Progetto e Piano di Formazione

Definizione e sviluppo di una piattaforma per l'accelerazione trasparente di applicazioni intelligenti e *time-sensitive* nel cloud continuum

Progetto

I recenti progressi nell'ambito dell'Intelligenza Artificiale, in particolare rispetto alle capacità dei modelli di *Machine Learning (ML)*, stanno portando alla progressiva adozione di logica intelligente in qualunque dominio applicativo, inclusi ambiti quali l'automazione industriale, le moderne reti di telecomunicazioni o l'industria veicolare. Tuttavia, la necessità di avere cicli di controllo con rigidi vincoli di prestazioni (es. risposte a eventi esterni in tempi quasi immediati) - *continuum*, che prevede la prossimità fisica di risorse di calcolo relativamente potenti alle sorgenti degli eventi monitorati, rappresenta una promettente direzione in supporto ai requisiti di queste applicazioni, ma non è di per sé sufficiente a garantire il rispetto dei vincoli più stringenti.

L'attività di ricerca di questo progetto prevede che l'assegnista si concentri sulle più moderne tecnologie di comunicazione e sulle più recenti piattaforme disponibili allo stato dell'arte per consentire l'accelerazione dei tempi di risposta di applicazioni nel cloud continuum, consentendo così l'impiego di logica intelligente senza metterne a rischio la funzionalità. In particolare, l'assegnista dovrà occuparsi delle seguenti tematiche:
i) analisi e valutazione delle moderne tecnologie di accelerazione dell'I/O, con particolare attenzione a tecniche di *kernel bypassing* (es. RDMA, DPDK, SPDK) e innovativi supporti hardware (es. GPU, DPU, SmartNIC, NVMe) in grado di minimizzare la variabilità e i ritardi associati alla comunicazione e all'archiviazione dei dati sotto stretti vincoli temporali.
ii) analisi e valutazione dello stack di protocolli necessario a utilizzare tali tecnologie e della loro compatibilità con strumenti e tecniche standard, allo scopo di minimizzare i cambiamenti necessari alle applicazioni esistenti per utilizzare in modo efficace tali tecnologie.
iii) analisi e valutazione della fattibilità tecnica dell'integrazione di queste tecnologie con gli strumenti standard di virtualizzazione e isolamento (es. VMs, containers) e con gli orchestratori (es. Kubernetes) correntemente utilizzati nell'ambito del cloud continuum.
iv) analisi e valutazione delle piattaforme esistenti in ambito cloud continuum per lo sviluppo, il collocamento e la manutenzione di applicazioni per il controllo di eventi esterni, allo scopo di definirne i limiti e di individuare possibilità di integrazione con le tecnologie di accelerazione precedentemente individuate. Particolare attenzione dovrà essere rivolta a piattaforme *serverless*, che attualmente sono considerate le più adatte a supportare questo tipo di requisiti.
v) definizione e implementazione di una architettura che integri le tecnologie di accelerazione precedentemente individuate e (1) offra un'unica interfaccia di accesso, semplice da utilizzare e sufficientemente flessibile da adattarsi a diversi casi applicativi, e (2) possa dinamicamente adattarsi ai requisiti di prestazioni, utilizzo di risorse e ambienti di esecuzione del cloud continuum.
vi) identificazione di casi d'uso reali per la valutazione dell'efficacia dell'architettura, considerando come metriche principali (1) le prestazioni, (2) l'utilizzo di risorse, (3) la possibilità di integrare logica intelligente nei cicli di controllo senza comprometterne le prestazioni.

Piano di Formazione

Il piano di formazione associato alle attività dell'assegno di ricerca prevede le seguenti attività organizzate nel periodo di un anno.

I semestre

Identificazione di casi d'uso rilevanti, preferibilmente in ambito industria, 5G, veicolare, e definizione dei requisiti di prestazioni, QoS, consumo di risorse e grado di intelligenza desiderato di tipiche applicazioni in questi domini.

Analisi e valutazione dello stato dell'arte nel settore dell'accelerazione dell'I/O per applicazioni distribuite nel cloud continuum, sia da un punto di vista delle tecnologie (hardware e software) correntemente utilizzate, che delle piattaforme utilizzate allo stato dell'arte per lo sviluppo e la gestione di tali applicazioni.

Valutazione sperimentale di diverse potenziali tecnologie (hardware e software) e piattaforme che possano consentire il rispetto dei requisiti nei casi d'uso individuati; discussione di tali risultati e individuazione delle possibili direzioni di integrazione.

Analisi delle possibilità di integrazione delle tecnologie individuate in ambienti cloud, caratterizzati dalla necessità di virtualizzazione e isolamento delle applicazioni.

II semestre

Definizione di una architettura per l'integrazione delle tecnologie individuate in un modello di piattaforma esistente, con preferenza per il paradigma *serverless*; definizione di principi di funzionamento comuni che consentano alle applicazioni di attare le specifiche tecnologie ai propri requisiti di QoS e alle condizioni dinamicamente individuate nel cloud continuum.

Implementazione dell'architettura in modo incrementale, attraverso una prototipazione e valutazione sperimentale.

Valutazione del risultato ottenuto in almeno in uno dei casi d'uso precedentemente individuati.

Promozione della soluzione individuata presso potenziali stakeholder in modo allargato.

Discussione dei vantaggi/svantaggi dell'impiego delle tecnologie di accelerazione individuate, allo scopo di valutare se esse possano effettivamente portare vantaggio ad applicazioni *intelligenti* distribuite nel cloud continuum con stringenti vincoli di prestazioni.