IoT-Cloud Service Optimization in Next Generation Smart Environments

IEEE Journal on Selected Areas in Communications

Marc Barcelo*, Alejandro Correa*, Jaime Llorca†, Antonia M. Tulino†‡, Jose Lopez Vicario*, Antoni Morell*
*Universidad Autonoma de Barcelona, Spain, Email: {marc.barcelo, alejandro.correa, jose.vicario, antoni.morell}@uab.cat, †Nokia Bell Labs, NJ, USA, Email: {jaime.llorca, a.tulino}@nokia.com, ‡Universita degli Studi di Napoli Federico II, Italy, Email: antoniamaria.tulino@unina.it

Content

- **Introduction**
- Related Work
- IoT-Cloud Networks
- System Model
- The Internet of Things-Cloud Service Distribution Problem (IoT-CSDP)
- Evaluation
- Conclusions

Introduction

The Internet of Things (IoT) refers to the network of objects, devices, machines, vehicles, buildings, and other physical systems with embedded sensing, computing, and communication capabilities, that sense and share real-time information about the physical world.

However, with the predicted **explosion in the number of IoT services and connected devices**, traditional centralized cloud architectures, in which computing and storage resources are concentrated in a few large data centers, will inevitably lead to excessive network load, end-to-end service latencies, and unbearable energy costs.

In order to **meet the tight QoS requirements associated with real-time IoT applications while maximizing overall efficiency**, cloud architectures are becoming increasingly distributed, with the presence of small cloud nodes at the edge of the network, referred to as cloudlets, micro-clouds, fog nodes, or simply edge cloud nodes. Creating what we refer to as ***IoT-Cloud networks***.

Compared to traditional centralized clouds, IoT-Cloud networks provide increased flexibility in the allocation of resources to IoT services, and a clear advantage in meeting their stringent **latency**, **mobility**, and **location-awareness** constraints.

Introduction

The cloud service distribution problem (CSDP), where the goal is to find the placement of virtual functions and the routing of network flows that meets QoS requirements, satisfies re- source capacities and minimizes overall infrastructure cost.

In this paper, we formalize the **IoT-CSDP** as a service placement and resource allocation problem that goes beyond traditional information services and cloud architectures to include next generation IoT services and IoT-Cloud infrastructures.

Contributions:

- We introduce **a flexible mathematical model for IoT-Cloud networks**.
- Building on the cloud service model introduced in, we characterize a generic IoT service via a directed rooted graph that encodes the relationship between the service functions that act on the source information flows to create the final augmented information that needs to be delivered to the end users.
- We formally define the IoT service distribution problem **(IoT-CSDP)**.
- We **evaluate the solution to the IoT-CSDP** in an illustrative set of next generation smart environments.

Content

- Introduction
- **Related Work**
- IoT-Cloud Networks
- System Model
- The Internet of Things-Cloud Service Distribution Problem (IoT-CSDP)
- Evaluation
- Conclusions

Related Work

In the context of integrating end devices into the Cloud, special attention has been given to **wireless sensors networks (WSNs)** due to their relevance in the IoT.

In addition to wireless sensors, the **efficient integration of smart devices**, such as smartphones and connected vehicles, has also attracted the attention of the research community.

- ❖ software defined networking (SDN)
- ❖ vehicular ad hoc networks (VANETs)
- ❖ edge
- ❖ fog

While existent literature provides a significant amount of studies **describing alternative models and architectures that illustrate the advantages, limitations, and challenges associated with IoT-Cloud networks**, to the best of our knowledge, this is the first work that **mathematically formalizes the problem of optimal distribution of generic IoT services over IoT-Cloud networks**, taking into account the heterogeneous nature of sensing, transmission, and computing resources across the physical infrastructure, as well as the unique function inter- relationships, mixed-cast flow nature, and tight QoS require-ments of IoT services.

Content

- Introduction
- Related Work
- **IoT-Cloud Networks**
- System Model
- The Internet of Things-Cloud Service Distribution Problem (IoT-CSDP)
- Evaluation
- Conclusions

IoT-Cloud Networks

IoT-Cloud networks result from **the convergence of distributed cloud networks** and the IoT.

Service functionality can be **dynamically allocated** across the resulting highly distributed platform and flows can be **routed through the appropriate service functions** in order to maximize end devices' battery life, optimize service performance, and minimize overall operational cost.

The main advantages of IoT-Cloud networks:

- Low latency
- High reliability
- Reduced operational cost
- High flexibility
- Location awareness and mobility support
- Scalability

IoT-Cloud networks hence emerge as an ideal platform for the implementation of IoT services in the context of a wide range of smart environments, such as **smart grids, smart mobility, smart buildings, and smart cities.**

Content

- Introduction
- Related Work
- IoT-Cloud Networks
- **System Model**
  - **Network model**
  - **Service model**
  - **Flow model**
- The Internet of Things-Cloud Service Distribution Problem (IoT-CSDP)
- Evaluation
- Conclusions

Network Model

A. Network model

We refer to an IoT-Cloud network as the converged platform that results from the integration of programmable IoT devices into the cloud infrastructure (Fig. 1). We model an IoT-Cloud network as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with V vertices and E edges representing the set of network nodes and links, respectively. A node in $\mathcal{V}$ may represent an end device, an access point, or a cloud node (at possibly different hierarchical layers). Each node is characterized by its energy resources (e.g., power grid, battery), processing resources (e.g., processor, microprocessor), and data acquisition or sensing resources (e.g., camera, sensors, I/O interfaces). We denote by c_u^{pr} and c_u^{sn} the data processing and sensing capacities (in bits per second or bps) at node $u \in \mathcal{V}$, and by e_u^{pr} and e_u^{sn} the data processing and sensing unit energy costs (in Watts per bps) at node $u \in \mathcal{V}$, respectively. Nodes are interconnected via wireless or wireline links, each characterized by their transmission capacity and unit energy cost. We use c_{vu}^{tr} and e_{vu}^{tr} to denote the capacity (in bps), and the unit energy cost (in Watts per bps) of link $(v, u) \in \mathcal{E}$, respectively.

