A Cognitive Serverless Framework for the Cloud-Edge Continuum

D4.6 COGNIT Serverless Platform - Software Source - a

Version 1.0

31 October 2023

Abstract

COGNIT is an AI-enabled Adaptive Serverless Framework for the Cognitive Cloud-Edge Continuum that enables the seamless, transparent, and trustworthy integration of data processing resources from providers and on-premises data centers in the cloud-edge continuum, and their automatic and intelligent adaptation to optimise where and how data is processed according to application requirements, changes in application demands and behaviour, and the operation of the infrastructure in terms of the main environmental sustainability metrics. This document offers a catalogue of those open source software resources developed in WP4 “AI-enabled Distributed Serverless Platform and Workload Orchestration” during the First Research & Innovation Cycle (M4-M9) as part of the implementation of several of the main components of the COGNIT Framework (i.e. Cloud-Edge Manager and AI-Enabled Orchestrator).

This project is funded by the European Union’s Horizon Europe research and innovation programme under Grant Agreement 101092711 – SovereignEdge.Cognit

Deliverable Metadata

Project Title:	A Cognitive Serverless Framework for the Cloud-Edge Continuum
Project Acronym:	SovereignEdge.Cognit
Call:	HORIZON-CL4-2022-DATA-01-02
Grant Agreement:	101092711
WP number and Title:	WP4. AI-enabled Distributed Serverless Platform and Workload Orchestration
Nature:	R: Report
Dissemination Level:	PU: Public
Version:	1.0
Contractual Date of Delivery:	30/09/2023
Actual Date of Delivery:	31/10/2023
Lead Author:	Monowar Bhuyan (UMU) & Paul Townend (UMU)
Authors:	Malik Bouhou (CETIC), Aritz Brosa (Ikerlan), Idoia de la Iglesia (Ikerlan), Sébastien Dupont (CETIC), Joan Iglesias (ACISA), Tomasz Korniluk (Phoenix), Johan Kristiansson (RISE), Antonio Lalaguna (ACISA), Ignacio M. Llorente (OpenNebula), Marco Mancini (OpenNebula), Alberto P. Martí (OpenNebula), Philippe Massonet (CETIC), Nikolaos Matskanis (CETIC), Daniel Olsson (RISE), Michał Opala (OpenNebula), Per-Olov Östberg (UMU), Goiuri Peralta (Ikerlan), Samuel Pérez (Ikerlan), Bruno Rodríguez (OpenNebula), Juan José Ruiz (ACISA), Kaja Swat (Phoenix), Thomas Ohlson Timoudas (RISE), Iván Valdés (Ikerlan), Constantino Vázquez (OpenNebula).
Status:	Submitted

Document History

Version	Issue Date	Status[1]	Content and changes
0.1	20/10/2023	Draft	Initial Draft
0.2	27/10/2023	Peer-Reviewed	Reviewed Draft
1.0	31/10/2023	Submitted	Final Version

Peer Review History

Version	Peer Review Date	Reviewed By
0.1	27/10/2023	Erik Elmroth (UMU)
0.1	27/10/2023	Marco Mancini (OpenNebula)

Summary of Changes from Previous Versions

First Version of Deliverable D4.6

[1] A deliverable can be in one of these stages: Draft, Peer-Reviewed, Submitted, and Approved.

Executive Summary

This is the first version of Deliverable D4.6, theCOGNIT Serverless Platform Software Source report, produced in WP4 “AI-enabled Distributed Serverless Platform and Workload Orchestration”. It provides a short description, licence, version, code repository and user guide, as well as design, testing, and verification reference of each of the software requirements that have had active development tasks during the First Research & Innovation Cycle (M4-M9) in connection with these main components of the COGNIT Framework:

Cloud-Edge Manager

- **SR4.3** Serverless Runtime Deployment:

  Deploy Serverless Runtime as Virtualized Workloads (e.g. Containers or VMs/microVMs) on the cloud-edge infrastructure.

- **SR4.4** Metrics, Monitoring, Auditing:

  Edge-Clusters monitoring, Serverless Runtimes metrics collection and continuous security assessment.

AI-Enabled Orchestrator

- **SR5.2** Smart Deployment of Serverless Runtimes:

  Implement a Smart Workload Orchestrator (SWO) that exposes a REST API used by the Cloud-Edge Manager for requesting the deployment plans used for provisioning the Serverless Runtimes.

