Big Data Analytics for Smart Cities
The H2020 CLASS Project

Eduardo Quiñones[1], Marko Bertogna[2], Erez Hadad[3], Ana Juan Ferrer[4], Luca Chiantore[5], Alfredo Reboa[6]

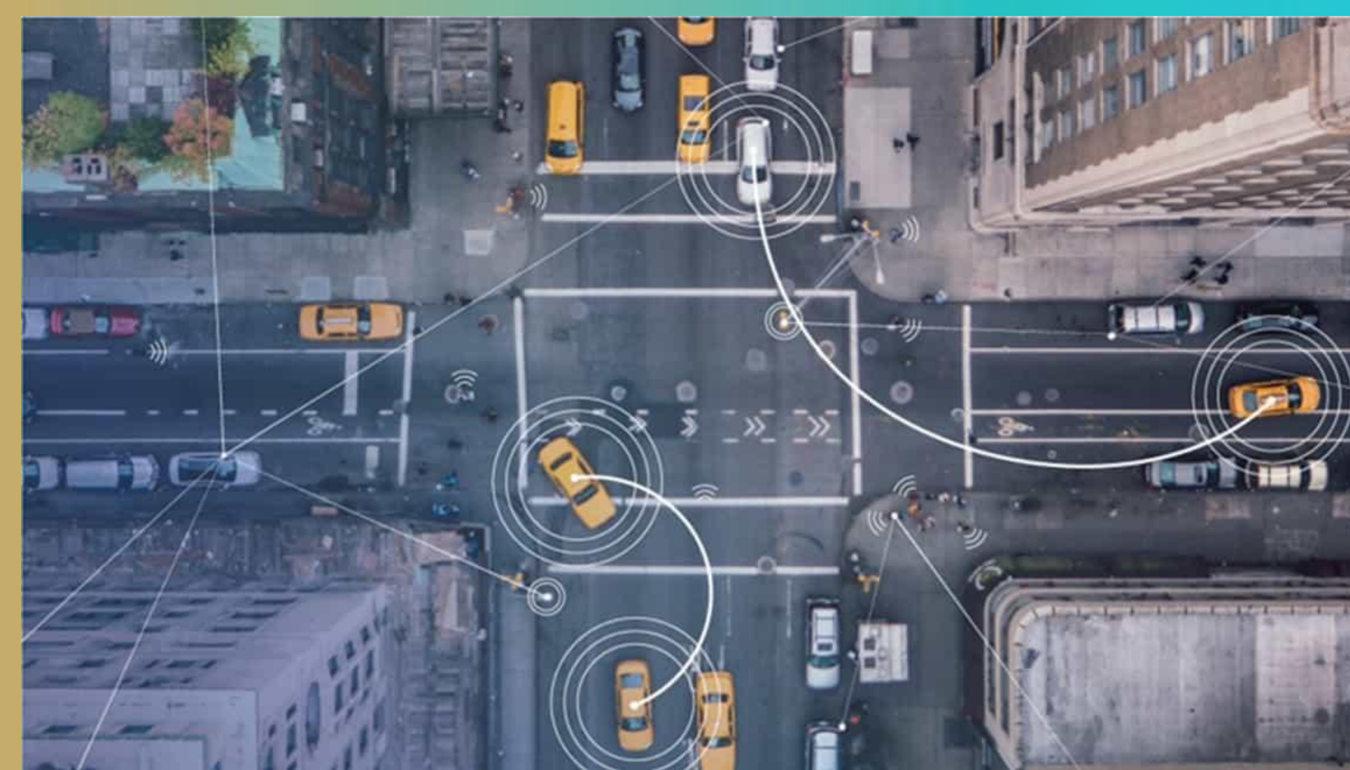
[1]Barcelona Supercomputing Center (BSC), [2]University of Modena, [3]IBM Israel, [4]Atos Research, [5]City of Modena, [6]Maserati

