L’Evoluzione del Cloud verso l’Edge

Valeria Cardellini

Università degli Studi di Roma "Tor Vergata"

www.ce.uniroma2.it/~valeria

12° Workshop CIPA – Il Cloud nelle Banche, 16 marzo 2023

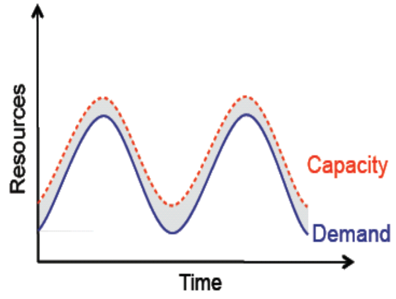
Cloud Benefits

- On-demand self-services
- Broad network access
- Resource pooling and virtualization
- Pay-per-use pricing model
- Rapid elasticity

Cloud Scalability and Elasticity

- Horizontal (scale-out/in) vs. vertical scaling (scale-up/down)

- Elasticity: system is able to adapt to workload changes by provisioning and de-provisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as clos



Data center in the cloud

⬜ Unused resources

The Cloud Evolution

- From a single data center
- To multiple geo-distributed data centers

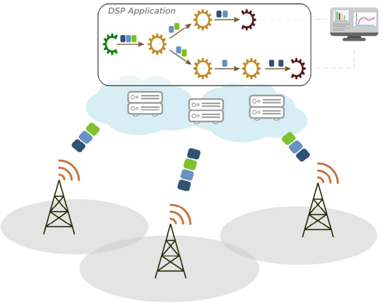
- From public/private Cloud
- To hybrid Cloud

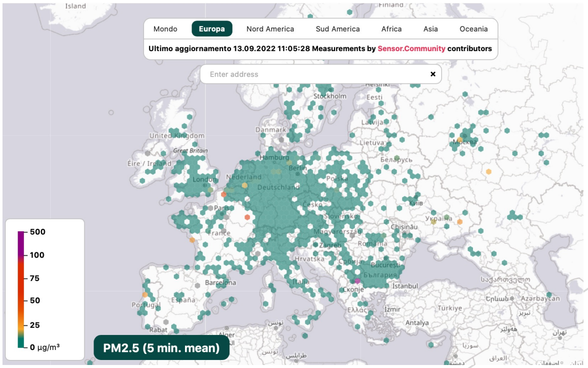
- From “long-term” pay-as-you-go resources and services
- To “short-term” pay-as-you-go resources and services with serverless computing

- From centralized infrastructures and services
- To **decentralized** infrastructures and services

The New Compute Continuum: from Cloud to Edge

- **Edge computing** as a strategic technology for Europe’s Digital Decade
  - Data - Edge & Cloud: 10,000 climate-neutral highly secure edge nodes
- Main benefits:
  - Reduce latency
  - Save energy
  - Bring AI and analytics where data are produced and consumed
  - Example: environmental data analysis, where processing can occur on edge resources

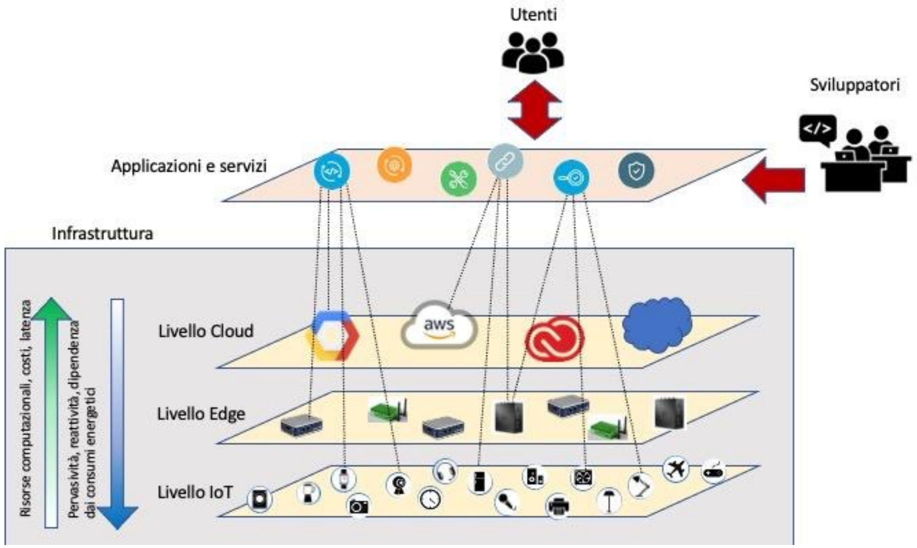




Ultimo aggiornamento 13.09.2022 11:05:28 Measurements by Sensor.Community contributors