- **[NEW] SR5.3** Scheduling Mechanisms:

  Implement a scheduler that will place the Serverless Runtimes on the Edge-Clusters resources according to the deployment plan provided by the AI-Enabled Orchestrator.

This deliverable has been released at the end of the First Research & Innovation Cycle (M9), and will be updated with incremental releases at the end of each research and innovation cycle (i.e. M15, M21, M27, M33).

Table of Contents

Abbreviations and Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
AWS	Amazon Web Services
DaaS	Data as a Service
DB	Database
FaaS	Function as a Service
GPU	Graphics Processing Unit
HTTP	Hypertext Transfer Protocol
IAM	Identity and Access Management system
IP	Internet Protocol
IoT	Internet of Things
JSON	Javascript Object Notation
ML	Machine Learning
OS	Operating System
QoS	Quality of Service
REST	Representational State Transfer
S3	Simple Storage Service
SDK	Software Development Kit
SLA	Service Level Agreement
SQL	Structured Query Language
VM	Virtual Machine
YAML	Yaml Ain’t a markup language

1. Cloud-Edge Manager

SR4.3 Serverless Runtime Deployment	
Description	The Serverless Runtime is the main management unit of the COGNIT Framework. It is defined by a document (a JSON file) that conveys all the information for its automatic deployment on the distributed cloud-edge continuum. The document containing the requirements is sent by the Device Client to the Provisioning Engine that communicates with the Cloud-Edge Manager. At the Cloud-Edge Manager level, Serverless Runtimes are managed by the OpenNebula OneFlow[2] component, which allows it to handle both FaaS and DaaS as a single entity.
Licence	Apache 2.0
Version	OpenNebula 6.6.3
Design	D4.1 → [SR4.3] Serverless Runtime Deployment
Code	Public Repository
User Guide	Repository README
Testing	D5.2 → 10.4 Cloud-Edge Manager
Verification	D5.2 → 10.4 Cloud-Edge Manager

SR4.4 Metrics, Monitoring, Auditing	
Description	Metrics collected by monitoring systems provide valuable information on the operational efficiency, resource utilisation and sustainability of data centres and severless environments. The OpenNebula monitoring system(link) includes several metrics related to each compute node involved in the operations managed by OpenNebula, including the monitoring of OpenNebula itself.
Licence	Apache 2.0
Version	e47dbbbb
Design	D4.1 → [SR4.4] Metrics, Monitoring, Auditing
Code	Public Repository

[2] https://docs.opennebula.io/6.8/management_and_operations/multivm_service_management/index.html

Testing	D5.2 → 10.4 Cloud-Edge Manager
Verification	D5.2 → 10.4 Cloud-Edge Manager

2. AI-Enabled Orchestrator

SR5.2 Smart Deployment of Serverless Runtimes	
Description	An improved FFD (First-Fit Decreasing) algorithm has been developed to prioritise the placement recommendation to the Cloud-Edge Manager. The placement algorithm is encapsulated into FastAPI and deployed using a Docker container.
Licence	Apache 2.0
Design	D4.1 → [SR5.2] Smart Deployment of Serverless Runtimes
Code	Public Repository
User Guide	Repository README
Testing	D5.2 → 10.5 AI-Enabled Orchestrator
Verification	D5.2 → 10.5 AI-Enabled Orchestrator

SR5.3 Scheduling Mechanisms	
Description	The default OpenNebula scheduler implements a matchmaking algorithm that assigns VMs in pending state to a suitable hypervisor host (node). A first step involves filtering the available nodes to remove those that are not suitable for a particular VM (this can happen for multiple reasons, for instance not enough capacity in terms of CPU and/or memory). In a second step, these shortlisted nodes are ordered by priority, and the one with higher priority is chosen to deploy the VM on it.
Licence	Apache 2.0
Version	OpenNebula 6.8.0
Design	D4.1 → [SR5.3] Scheduling Mechanism for Smart Deployment of Serverless Runtimes
Code	Public Repository
User Guide	OpenNebula official documentation
Testing	D5.2 → 10.5 AI-Enabled Orchestrator
Verification	D5.2 → 10.5 AI-Enabled Orchestrator