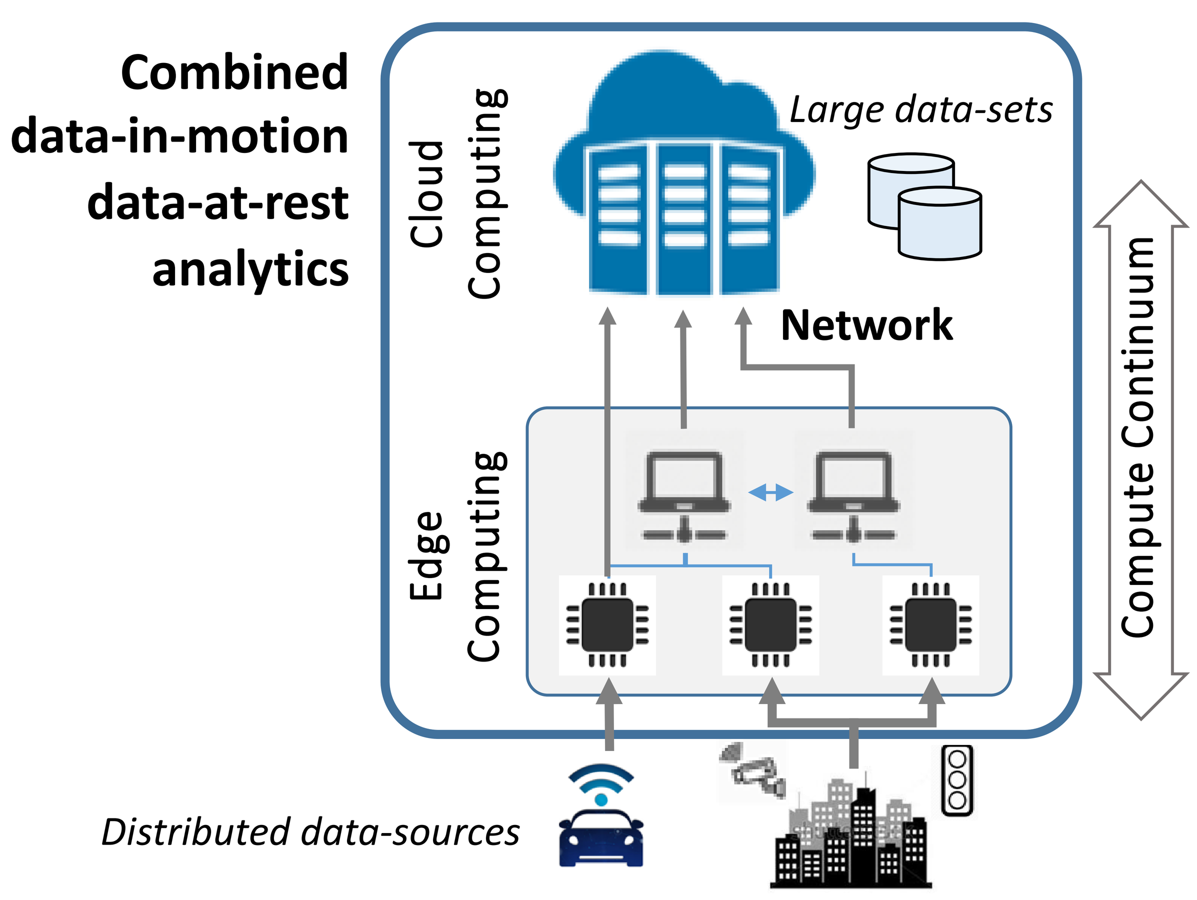
CLASS

Edge and Cloud Computation: A Highly Distributed Software for Big Data Analytics

Vision

- Computational challenges of smart cities can be effectively addressed by **coordinating** computing resources across the **compute continuum**
- Integration of technologies from **multiple computing domains** into a single development framework
  1. Advanced **data-analytics** solutions
  2. HPC techniques for an efficient workload distribution
  3. Timing analysis techniques
  4. Parallel heterogeneous **embedded** processor architectures





