UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Optimal placement of micro-services chains in a Fog infrastructure

Claudia Canali
DIEF, University of Modena and Reggio Emilia

Giuseppe Di Modica
DISI, University of Bologna

Riccardo Lancellotti
DIEF, University of Modena and Reggio Emilia

Domenico Scotece
DISI, University of Bologna

Motivation

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

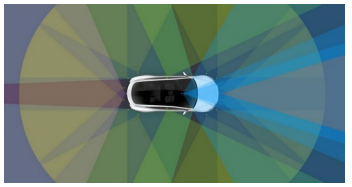
"... fog computing is a horizontal system-level architecture that distributes computing, storage, control, and networking functions closer to users along a cloud-to-things continuum"

- Fog can place elaboration tasks close to data sources
  - Data filtering / dimensionality reduction
  - Latency-sensitive tasks
- New challenge
  - Mapping services to fog nodes
  - Distributed nature of fog

OpenFog™

- Some examples:
  - Crowd-sourcing applications (data)

  - Support for autonomous driving (latency)



- Cloud not flexible enough

Contribution

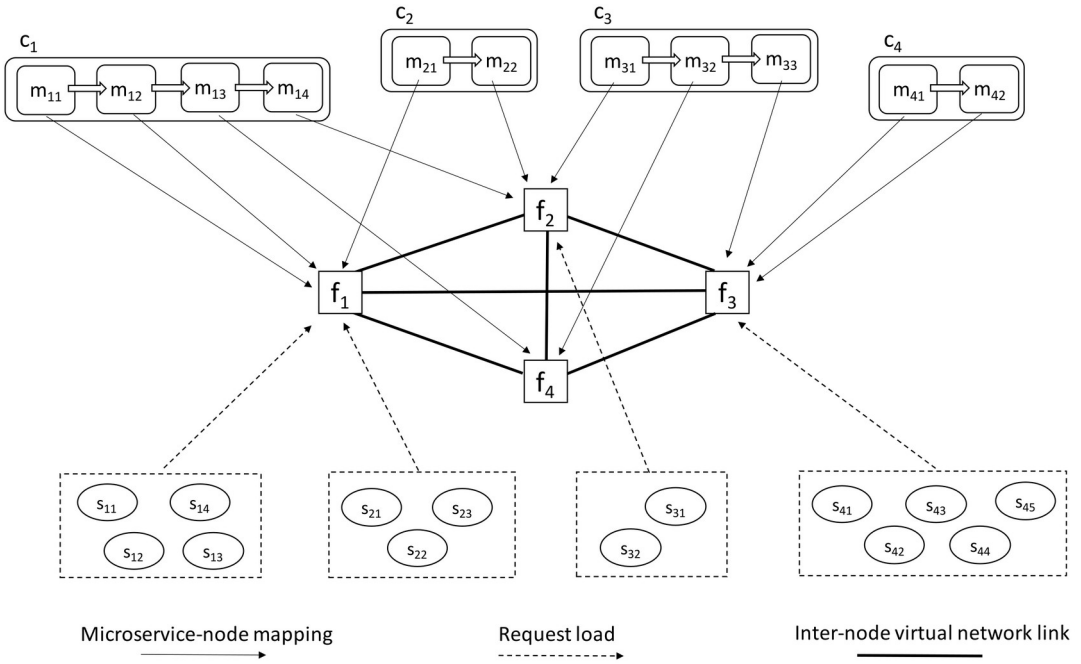
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

- Performance model
  - Applications composed of multiple micro-services
  - Micro-services of same applications on different nodes
  - Micro-services of different applications on same node
- Optimization problem
- Heuristic
  - Based on Genetic Algorithms
- Experimental results
  - Randomly generated problems
  - Sensitivity analyses to problem characteristics

Reference architecture

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

- Service chains
  - Applications
- Micro-services
  - Components of chains
- Fog nodes
  - Host micro-services
- Sensors
  - Input for service chains
  - First micro-service of the chain
- Placement problem
  - Micro-service → fog nodes