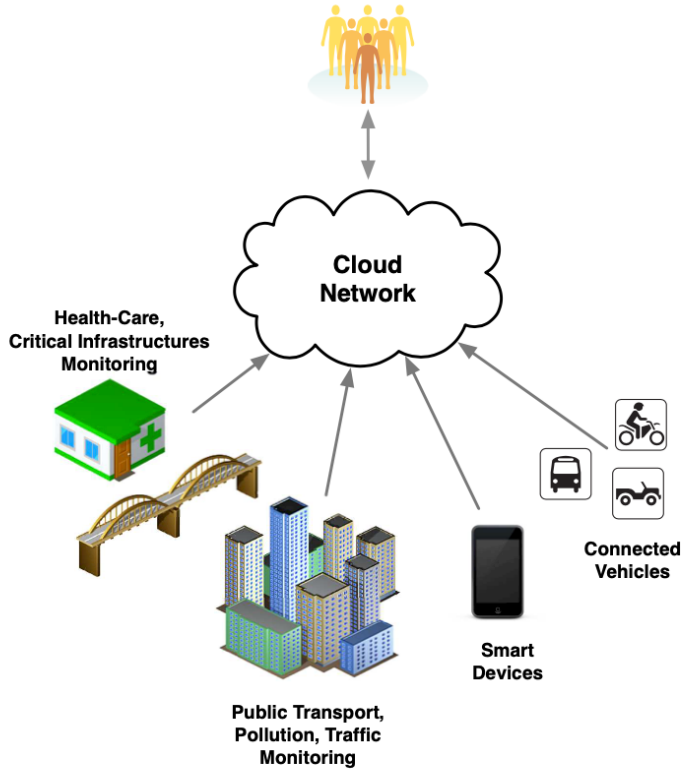


Fig. 1: IoT-Cloud network resulting from the integration of IoT devices into the cloud network infrastructure.

Service Model

We represent a generic IoT-Cloud service ϕ by a directed rooted graph $\mathcal{T}_\phi = (\mathcal{A}_\phi, \mathcal{O}_\phi)$. For any node $o \in \mathcal{O}_\phi$, there is a set of incoming edges $\{(z, o) \in \mathcal{A}_\phi : z \in \mathcal{Z}(o)\}$, as shown in Fig. 2. Hence, the set of objects $\mathcal{Z}(o) \subset \mathcal{O}$ required to generate object o via function p_o are represented as the children of o in the service graph $\mathcal{T}_\phi$. In particular, the root of the service graph $r_\phi \in \mathcal{O}$ represents the final information object that needs to be delivered to the end user(s). The service graph hence encodes the relationship between the virtual functions that act on the source objects to create the final object that needs to be delivered to the end users. When a user requests service ϕ, the user is, in essence, requesting the final information object or flow represented by the root of the service graph r_ϕ.

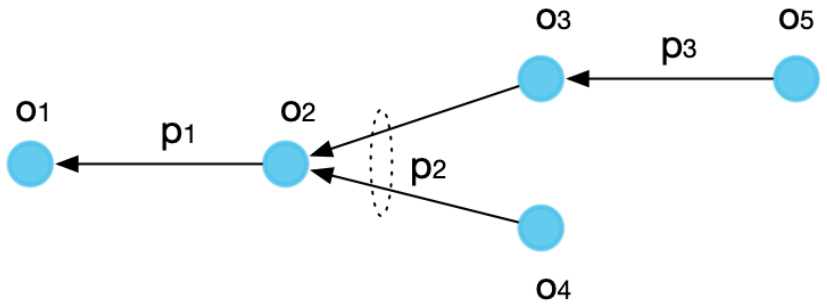


Fig. 2: An example of a service graph, $\mathcal{T}_\phi = (\mathcal{A}_\phi, \mathcal{O}_\phi)$, with $|\mathcal{O}_\phi| = 5$ objects, $|\mathcal{A}_\phi| = 4$ edges, and $|\mathcal{P}| = 3$ virtual functions.

Flow Model

- User-object flows: are characterized by a triplet (d, o, z), which indicates that the given flow is carrying information of object $z \in \mathcal{O}$ used to deliver final product $o \in \mathcal{O}$ at destination $d \in \mathcal{V}$. In particular, $f_{vu}^{tr,d,o,z}$, $f_{vu}^{sn,d,o,z}$, $f_u^{pr_i,d,o,z}$ and $f_u^{pr_o,d,o,z}$ indicate the fraction of object z carried/captured/processed by edge $(v, u) \in \mathcal{E}$ or node $u \in \mathcal{V}$ for final product $o \in \mathcal{O}$ at destination $d \in \mathcal{V}$, respectively. Note that we differentiate between $f_u^{pr_i,d,o,z}$ and $f_u^{pr_o,d,o,z}$, which denote the input and output flows of the processing unit at node $u \in \mathcal{V}$ associated with triplet (d, o, z). Fig. 3 illustrates the network flows associated with a given triplet (d, o, z) at node $u \in \mathcal{V}$, where pr represents the processing unit that hosts virtual functions and $q_u^{d,o,z}$ is a binary demand parameter that indicates if node $u \in \mathcal{V}$ requests object $z \in \mathcal{O}$. Note that $q_u^{d,o,z} = 0$ if $u \neq d$ or $z \neq o$, since users only request final information objects for themselves.
- Global flows: f_{vu}^{tr}, f_u^{sn} and f_u^{pr} determine the total amount of information flow carried, captured, and processed at a given physical link or node, respectively.

