ID contributo: **5** Tipo: **non specificato**

AI-based approach for provider selection in the INDIGO PaaS Orchestration System of INFN Cloud

mercoledì 12 giugno 2024 09:15 (25 minuti)

INFN Cloud provides scientific communities supported by the Institute with a federated Cloud infrastructure and a dynamic portfolio of services based on the needs of the supported use cases. The federative middleware of INFN Cloud is based on the INDIGO PaaS orchestration system, consisting of interconnected open-source microservices. Among these, the INDIGO PaaS Orchestrator receives high-level deployment requests and coordinates the process of creating deployments on the IaaS platforms provided by federated providers.

In the default configuration, the INDIGO PaaS Orchestrator determines the provider to submit the deployment creation request to from an ordered list of providers, selection based on the user's group affiliation. This list is provided by the Cloud Provider Ranker service, which applies a ranking algorithm using a restricted set of metrics related to deployments and defined Service Level Agreements for providers. The INDIGO PaaS Orchestrator submits the deployment to the first provider in the list and, in case of failure, scales to the next provider until the list is exhausted.

This contribution presents the activity aimed at improving the ranking system and optimizing resource usage through an approach based on the use of artificial intelligence techniques. In this context, significant preparatory work was carried out to identify the most meaningful metrics, as well as the sources from which to retrieve these metrics. The subsequent dataset preparation allowed us to study the case in detail, identifying and comparing different artificial intelligence techniques. The proposed approach involves creating two models: one for predictive classification of deployment success/failure and one for deployment creation time regression. A linear combination of the output of the two models, along with training on recent and mobile time windows, allows for the definition of an ordered list of providers that the orchestrator can use for deployment submission.

Autore principale: GIOMMI, Luca (Istituto Nazionale di Fisica Nucleare)

Relatore: GIOMMI, Luca (Istituto Nazionale di Fisica Nucleare)

Classifica Sessioni: Wednesday morning: Part I