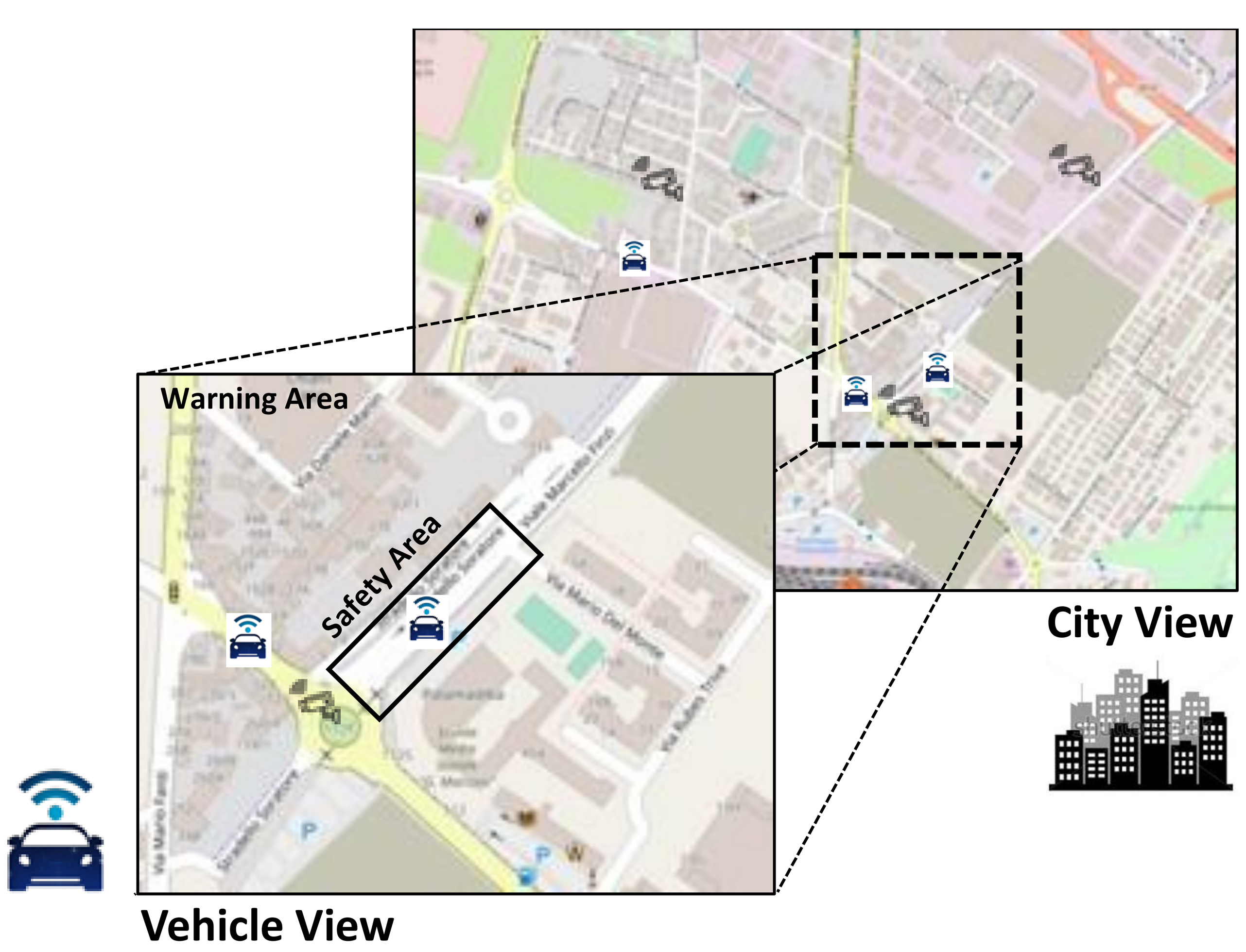
The CLASS Software Architecture

1. **Coordinate** edge and cloud computing resources
2. **Distribute** big-data workloads with **real-time requirements** along the compute continuum
3. **Combine** data-in-motion and data-at-rest analytics
4. **Increase productivity** in terms of programmability, portability/ scalability and (guaranteed) performance

Smart City Use Case

- Deployed in 1 Km^2 urban area of the city of Modena populated with IoT devices that exchange information
- Three vehicles equipped with *V2X connectivity* and *sensors*

City of Modena (Italy)



Applications Use Cases

- **Intelligent traffic management**, acting on traffic lights and smart road signals
  1. "Green routes" for emergency vehicles
  2. Traffic enhancement based on intelligent cross road management
- **Advanced driving assistance systems (ADAS)**
  1. Intelligent cross road management based on obstacle detection
  2. Automated valet parking systems
- **Knowledge extracted from sensor fusion** from geographically disperse data-sources coming from city and vehicle sensors

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación | UNIMORE Università degli studi di Modena e Reggio Emilia | IBM | Atos | Comune di Modena | MASERATI

The CLASS project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 780622.

www.class-project.eu

Big Data Analytics for Smart Cities: The H2020 CLASS Project[1]

Eduardo Quiñones*, Marko Bertogna†, Erez Hadad‡, Ana Juan Ferrer§, Luca Chiantore¶, Alfredo Reboa‖
*Barcelona Supercomputing Center, Spain
†University of Modena, Italy
‡IBM, Haifa, Israel
§Atos Research, Barcelona, Spain
¶City of Modena, Italy
‖Maserati, Modena, Italy