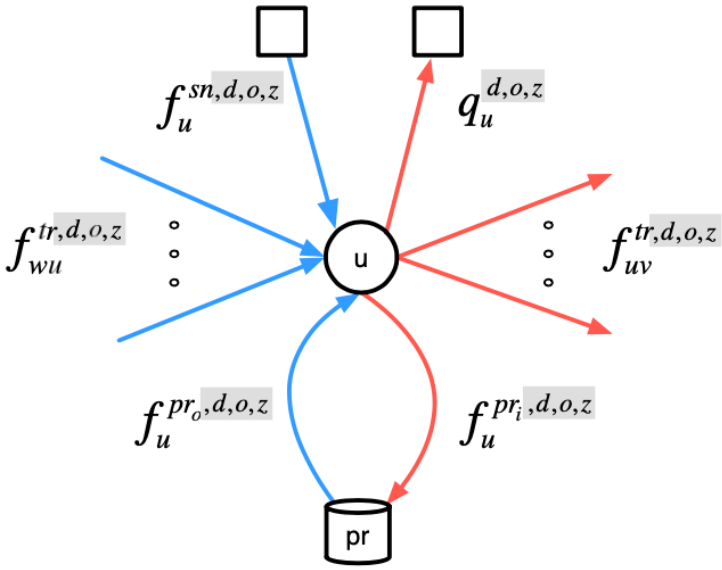


Fig. 3: Input/output user-object flows at node $u \in \mathcal{V}$.

Content

- Introduction
- Related Work
- IoT-Cloud Networks
- System Model
- **The Internet of Things-Cloud Service Distribution Problem (IoT-CSDP)**
  - **How the IoT-CSDP complements the CSDP**
  - **Mathematical formulation**
- Evaluation
- Conclusions

How IoT-CSDP complements CSDP

- **The sensing or data acquisition capabilities** of end devices, such as sensors, video cameras or RFID tags, **are considered**.
- **The capacities and energy costs of processing and transmission resources are modeled** across the entire heteroge- neous IoT-Cloud network, including core, metro, access, and end devices.
- **The limited energy resources of battery powered devices are taken into account** in order to guarantee their minimum lifetime requirements.
- **The reliability of links is considered** in order to characterize the possible packet losses and associated retransmis- sions, particularly relevant in low power wireless links.
- **The end-to-end latency is modeled** by considering the delay contributions along the entire service path, from the nodes that generate the source data to the delivery of the final augmented information to the end users.
- **The IoT-CSDP captures the unique nature of IoT services**, typically characterized by a multicast upstream phase in which sensed data that can be used for multiple services and end users is uploaded to edge cloud nodes, and a typically unicast downstream phase in which specific information resulting from the processing of sensed data is delivered in a personalized manner to the end users.
- Given that the cloud network infrastructure is shared among multiple services, **the IoT-CSDP assumes a load-proportional cost (e.g., energy) model**, in which the cost of a given service is proportional to its use of the physical infrastructure. This results in a significantly reduced complexity linear program that enables faster reactions to variations of users' service demands.

Mathematical Formulation

❖ Objective Function

$$\underset{f^{tr},f^{sn},f^{pr}}{\text{minimize}} \quad \sum_{(v,u)\in\mathcal{E}} e^{tr}_{vu} f^{tr}_{vu} + \sum_{u\in\mathcal{V}} (e^{sn}_{u} f^{sn}_{u} + e^{pr}_{u} f^{pr}_{u})$$

❖ Generalized Flow Conservation Constraints

$$q^{d,o,z}_u + f^{pr_i,d,o,z}_u + \sum_{w\in\mathcal{N}^+(u)} f^{tr,d,o,z}_{uw} =$$

$$f^{sn,d,o,z}_u + f^{pr_o,d,o,z}_u + \sum_{v\in\mathcal{N}^-(u)} f^{tr,d,o,z}_{vu} \quad \forall u,d,o,z.$$

$$f^{pr_o,d,o,z}_u \le f^{pr_i,d,o,y}_u \quad \forall u,d,o,z,y\in\mathcal{Z}(z).$$

❖ Function Availability Constraints

$$f^{pr_o,d,o,z}_u = 0 \quad \forall u,d,o,z,p_z \notin \mathcal{P}_u$$

❖ Source Constraints

$$f^{sn,d,o,z}_u = 0 \quad \forall u,d,o,z \notin O_u$$

$$f^{pr_o,d,o,z}_u = 0 \quad \forall u,d,o,z \in \mathcal{S}$$

❖ Mixed-cast Constraints

$$f^{sn,d,o,z}_u \le f^{sn,z}_u \quad \forall u,d,o,z,$$

$$\sum_{z\in\mathcal{O}} f^{sn,z}_u B_z = f^{sn}_u \quad \forall u,$$

❖ QoS Constraints

$$\sum_{(u,w)\in\mathcal{E}} f^{tr}_{uw} e^{tx}_u + \sum_{(v,u)\in\mathcal{E}} f^{tr}_{vu} e^{rx}_u +$$

$$f^{sn}_u e^{sn}_u + f^{pr}_u e^{pr}_u \le E_u \quad \forall u,$$

❖ Capacity Constraints

$$f^{tr}_{vu} \le c^{tr}_{vu} \quad \forall (v,u),$$

$$f^{sn}_u \le c^{sn}_u \quad \forall u,$$

$$f^{pr}_u \le c^{pr}_u \quad \forall u.$$

❖ Integer/Fractional Flow Constraints

$$f^{tr,d,o,z}_{vu}, f^{pr_i,d,o,z}_u, f^{pr_o,d,o,z}_u, f^{sn,d,o,z}_u \in \{0,1\},$$

or

$$f^{tr,d,o,z}_{vu}, f^{pr_i,d,o,z}_u, f^{pr_o,d,o,z}_u, f^{sn,d,o,z}_u \in [0,1].$$

