ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXVI Cycle

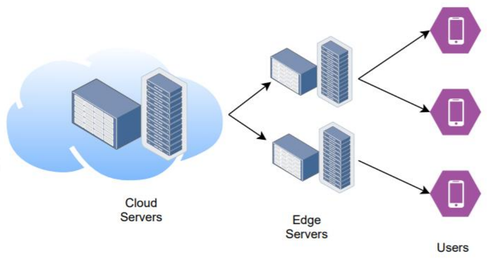
Reducing Interference and Deployment Time of Microservices at the Edge

Madhura Adeppady

Supervisor: Prof. Carla Fabiana Chiasserini, Prof. Paolo Giaccone

Research context and motivation

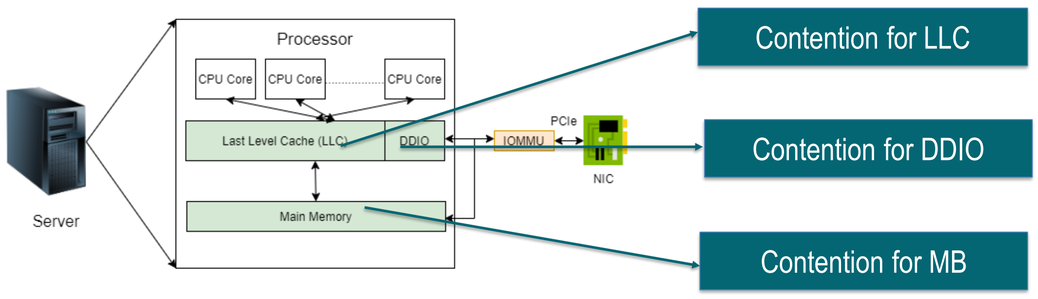
- **Edge computing** enables offloading of service tasks from either mobile devices or the core network to the edge
  - ❑ Reduces end-to-end latency and network traffic
  - ❑ Edge nodes have limited resources
- **Microservices (MSs)** implementing mobile services deployed on edge nodes
  - ❑ Using **containers**
  - ❑ **Consolidation** of multiple containers on same hardware
  - ❑ Containers run on dedicated cores



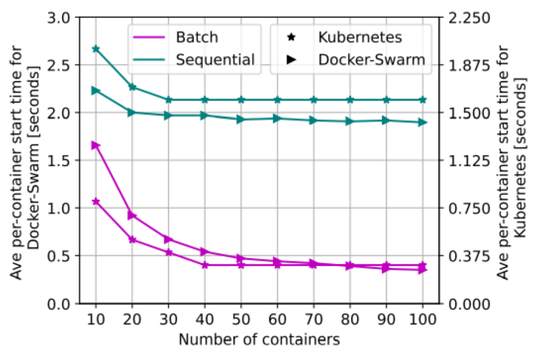
- **Orchestrators** receive multiple requests for service deployments simultaneously
  - ❑ Need for **batch deployment** of MSs
  - ❑ Edge services demand **faster deployment time**
- MS placement problem studied in literature ignored
  - ❑ **MS deployment time** incurred while using real-world orchestrators
  - ❑ **Performance interference** experienced by the consolidated MSs running on the same edge node

Addressed research questions/problems

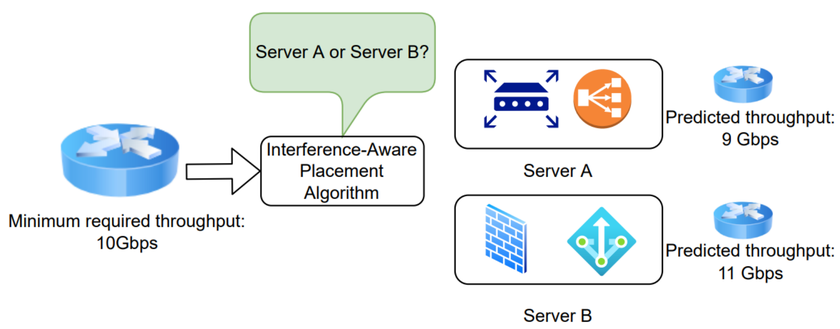
- MSs running on the same hardware share and compete for memory subsystem resources
  - ❑ Results in **performance interference**



- Current orchestrators deploy MSs sequentially
  - ❑ Sequential deployment has 80% more deployment time than batch deployment

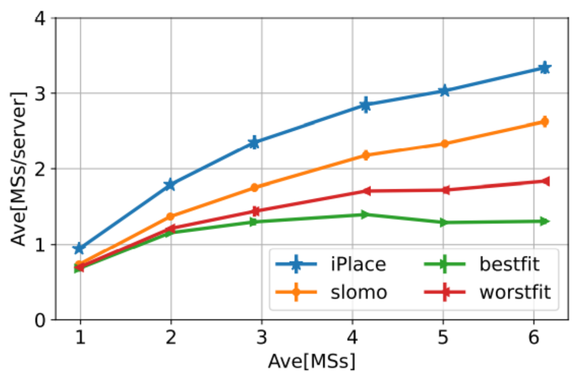


- Addressed **Interference-aware MS Placement (IMSP) problem** to minimize the number of used servers for placement