The New Compute Continuum: from Cloud to Edge

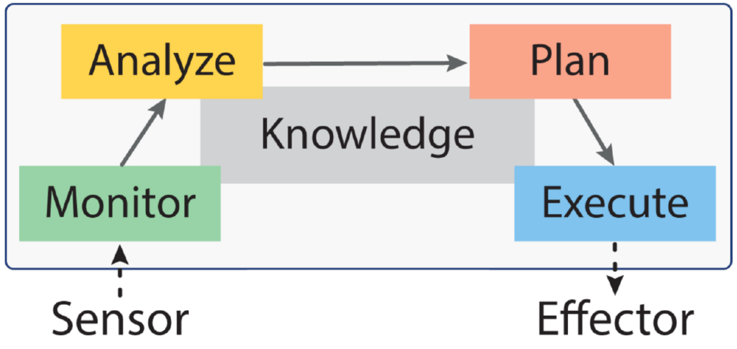
- But edge computing alone is not enough!
- A continuum of computing resources from the edge to the cloud


Utenti
Sviluppatori
Applicazioni e servizi
Infrastruttura
Risorse computazionali, costi, latenza
Pervasività, reattività, dipendenza dai consumi energetici
Livello Cloud
aws
Livello Edge
Livello IoT

Source: https://www.agendadigitale.eu/infrastrutture/compute-continuum-nuove-opportunita-di-calcolo-efficiente-e-pervasivo-sfide-e-vantaggi/

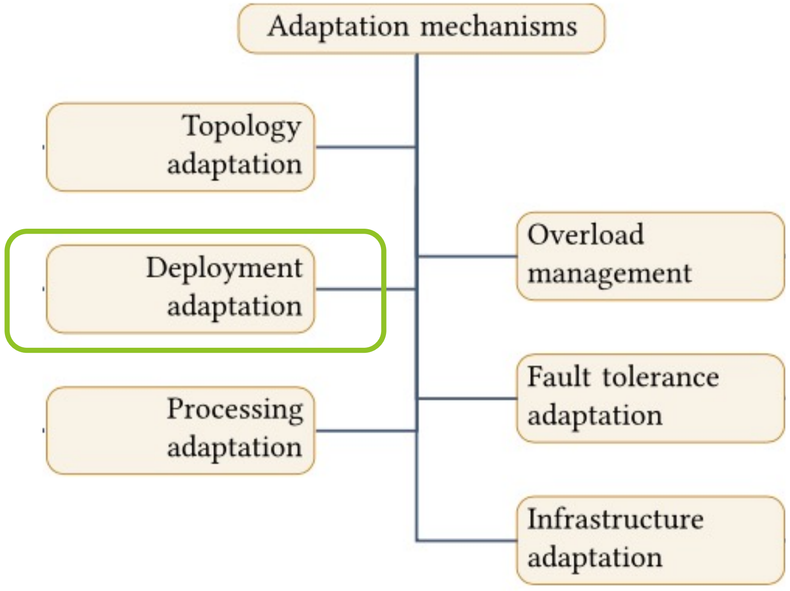
Challenges in the Compute Continuum

- Several sources of uncertainty in the Cloud-Edge continuum:
  - Unpredictable workloads
  - Unstable network conditions
  - Resource heterogeneity
  - Variable monetary costs
  - Security attacks
- How to cope with? Ability to **self-adapt** at runtime



Main Choices: Adaptation Mechanisms

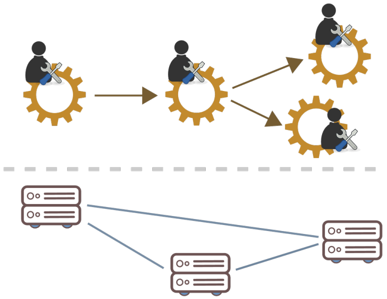
- Many adaptation mechanisms for Cloud-native apps, including

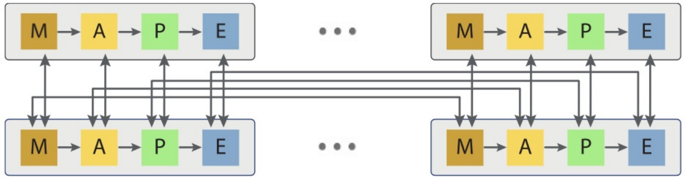


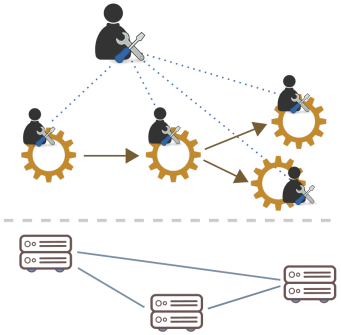
- Focus on deployment adaptation: **auto-scaling and placement**