Content

- Introduction
- Related Work
- IoT-Cloud Networks
- System Model
- The Internet of Things-Cloud Service Distribution Problem (IoT-CSDP)
- **Evaluation**
  - **Simulation details**
  - **Smart city**
  - **Smart Building**
  - **Smart Mobility**
- Conclusions

Simulation Details

- Present results from the solution to the IoT-CSDP for illustrative IoT services in smart environments, obtained via the linear programming solver Xpress-MP.

- We analyze and compare the efficiency of the IoT-Cloud solution, which finds the optimal location of IoT service functions exploiting the full flexibility of the IoT-Cloud infrastructure with
  i) a conventional cloud approach, in which all service functions are centralized at the highest cloud layer
  ii) the recently proposed cloudlet or fog approach, in which the processing of IoT services is handled by micro clouds located one hop away from the end devices
  iii) a fully distributed approach, in which all service functions are executed at the end devices.

- We consider a hierarchical IoT-Cloud network architecture composed of three main layers:
  i) a cloud layer, in which cloud nodes are organized into 3 tiers, i.e., a head office (HO) node representing the largest centralized data center, intermediate offices (IOs), and end offices (EOs),
  ii) an access layer, composed of base station (BS) nodes hosting micro-clouds (MCs) or cloudlets,
  iii) a device layer, containing wireless sensors, smart devices (e.g., smartphones, tablets, smart glasses), and connected vehicles.

Simulation Details

We compute approximate values using information extracted from [37] (for cloud equipment) and [38] (for wireless sensors/actuators), which are presented in Table I.

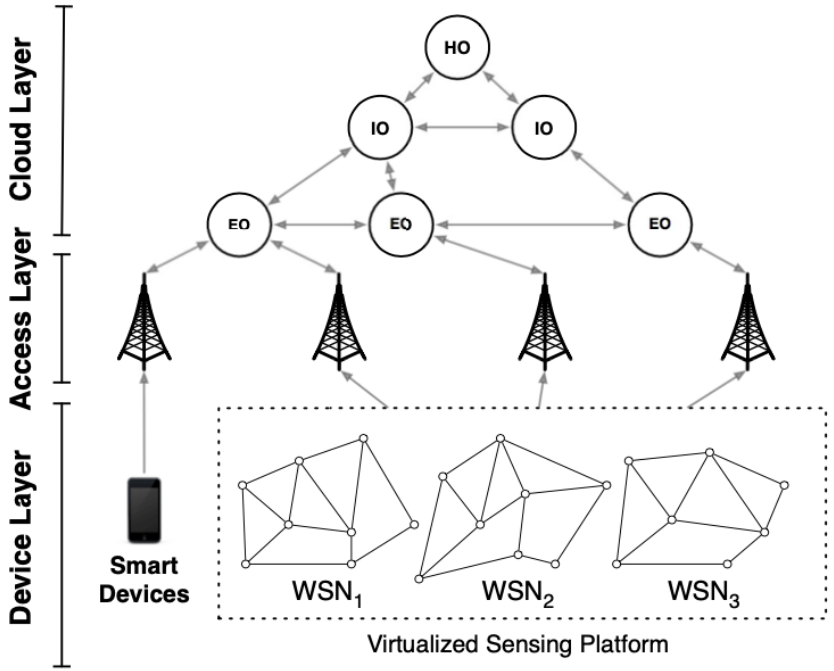
In terms of communication technologies, cloud nodes and base stations are connected via optical links, smart devices use 4G or WiFi, and wireless sensors communicate using the ZigBee protocol.

The average transport delay values considered in the sim- ulations include the effect of both queuing and propagation delays.

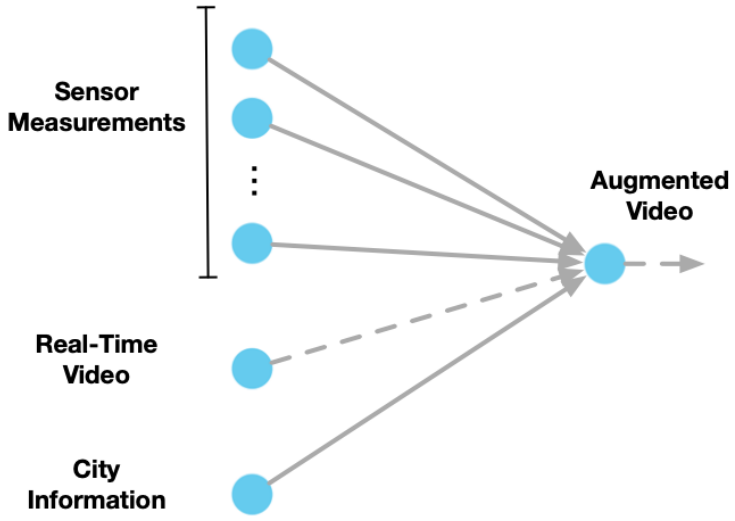
TABLE I: Typical capacities and efficiencies of IoT-Cloud network resources