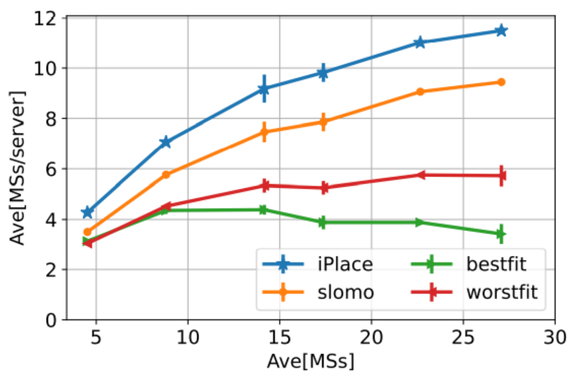


Novel contributions

- Built a machine learning model for predicting MS performance
- Formulated IMSP as an optimization problem minimizing the number of servers needed to place MSs
- Proposed a low complexity heuristic **'iPlace'** based on MS clustering
- Extensive simulation results show that,
  - ❑ iPlace uses **lower number of servers** to place the requests
  - ❑ **Number of consolidated MSs per node** is higher in iPlace



Low load scenario

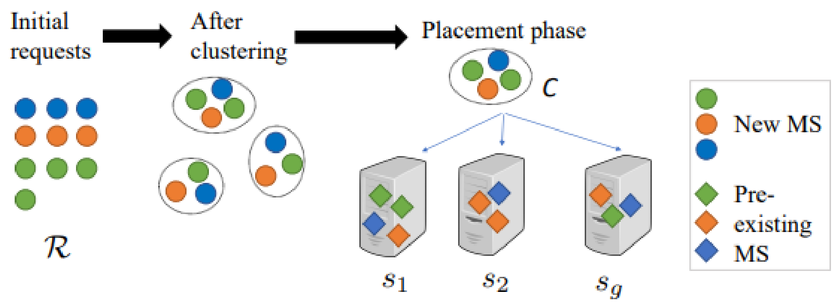


High load scenario

- Clustering approach followed in iPlace reduces **per-container deployment time** compared to the benchmarks.

Adopted methodologies

- Prediction model using Gradient Boosting Regressor built to predict target MS performance using,
  - ❑ **Contentiousness Vector:** Consists of various system level metrics
  - ❑ **Sensitivity Model:** Models target MS performance as a function of its contentiousness
- iPlace algorithm works in two phases,
  - ❑ A k-means based **clustering approach:** Clustering done using contentiousness vector of MSs
  - ❑ Iterative **placement phase** to deploy each created cluster: Minimizes interference effect among MSs running on same hardware



Future work

- MSs still can suffer from high start-up latency
  - ❑ Handling a service request requires creating a container, downloading and installing necessary libraries before starting
  - ❑ Short-lived services suffer from this high start-up latency
- Propose a solution to reduce the **start-up latency** by utilizing various container states (pause, pre-warm, warm, etc.) at the edge nodes

Submitted and published works

- M. Adeppady, C. F. Chiasserini, H. Karl, P. Giaccone, "iPlace: An Interference-aware Clustering Algorithm for Microservice Placement," ICC 2022 International Conference on Communications, 2022, pp. 5457-5462
- M. Adeppady, C. F. Chiasserini, H. Karl, P. Giaccone, "Reducing Interference and Deployment of Microservices at the Edge," submitted to IEEE Transactions on Network Service Management
- M. Adeppay, C.F Chiasserini, P. Giaccone, "Building Dataset for Predicting VNF Interference," Meditcom 2021 special session on SEMANTIC
- M.Adeppady, P. Giaccone, A. Conte, H. Carl, C. F. Chiasserini, "Efficient Container Retention Strategies for Serverless Edge Computing," submitted to IEEE CAMAD special session SEMANTIC 2022

List of attended classes

- 01QTEIU – Data mining concepts and algorithms (1/2/2021, credits: 4)
- 01DTPRV – Connected Vehicles (didattica di eccellenza) (23/6/2022, credits: 4)
- 02SFURV – Programmazione scientifica avanzata in matlab (25/5/2021, credits: 6)
- 01DNBIU – Security of next generation networks (18/7/2022,4)
- Summer School - Machine learning, sustainable edge computing, and networking (15/7/2022, credits: 5)
- Research Integrity (20/9/2021, credits: 1)
- SEMANTIC ITN Training Activities (Total hours: 72)