Performance model

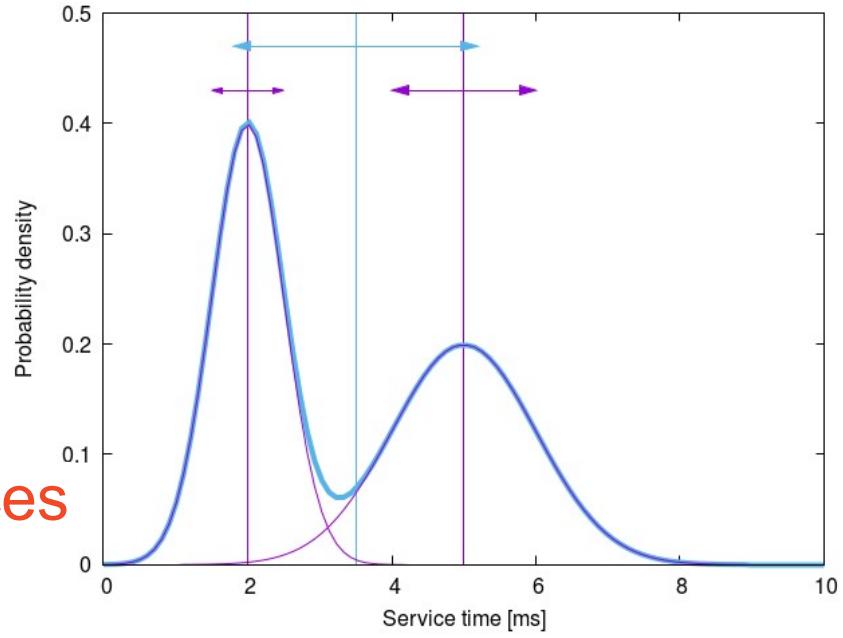
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

- Multiple micro-services on same Fog node
  - Assuming Gaussian distribution for micro-service execution
  - Service time of node: Mixture of Gaussians
  - Can compute average and standard deviation

$$S_f = \frac{1}{P_f} \cdot \sum_{m \in \mathcal{M}} x_{m,f} \frac{\lambda_m}{\lambda_f} S_m$$

$$\sigma_f^2 = \left(\frac{1}{P_f^2} \cdot \sum_{m \in \mathcal{M}} x_{m,f} \frac{\lambda_m}{\lambda_f} (S_m^2 + \sigma_m^2) \right) - S_f^2$$

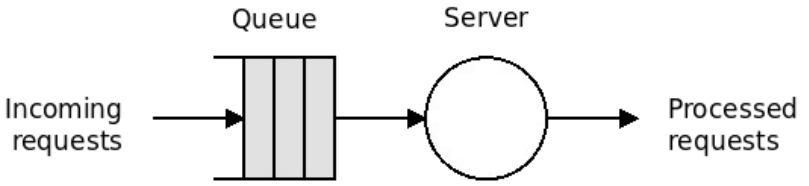
- Variance of service time
  - Co-location of different micro-services
  - Heterogeneous fog nodes (P_f)



Performance model

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

- Must take into account high variance
  - M/G/1 model for Fog nodes
  - Response time based on Pollaczek Kinchin formula



$$R_f = S_f + \frac{S_f^2 + \sigma_f^2}{2} \cdot \frac{\lambda_f}{1 - \lambda_f S_f}$$

- Response time of service chain must include:
  - Response time of micro-services (depends on fog nodes)
  - Network delays (consecutive services on different nodes)

$$R_c = \sum_{m \in c} x_{m,f} \cdot R_f + \sum_{f_1, f_2 \in \mathcal{F}} \sum_{m_1, m_2 \in c} o_{m_1,m_2} \cdot x_{m_1,f_1} \cdot x_{m_2,f_2} \cdot \delta_{f_1,f_2}$$

Optimization problem

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

- Goal
  - Minimize average response times of service chains
- Decision variable:
  - Placement of micro-service on fog nodes $x_{m,f}$
- Constraints
  - Services on exactly one node
  - Avoid overload on every fog node
  - SLA for service chains
  - Boolean decision variable

$$\min obj(X) = \sum_{c \in \mathcal{C}} w_c R_c$$

subject to:

$$\sum_{f \in \mathcal{F}} x_{m,f} = 1 \quad \forall m \in \mathcal{M},$$

$$\lambda_f < \frac{1}{S_f} \quad \forall f \in \mathcal{F},$$

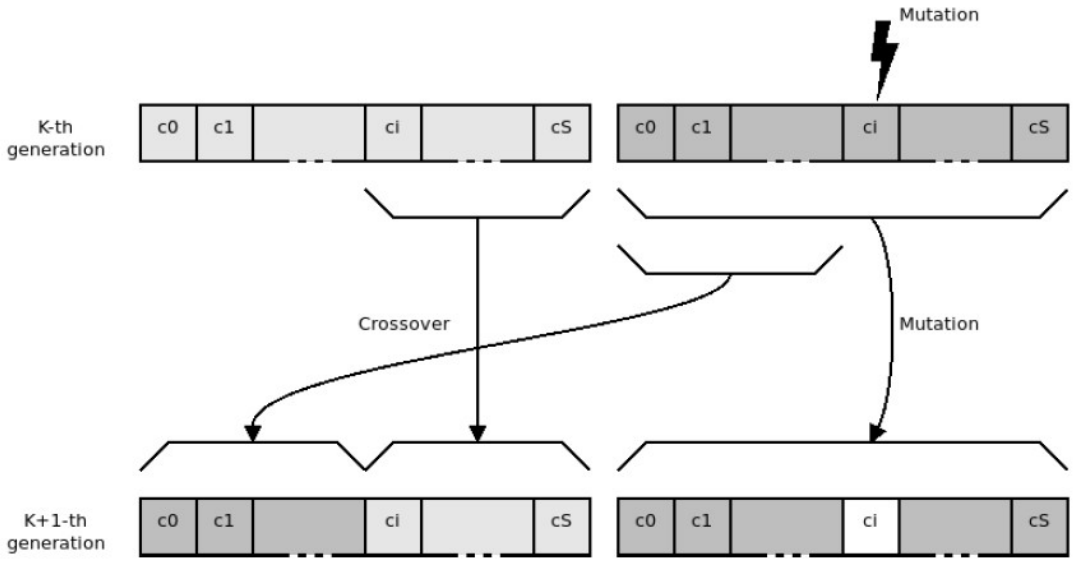
$$R_c < T_c^{SLA} \quad \forall c \in \mathcal{C},$$