	Capacity	Efficiency
Cloud Node (HO)	53.5 Million MIPS	500 MIPS/W
Cloud Node (IO)	26 Million MIPS	200 MIPS/W
Cloud Node (EO)	13 Million MIPS	133 MIPS/W
Micro Cloud Node (MC)	6.5 Million MIPS	100 MIPS/W
Wireless Sensor/Actuator	1 MIPS	480 MIPS/W
Smart Device	2000 MIPS	50-1000 MIPS/W
Connected Vehicle	2000 MIPS	1000 MIPS/W
Optical Link	4480 Gbps	12.6 nJ/bit
4G Link (Down/Up)	72/12 Mbps	76.2/19 μJ/bit
WiFi Link	150 Mbps	300 nJ/bit
ZigBee Link	250 kbps	100 nJ/bit

Smart City

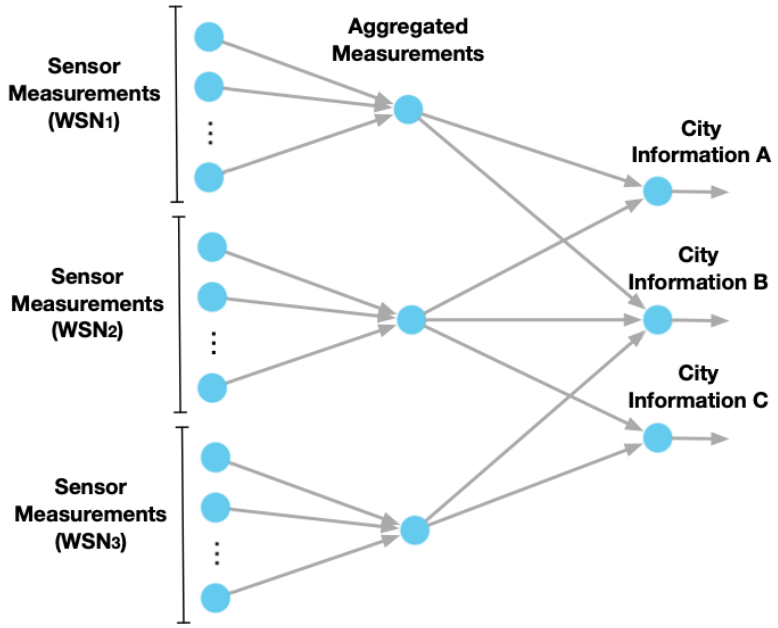


(a) Smart city architecture. The device layer is composed of smart devices, which provide live video streams and request augmented videos, and 3 WSNs collecting environmental information around the city.



(b) Service graph of the augmented reality application. This combines measurements from wireless sensors, live video streams from the smart devices and city information coming from the HO. Solid and dashed arrows indicate multicast and unicast flows, respectively.

Smart City



(c) Service graph of the city monitoring application. The city information is obtained through the analysis of the aggregated measurements collected by 3 different WSNs. Sensor readings are allowed to be sent via multicast, but the city information can only be sent via multicast for multiple users requesting the same information simultaneously.

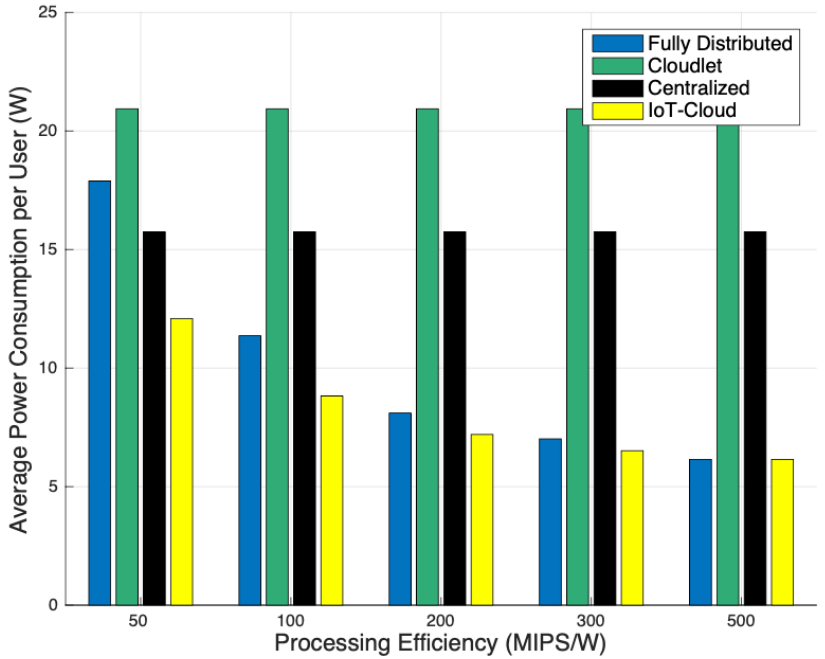


Fig. 5: Average power consumption per user of the augmented reality application for different smart device processing efficiencies.