***Abstract*—The digitalization process is making cities to rapidly increase the amount of data to be processed. On one side, the advent of autonomous vehicles challenges big data analytics due to the need of accomplishing safety requirements. On the other side, the dispersion nature of data sources makes current data analytics systems not suitable for smart cities, commonly designed following two conflicting priorities: (i) a quick and reactive response (*data-in-motion analysis*); or (ii) a thorough and more computationally intensive feedback (*data-at-rest analysis*). These approaches have been tackled separately although they provide complementary capabilities.**

This paper presents the CLASS project, whose objective is to develop a novel software architecture to help big data developers to fully benefit from a combined data-in-motion and data-at-rest analysis by efficiently distributing data and process mining along the *compute continuum* (from edge to cloud resources) in a complete and transparent way, while providing sound real-time guarantees imposed by autonomous vehicles. The capabilities of this novel software architecture will be demonstrated on a real smart-city use case in the city of Modena, featuring a heavy sensor infrastructure to collect real-time data across a wide urban area, and prototype cars equipped with heterogeneous sensors/actuators, V2I and V2V connectivity.

I. Introduction

Big data analytics has become a key enabling technology across multiple application domains, to address societal, economic and industrial challenges. Providing the required computational capacity for absorbing in real-time extreme amounts of complex and distributed data being continuously collected, is fundamental to allow converting the data into few concise and relevant facts that can be then acted upon. These complex systems with analytics-augmented control are often regarded as smart, e.g., smart cities.

For the purpose of providing relevant big data analytics, the current systems are designed following two different and apparently opposite priorities, which are typically transparent to the analytics consumer. On one hand, the priority may be to provide quick and reactive information, possibly in real-time, based on the flowing stream of data to focus on the most relevant aspects of the stream. On the other hand, the priority may be to provide thorough and more time consistent responses, which typically implies aggregating as much information as possible into large models. The former is commonly categorized as *data-in-motion* analysis; the latter is commonly categorized as *data-at-rest* analysis [4]. Historically, both approaches have been tackled separately, given their apparently incompatible requirements. Their capabilities however, are complementary, and so future data analytics systems can take full advantatge of their integration.

One of the smart computing domains that can take full benefit of such integration is the smart city domain. The use of combined data-in-motion and data-at-rest analysis provides to cities efficient methods to exploit the massive amount of data generated from heterogeneous and geographically distributed sources including citizens, traffic, vehicles, city infrastructures, buildings, IoT devices, etc. Moreover, smart cities are considered as an enabler for the advent of more complex *Advanced Driving Assistance Systems* (ADAS) and lattely *Autonomous Driving* (AD) vehicles.

Clearly, the huge amount of heterogeneous and geographically distributed data sources makes difficult (if not impossible) to implement efficent data-analytics applications into a centralized computing system. Moreover, these types of systems cannot provide the required timing guarantees.

II. The H2020 CLASS Project

The vision of the CLASS[1] project, is that the pressure that the newest smart systems requiring big data analytics and real-time requirements will put on computing systems, can be efficiently addressed by devising a full distributed system architectures in which a combined data-in-motion and data-at-rest analytics can be performed by efficiently coordinating resource across the *compute continuum*, from cloud in which data-servers are located, to data sources and edge devices that are co-located with them (see Figure 1).

To that end, the CLASS project, coordinated by BSC and participated by University of Modena, the City of Modena, Maserati, IBM and Atos, aims to develop a novel software architecture to help programmers and big data practitioners to combine data-in-motion and data-at-rest analysis, by efficiently distributing data and process mining along the compute continuum, while providing real-time guarantees.

The ability of providing distributed and trustable real-time properties to data analytics, opens the door to the use of big data into critical real-time systems, providing to them superior

[1]CLASS stands for *"Edge&**CL**oud Computation: **A** Highly Distributed **S**oftware Architecture for Big Data Analytic**S**"*, Grant No. 780622, www.class-project.eu

data analytics capabilities to implement more intelligent and autonomous control applications, e.g., full autonomous cars, manufacturing robots, unmanned aerial vehicles (UAV).

