ENI ISG - PoC Proposal

A.1 PoC Project Details

A.1.1 PoC Project

PoC Number (assigned by ETSI):

PoC Project Name: Intent-based Cloud Management

PoC Project Host: NTT

Short Description: This PoC will provide an Intent-Based Cloud Management (IBCM) solution that assists the cloud provider with decision making concerning cloud computing resources to meet the high-level cloud performance goal, i.e. the intent. In the PoC, we will demonstrate abstracting knowledge(building AI models) from cloud telemetry data, and making decisions of necessary cloud computing resources that meets the cloud performance goal using the knowledge (the models). Consequently, reduction of OPEX including the human resource cost, time cost and cloud resource cost can be expected by autonomizing the cloud resource decision process by the IBCM. In addition, the PoC will be designed to comply a sets of the use case specified in ENI 001, goals specified in ENI 002, intent translation architecture specified in ENI 008 and system architecture specified in ENI 005.

A.1.2 PoC Team Members

Table A.1

	Organization name	ISG ENI participant (yes/no)	Contact (Email)	PoC Point of Contact (see note 1)	Role (see note 2)	PoC Components
1	NTT Corporation	yes	Chao.wu.ex@hco.ntt.co.jp		Network operator/service provider	Design the PoC including the objective, architecture, etc., identify use cases Work with other member to develop the PoC
2	Intel	yes	emma.collins@intel.com haining.wang@intel.com		Manufacturer /infrastructure provider	Provide cloud hardware features, as well as platform telemetry solution, which are enablers for data collection and resource allocation.
3	NTT- AT	no	takaaki.tanaka@ntt-at.co.jp		Application provider	Work together with NTT Lab to help with implementation and optimization of AI algorithm, testbed setup and implementation of the demo
4	Intracom Telecom	no	nanastop@intracom-telecom.com		Application provider	Design AI based Cloud resource allocation solution.

NOTE 1: Identify the PoC Point of Contact with an X.
NOTE 2: The Role will be network operator/service provider, infrastructure provider, application provider or other as given in the Definitions of ETSI Classes of membership.

All the PoC Team members listed above declare that the information in this proposal is conformant to their plans at this date and commit to inform ETSI timely in case of changes in the PoC Team, scope or timeline.

A.1.3 PoC Project Scope

A.1.3.1 PoC Goals

- PoC Project Goal #1: Identify representative scenarios of intent-based cloud management.

- PoC Project Goal #2: Demonstrate intent-based cloud management's application in one or multiple identified scenarios, i.e., demonstrate the use of intent and AI algorithms to enable autonomous cloud resource decision according to the cloud performance goal (intent), and assess its for cloud service OPEX reduction.

- PoC Project Goal #3: Demonstrate telemetry collection for intent-based cloud management. Demonstrate the collection of multi-layer telemetry data from the cloud environment and using these telemetry data to train AI models for intent-based cloud management.

A.1.3.2 PoC Topics

Table A.2

PoC Topic Description (see note)	Related WI	Expected Contribution	Target Date
Network Operations > Automatic service and resource design framework for cloud service	GS ENI 001 Use Cases	Contribute to existing use case or new use case	March, 2022
	DGR/ENI-0023 ENI Intent Policy Model (GR ENI 013)	Provide contributions about Intent Policy Model related to cloud management scenarios	
	DGR/ENI-0025Processing and Management of Intent Policy (GR ENI 015)	Provide contributions about Intent Policy Processing and Management related to cloud management scenarios	
	Following-up work item of GR ENI 009 Data mechanisms if any	Provide contributions about data mechanism related to cloud management scenarios	

NOTE: This column should be filled according to the contents of table 1.

A.1.3.3 Other topics in scope

NA.

A.1.4 PoC Project Stages/Milestones

Table A.4

PoC Milestone	Stages/Milestone description	Target Date	Additional Info
P.S	PoC Project Start		June 2021, ENI #18
P.U	PoC user story		September 2021, ENI #19
P.D1	PoC Demo		December 2021, ENI#20
P.C	PoC Contribution		March 2022
P.R	PoC Report		March 2022
P.E	PoC Project End		June 2022

NOTE: Milestones need to be entered in chronological order.

A.1.5 Additional Details

For example, URL, planned publications, conferences, etc.

A.2 PoC Technical Details

A.2.1 PoC Overview

An increasing number of individuals, enterprises and organizations are utilizing cloud for conducting computing tasks, saving and sharing contents, hosting end user-facing services, etc.