Smart City

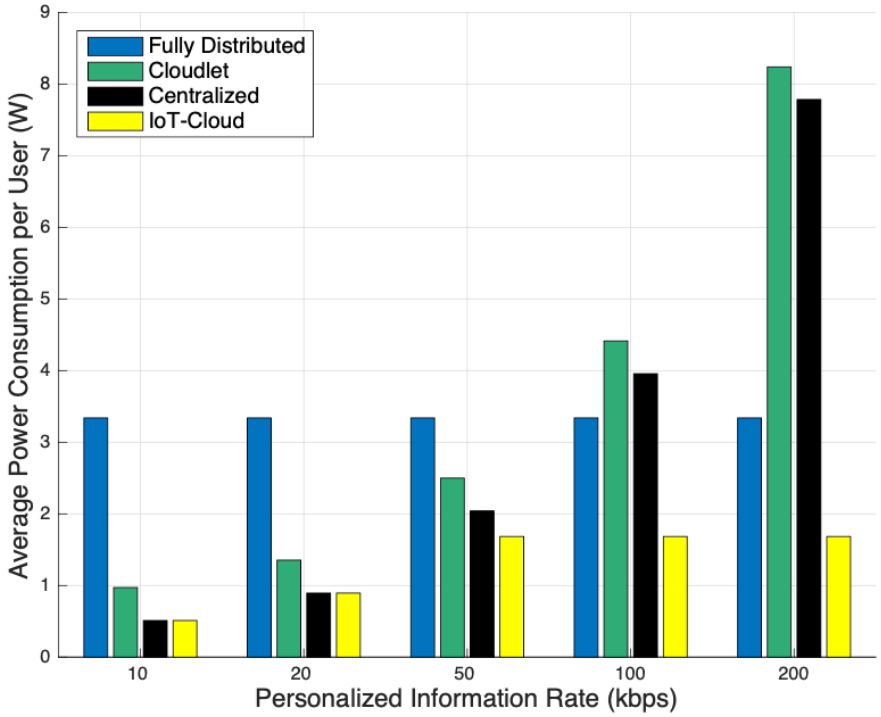
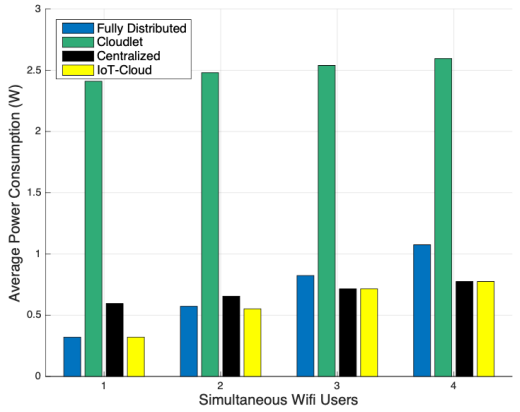
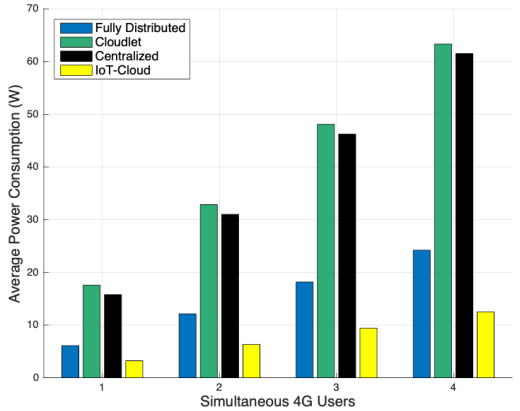


Fig. 6: Average power consumption per user of a city monitoring application for different personalized information rates.