This computation paradigm, also referred to as *fog computing*[1], has a system-level horizontal architectural view that distributes resources and services of computing located anywhere along the compute continuum. The principle behind distributing the computation along the continuum is that decision making occurs as close as possible to where the data is originated (either at edge or cloud side), enabling faster processing time with timing guarantees and lowering network costs. In that respect, an efficient use of novel parallel software and hardware architectures to distribute, optimize and combine data-in-motion and data-at-rest processing is of paramount importance at both edge and cloud computing side, to exploit the full potential benefits of big data analytics:

On the edge computing side, the availability of new highly parallel heterogeneous embedded processor architectures, (e.g. GPUs, many-core fabrics, SoC-FPGAs) may enable the processing of complex data-in-motion analytic processes with a reduced cost and power consumption. Recently proposed techniques for the real-time elaboration of heavy streams of data collected from the physical world [5] may allow for smart edge computing paradigms that significantly reduce the pressure on communication, processing and storage capabilities at cloud and data-center level. By doing so, short span responses (both in time and space) can be produced using low-cost data analytics algorithms.

On the cloud computing side, data-in-motion analysis can be completed (and enriched) with data-at-rest analysis with a more efficient usage of server resources for executing big-data workloads implying higher throughput of execution (more workloads completed per time-unit) and higher resource utilization. Combined with the smart distribution along the compute continuum envisioned by CLASS, these techniques can significantly reduce the pressure on data-centers. By doing so, reliable models can be built, which are periodically updated and queried using high-accuracy data analytics algorithms.

Finally, *on the software architecture side*, CLASS will enrich the newest distributing computing frameworks used in the high-performance domain, e.g., COMPSs[2], Spark[3], with timing and schedulability analysis methods, to fully exploit the computation capabilities of the overall compute continuum, while providing real-time guarantees.

III. Smart City Use-Case

The novel *CLASS software architecture* will be evaluated on the *Automotive Smart Area* (ASA) from the municipality of the City of Modena (Italy). The ASA represents a real urban laboratory of one square-kilometre equipped with multiple (IoT) devices (e.g., smart cameras, traffic scanner and counter, smart parking, weather conditions) and network connectivity (4G, LTE) that enables to exchange massive amounts of information to jointly cooperate for a so-called *distributed awareness* of what happens in a wide city area. Moreover, Maserati will provide four prototypes of highly-connected

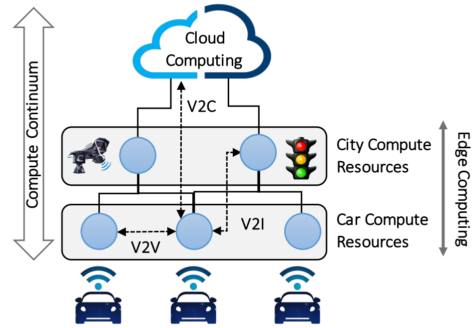


Fig. 1. Distributed computing paradigm used for the smart city use-case.

cars with the necessary *vehicle-to-vehicle* (V2V), *vehicle- to-infrastructure* (V2I) and *vehicle-to-cloud* (V2C) connectivity, tailored to the CLASS use-case. Cars will be further equipped with multiple cameras @4K resolution, long-range and middle range radars and ultrasound sensors for short range detection.

Figure 1 shows a schematic view of the distributed computing paradigm considered in the CLASS project composed of compute resources distributed across the city (including those within cars), connected to cloud computing resources.

We aim to efficiently process the data-sets generated by the ASA and the prototype cars with the CLASS architecture, so the envisioned data-analytics services can be developed and executed to extract valuable information in real-time for an improved traffic management at different resolutions, e.g., at the level of a single intersection, a city block, a neighbourhood, or even the whole city area; and a trustworthy and safe communication of real-time information for enabling next-generation ADAS and AD vehicles. While the main target of the project is not developing new smart city applications, rather than preparing the challenging computing infrastructure necessary to allow the development of next-generation applications, a set of applications has already been identified for the ASA setting to test and highlight the benefits of our distributed big-data analytics architecture:

- Intelligent traffic management, driving traffic lights and smart road signals based on traffic and emergency conditions. "Green routes" will be dynamically created for public vehicles (e.g., ambulances, fire-fighters and police vehicles) and traffic will be adapted to the monitored car and pedestrian flow and/or to reduce overall fuel consumption and CO2 emission at crossroads.
- Advanced driving assistance to highly connected cars supporting pedestrian and obstacle avoidance, path/route planning depending on traffic conditions, or automated parking systems driven by smart cameras.

References

[1] OpenFog Consortium, https://www.openfogconsortium.org

[2] COMP Superscalar, "https://www.bsc.es/research-and-development/software-and-apps/software-list/compsuperscalar"

[3] Apache Spark, "http://spark.apache.org"

[4] Big Data Value Strategic Research and Innovation Agenda (SRIA), Version 3.0, Jan 2017

[5] EC FP7 project P-SOCRATES (Parallel Software Framework for Time-Critical Many-core Systems)