 |
| Network bandwidth between edges |  |
| Location | Known location |

#### Latency

The service will have a 70ms latency while performing the tasks.

#### Service Availability

The service (edges, PMs and VMs) will be available 24h and will be up and running 99.95%.

#### Migration of tasks

VM physical memories are migrated as images and will be re-sent till the page dirtiness falls under the threshold, which is 99%

#### Failure recovery time

Any failure will be recovered depending on the severity of the failure (please refer to the section Measurements and Penalties).

|  |  |
| --- | --- |
| Severity | Recovery time |
| Low | 6hrs |
| Medium | 4hrs |
| High | 2hrs |

## Service Management

This section describes how the service is managed in case of failures. We have identified 3 types of potential failures.

VM failures

PM failures

Edge failures

These failures are handled using an algorithm to make sure the task is completed smoothly without any disruption. To minimize the failures all tasks re divided into subtasks and distributed so failed components will be avoided. Furthermore, the system will be rebooted periodically and also in case of an unavoidable failure, the tasks will be migrated to a different PM.

## Exclusions

App failures are excluded from this SLA document.

## Measurements and Penalties

|  |  |
| --- | --- |
| Failure | Severity |
| VM failures | Low |
| PM Failures | Medium |
| Edge Failures | High |

There will be certain compensations according to the occurrence and the severity of the failures.

If an edge fails more than twice or a PM fails more than 10 times a day, 50% of the daily fee will be reduced.

If an edge fails more than 5 times a day the 100% of the daily fee will be reduced.

## Right of Termination

If the service constantly fails constantly to deliver the agreed service levels described in this document, the client may terminate the service contract with no penalty. However, the termination right is only available if the failures of the service level occurs more than 20 times in a single calendar month.