Between the cloud user and the cloud provider, a gap has been existing that when requesting cloud services, the cloud user tell the cloud provider about their "intent" about cloud functional and non-functional goals, e.g., their goals about ability to handle workload, security level, and reliability level. For example, a cloud user from an e-commerce company will be concerned about whether the e-commerce service that runs on the cloud is able to handle 1000 transactions per second, rather than how these goals are met by the cloud resources and how the cloud resources is implemented. In contrast, cloud providers, when providing the cloud services to the cloud user, need to decide concrete cloud resource configurations, e.g., the amount of VMs, vCPU, and memory to be allocated to meet the cloud user's intent. In other words, the cloud provider needs to translate the user's intent ("what") into "how" the service is implemented.

There are currently two main approaches to translate the user's intent about the cloud service into "how" the service is implemented: the *cloud-consultant approach* and *self-service approach*. In the *cloud-consultant approach*, cloud users are supported by cloud consultants from the cloud provider, who collect users' goals and determine the resource details accordingly. However, this puts a huge workload on cloud providers, causes high OPEX and needs a relatively long lead time. In the *self-service approach*, cloud users are provided with a cloud resource management interface, and need to decide the necessary resource by themselves on the basis of their service goals. Thus, users need to be skilled in cloud issues in order to decide cloud resources. In both approaches, translation of service goals into resources heavily relies on human decisions.

The translation of intent into amount of cloud resource involves understanding both the cloud-based applications' workload patterns, and the underlying cloud resource configuration patterns, and most importantly, the complex relationship between the two. What is more, the increasing variety of service goals and high flexibility in cloud resource configuration add to the complexity of the decision-making process.

In this PoC, we provide the IBCM solution to address the challenges. Specially, we put our focus on how to translate the user's intent about performance (the ability to handle workload) to the amount of resource that meets these goal. We will mainly demonstrate the scenarios where IBCM decides the necessary cloud computing resource amount in accordance with the cloud user's intent about performance goal.

A.2.2 PoC Architecture

To support the cloud resource decisions, lower the barriers for the cloud users, and reduce the service delivery lead time and cost for cloud providers, we provide an IBCM solution in this PoC to realize autonomous cloud resource design in accordance to intent. The architecture of IBCM is shown in figure 1.

The IBCM system is a closed-loop system. Firstly, telemetry logs about the performance, etc., are collected from the cloud environment, and knowledge about the resource and performance causal relationship is abstracted from the logs. When cloud users input their cloud performance goal as intent, the knowledge is then used to make decision about the resource amount configuration for the intent. Finally, the resource configuration is implemented in the cloud environment and provided to the cloud users, and logs about the performance, etc. are collected.

IBCM has five main functional blocks (FB) as follows:

Cloud telemetry collection and processing

This FB collaborates with telemetry collection agents in the cloud environment to collect the cloud telemetry log data, such as cloud resource amount configurations, cloud application workload amount, cloud and application performance log data from various level of the cloud environment, and conducts combinations, normalizations, etc. of the log data.

Cloud performance knowledge abstraction

In the FB, models are trained on the basis of the collected log data. It trains models which infers the performance for the given the workload amount and the resource configuration information. The trained models are passed to the cloud performance inference FB.

Cloud performance inference

In this FB, first, the workload requirement, and configurable patterns of cloud resource amount are inputted to the generated models to infer the cloud performance. The pairs of inferred performance and the corresponding resource amount are passed to the intent-based cloud resource decision FB.

Intent-based cloud resource decision

This FB decides the resource amount that meets the intent on the basis of the performance inference results. For each pair of inferred performance and the corresponding resource amount, the FB examines whether the inferred performance satisfies the intent, if yes, the corresponding pattern of resource amounts is outputted as resource solutions.

Cloud resource configuration generation for the intent

In this FB, the selected resource amount is embedded in a resource orchestration template/command e.g. yaml and sent to resource orchestrators.

The recommend resource configuration is also sent back to the cloud provider for confirmation or manual adjustment if necessary. The provider is able to confirm/revise the resource decision and instruct launching of the VMs accordingly. In accordance with the resource decision, the resource orchestrators allocate the resources and activate the VMs, and the service is provided to the cloud user. The performance and system logs are collected and potentially used to update the models.