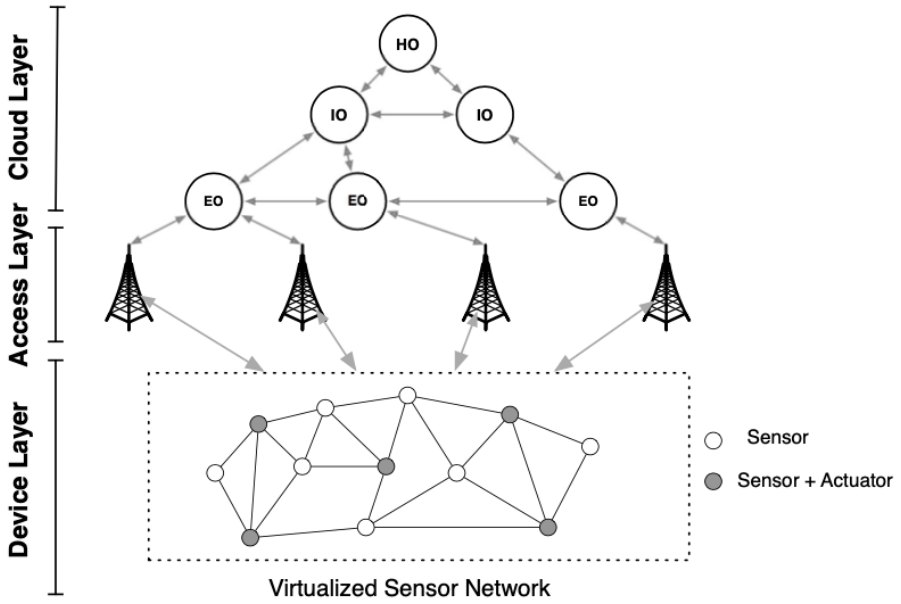
(a) Simultaneous WiFi users



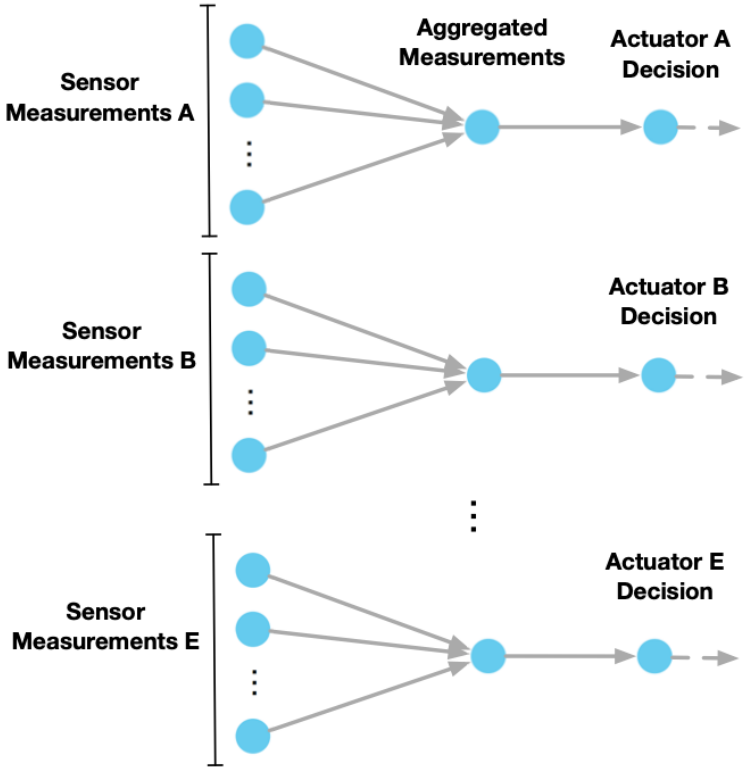
(b) Simultaneous 4G users

Fig. 7: Average power consumption per user of a city monitoring service for different number of simultaneous users.

Smart Building



(a) Smart building architecture. The device layer is composed of wireless sensors and actuators. The sensors collect the environmental measurements, while the actuators control the building automation system.



(b) Service graph of the smart building application. The sensor readings are first aggregated, to reduce the transport flow, and then analyzed to control the building automation system using the wireless actuators. Solid and dashed arrows state for multicast and unicast flows, respectively.

Smart Building

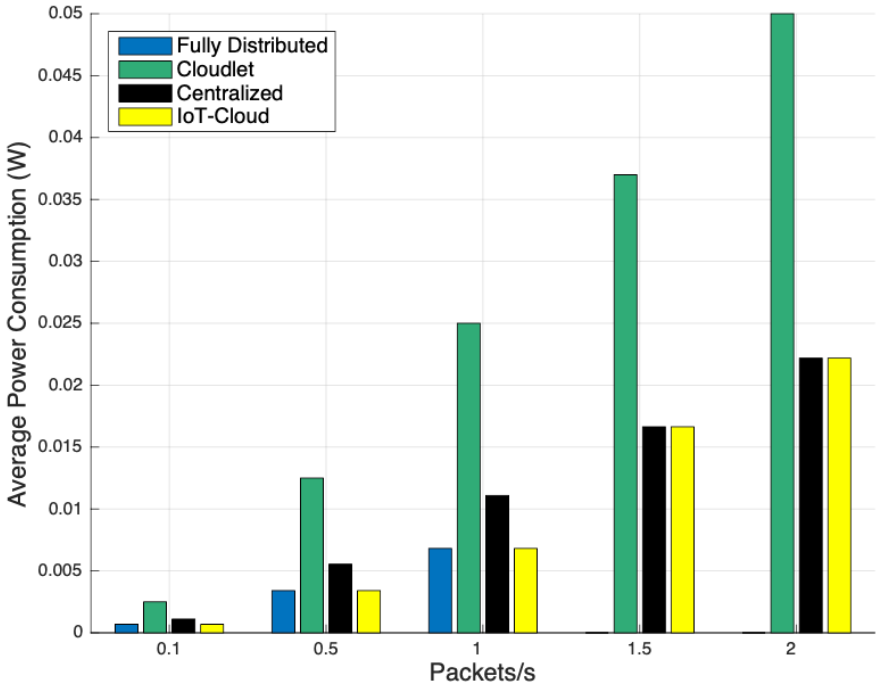
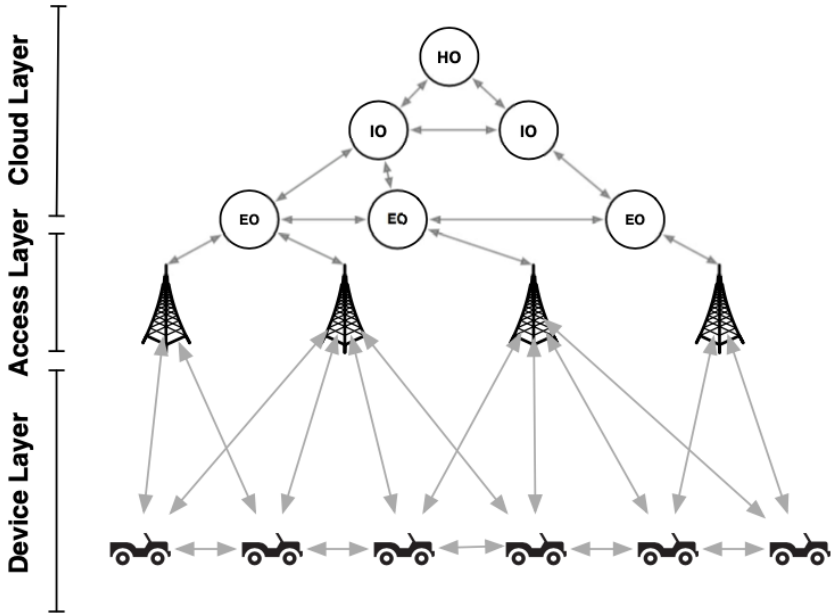
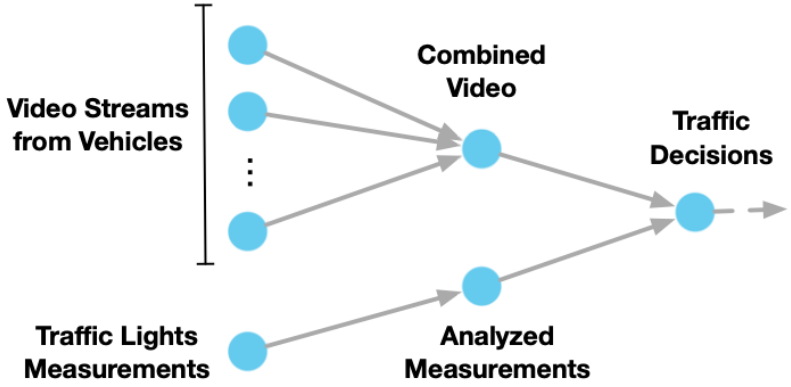


