Metodi innovativi basati sul machine learning per la workload mobility.

DESCRIZIONE DEL PROGETTO

Le soluzioni allo stato dell'arte permettono di utilizzare le risorse di calcolo disponibili sui vari dispositivi, quali smart gateway, edge-private-public cloud, seguendo però un approccio statico. Tali soluzioni associano staticamente, ad ogni tipologia di task, i nodi della rete dedicati al relativo compito. Questo approccio non si adatta a contesti in cui ci possano essere vantaggi, dal punto di vista dell'utilizzo delle risorse di calcolo o della riduzione dei tempi di esecuzione dei task, derivanti da una allocazione dinamica dei carichi di lavoro. A tal fine ci si prefigge di progettare ed implementare opportune tecniche di workload mobility in ambito smart gateway/edge/cloud, permettendo la definizione di logiche di allocazione/migrazione dei servizi e supportando un relativo sistema di deployment automatico. Più precisamente ci si prefigge di utilizzare tecniche di deployment tramite container technologies, quali Docker e Kubernetes, per realizzare architetture altamente modulari e dinamicamente ristrutturabili. Le tecniche proposte si basano su architecture-level reconfiguration tramite container orchestration plans in ambito di architetture con scalabilità fine-grained, come quelle a microservizi. Tali plan vengono generati tramite monitoraggio e predizione del workload. Quest'ultima è resa possibile dall'utilizzo di tecniche di machine learning che stabiliscono, sulla base di dati relativi all'attività passata, quali componenti del sistema sono soggetti a elevato workload e in quali periodi temporali.

PIANO DI ATTIVITA'

Le attività saranno svolte nel contesto del progetto SeaWall, derivante da attività commissionata finanziata dal consorzio BiRex.
In particolare, il piano di attività include:

- Analisi dello stato dell'arte di: soluzioni tecnologiche a supporto della migrazione di servizi software tramite architecture-level reconfiguration, finalizzati all'implementazione di scenari di edge/cloud continuum; e di tecniche di machine learning per l'analisi dell'evoluzione del workload in ambito di sistemi a componenti.
- Progettazione e sviluppo di algoritmi per la dynamic architecture-level reconfiguration tramite riallocazione/replicazione dei servizi, in scenari di edge/cloud continuum. In particolare algoritmi basati sul mantenimento di vincoli di Qualità del Servizio (QoS) delle applicazioni e sul monitoraggio del carico computazionale dei nodi/prestazioni di rete.
- Utilizzo di tecniche di machine learning per la predizione delle variazioni del workload sulla base di dati relativi all'attività passata: quali componenti del sistema sono soggetti a elevato workload e in quali periodi temporali.