The IBCM framework can be mapped to the ENI reference architecture (figure 2). The cloud user/provider, cloud monitoring and management system are treated as Assisted and / or Governing users and systems in ENI system respectively. The cloud telemetry collection and processing of IBCM and the cloud resource configuration generation of IBCM play the roles of Input Processing and Normalization, Demoralization *and* Output generation of ENI system respectively. The cloud performance knowledge abstraction, cloud performance inference and intent-based cloud resource decision largely involve analysis process that is similar with ENI analysis parts, including the context awareness, knowledge management, situation awareness, policy management etc.

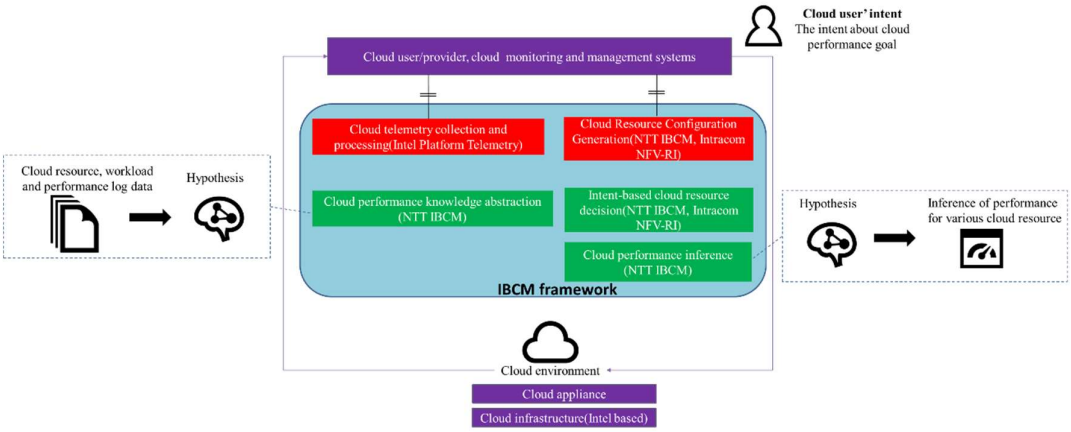

Figure 1. Overview of IBCM

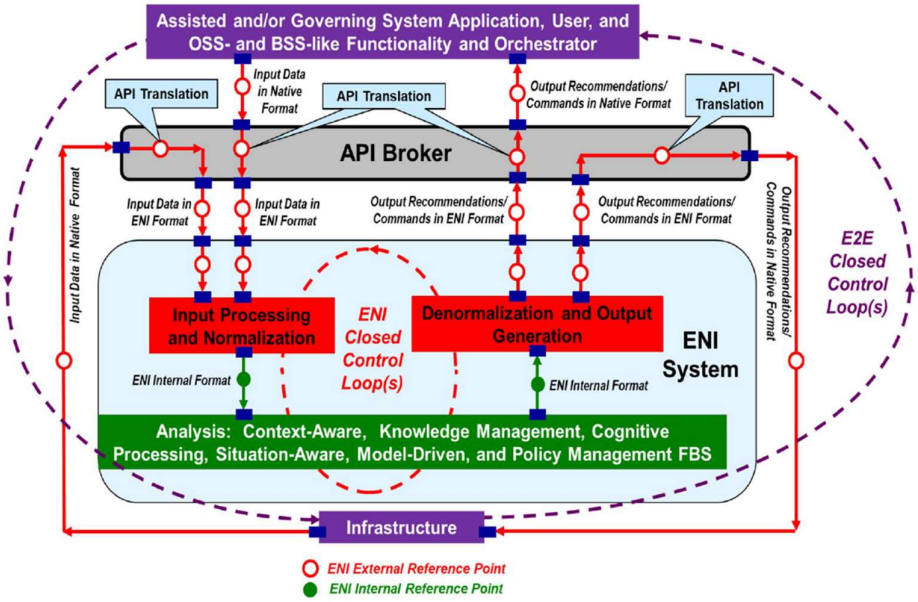

Figure 2. ENI system architecture

.

Figure2. ENI System Architecture

A.2.3 PoC Success Criteria

The PoC is considered success if the following criteria are meet

1. Demonstrate that IBCM is able to decide the resource amount in accordance to cloud user's intent

2. Qualitatively show the effect of reduction of cloud resource decision cost by applying IBCM.

A.2.4 Additional information

Include additional information as useful.