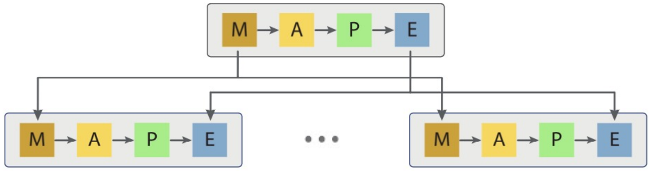
Main Choices: Adaptation Architectures

- Large-scale apps and environments: need to decentralize
- Fully decentralized vs. hierarchical control loops







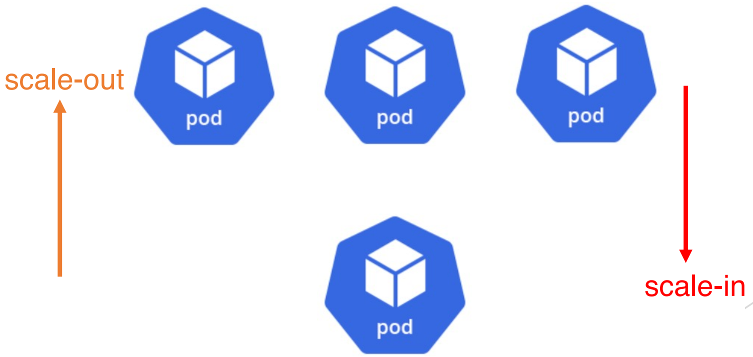


Main Choices: Adaptation Policies

- From simple heuristics control policies
- To more complex policies that exploit a variety of methodologies, among which:
  - Mathematical optimization
  - Control theory
  - Machine learning and reinforcement learning

Adaptation Policies: Example

- How Kubernetes controls auto-scaling
- Multiple auto-scalers at different control layers
  - Cluster auto-scaling with node granularity
  - Horizontal auto-scaling with pod granularity
  - Vertical auto-scaling with pod granularity
- Horizontal Pod Autoscaler (HPA)
  - Threshold-based policy
  - Scales number of pods according to ratio between observed value and target value



$$desiredReplicas = \left\lceil currentReplicas \frac{currentMetricValue}{desiredMetricValue} \right\rceil$$

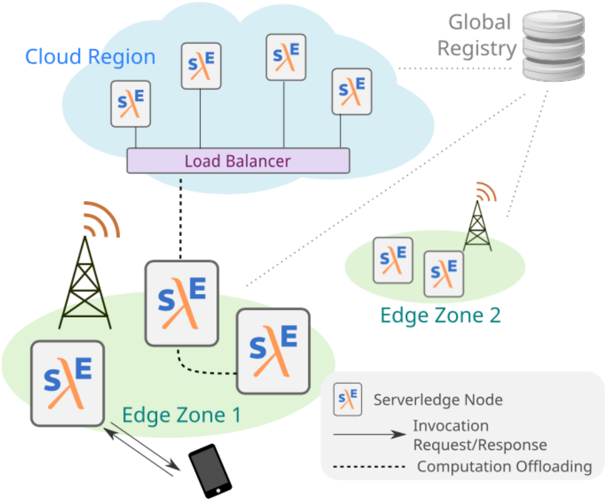
Adaptation Policies: Example

- HPA policy pros and cons
  - ✓ Simple and easy to understand policy: select metrics and thresholds
  - ✗ How to set thresholds values? Can be application-dependent
  - ✗ Not robust against varying load
- Alternative: use reinforcement learning to adapt threshold values at run-time
  - ✓ Improve application performance
  - ✓ Reduce resource wastage

F. Rossi, V. Cardellini, F. Lo Presti, M. Nardelli, "Dynamic multi-metric thresholds for scaling applications using reinforcement learning", *IEEE Transactions on Cloud Computing*, 2022.

Our Research Work at Rome Tor Vergata

- **Deploy** and **manage** at **runtime** distributed applications in the Cloud-Edge continuum satisfying Quality of Service (**QoS**) requirements
  - Which apps? Data stream processing, microservices, serverless
- E.g., Serverless in the Cloud-Edge Continuum
- Our solution: Serverledge, a new FaaS framework
  - Decentralized architecture
  - Horizontal and vertical offloading



G. Russo Russo, V. Cardellini, F. Lo Presti, T. Mannucci, "Serverledge: Decentralized Function-as-a-Service for the edge-cloud continuum", IEEE PerCom 2023.

Summing Up

- The new Compute Continuum opens up new challenges for academic and industry
- From seamless infrastructure and platform management
- To application design and run-time management