Fig. 9: Average power consumption of an autonomous sensing actuation service for different packet generation rates. Note that missing values indicate that the solution is infeasible.

Smart Mobility



(a) Smart traffic architecture. The device layer is composed by the connected vehicles, which provide life video streams to the traffic management system.



(b) Service graph of the smart mobility application. The measurements collected at the traffic lights are combined with the video streams of vehicles. Solid and dashed arrows indicate multicast and unicast flows, respectively.

Smart Mobility

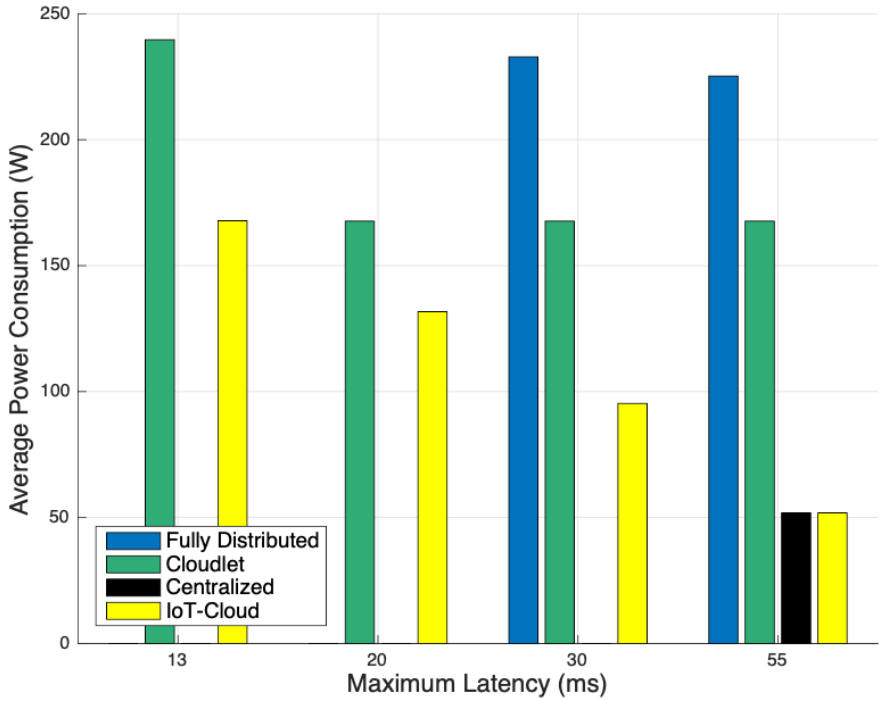


Fig. 11: Average power consumption of a traffic management service for different latency requirements. Note that missing values indicate that the solution is infeasible.

Content

- Introduction
- Related Work
- IoT-Cloud Networks
- System Model
- The Internet of Things-Cloud Service Distribution Problem (IoT-CSDP)
- Evaluation
- **Conclusions**

Conclusions

The confluence of distributed cloud networking and the Internet of Things (IoT) enables a new class of services that create augmented information from the cloud analysis of IoT data. IoT-Cloud networks reduce the distance between end users and cloud resources using edge cloud nodes distributed across the network, in order to support the key low latency, mobility, and location-awareness requirements of IoT services. **We proposed a comprehensive optimization framework** to enhance the efficiency of IoT services in next generation smart environments. We **formulated the service distribution problem (SDP)** in IoT-Cloud networks (IoT-CSDP) as a min-cost mixed-cast flow problem. The solution to the IoT-CSDP determines the optimal placement of service functions and the routing of network flows, taking into account the heterogeneous capacities and efficiencies of sensing, computing, and transport resources across the distributed IoT-Cloud infrastructure. We **solved the IoT-CSDP for the optimization of IoT services** in multiple smart environments. **Results show** that the IoT-CSDP solution captures the critical tradeoffs that appear in IoT-Cloud platforms due to the heterogeneity of IoT services, cloud network technologies, and end user devices. When compared to current solutions, smart IoT services optimized over a fully virtualized IoT-Cloud platform are shown to guarantee stringent QoS requirements in terms of reliability, battery lifetime, and end-to-end latency, while reducing overall power consumption by more than 80%.

Conclusions

Motivated by these promising results, interesting directions for future work include:

- implementing the proposed solution in a cloud computing testbed
- designing approximation algo-rithms of improved computational complexity
- studying the benefit of distributed optimization techniques to enable local reactions to fast system dynamics

Thanks.