$$x_{m,f} = \{0, 1\}, \quad \forall m \in \mathcal{M}, f \in \mathcal{F},$$

Genetic algorithm

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

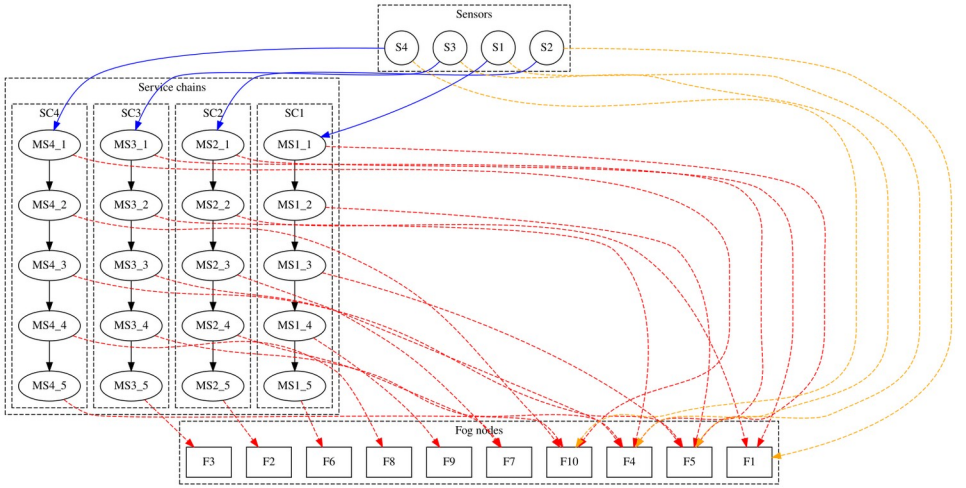
- Coding of a problem solution into a string (chromosome)
  - i-th gene in a chromosome
  - Value in range [1..|F|]
  - Placement of i-th micro-service
- Population of individuals evolving through generations
  - Fitness → Obj function
- Genetic operators
  - Selection (tournament)
  - Mutation (random)
  - Crossover (uniform)



Experimental setup

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

- Generation of multiple random problems
  - Service chain length L_c
  - Service time of chain S_c
  - Average network delay δ
  - Number of nodes $|F|$ and of service chains $|C|=0.4\times|F|$
  - Average system utilization $\rho=60\%$; $T_{SLA}=10\times S_c$
- Problems solved with GA
  - Sensitivity to problem parameters

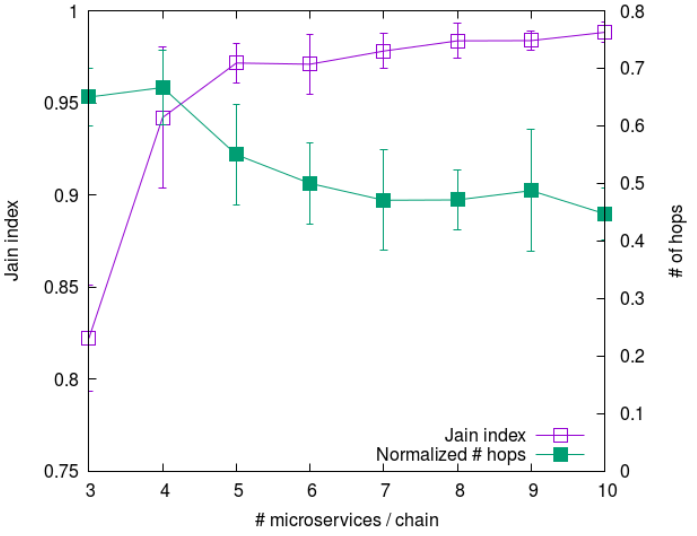


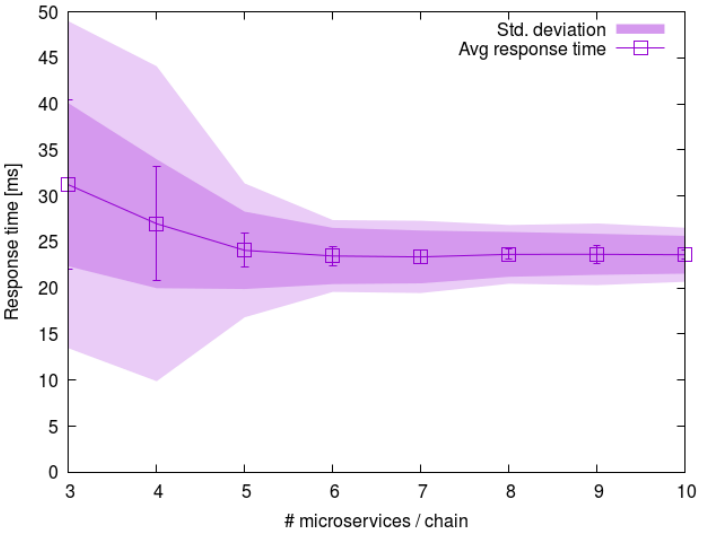
- Main metrics:
  - Jain index (load balancing)
  - $Nhops/L_c$
  - Response time R_c
  - GA execution time

Sensitivity to chain length

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

- Varying number of micro-services in a chain
  - Long chains: many small services
  - Short chains: few heavy services
- Load balancing in short chains
  - Low Jain index
  - High variance in resp. time
    Both within and across problems
- Network delay for long chains
  - Remains limited
  - Obj function reduces # of hops

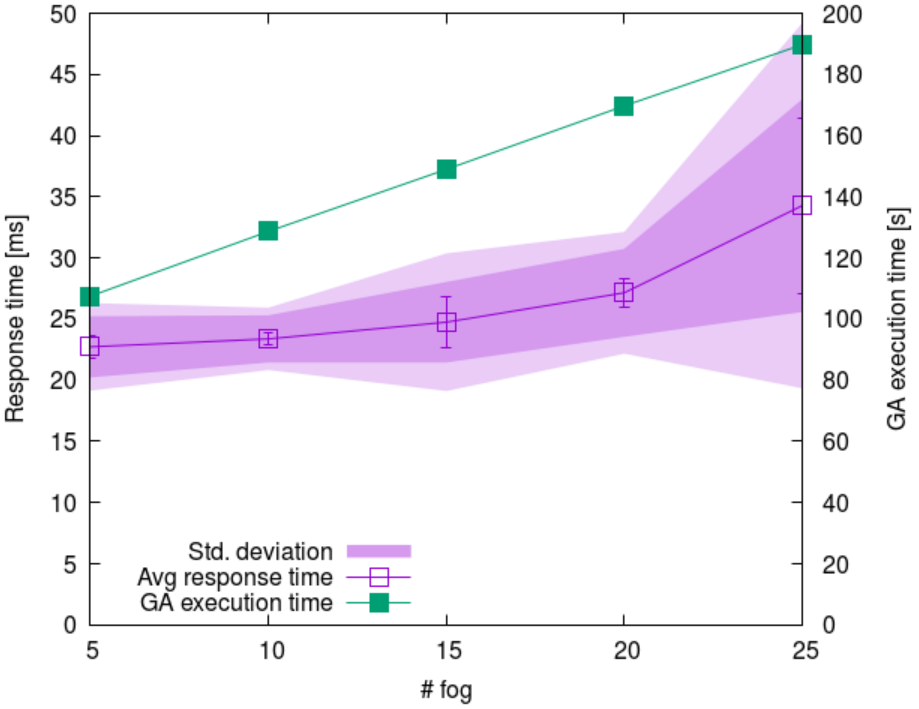




Scalability analysis

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

- Scalability analysis
  - Fog nodes 5 → 25
  - Micro-service 10 → 50, L_c=5
- Execution time
  - Grows linearly with problem size
    → *longer chromosome*
- Solution quality
  - Reduced for larger problem
  - Solution space too large
  - Limited impact of GA tuning



Conclusions and future work

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

- Applications composed of multiple micro-services in fog
  - Complex management of micro-services placement
- Proposal of a performance model
  - Heterogeneous fog nodes
  - Multiple services on same node/ App. on multiple nodes
- Optimization problem and Heuristic (based on GA)
  - Testing on synthetic problems
- Future work
  - Improve scalability (new heuristics, dynamic programming)
  - Testing in real application scenarios
  - More complex applications (DAG)

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Optimal placement of micro-services chains in a Fog infrastructure

Claudia Canali
DIEF, University of Modena and Reggio Emilia

Giuseppe Di Modica
DISI, University of Bologna

Riccardo Lancellotti
DIEF, University of Modena and Reggio Emilia

Domenico Scotece
DISI, University of Bologna