Challenges in Resource Management in the ELASTIC and AMPERE European Projects

Luis Miguel Pinho
LMP@isep.ipp.pt

CERCIRAS COST Action Meeting

September 2-3, 2021

isep Instituto Superior de Engenharia do Porto

P.PORTO

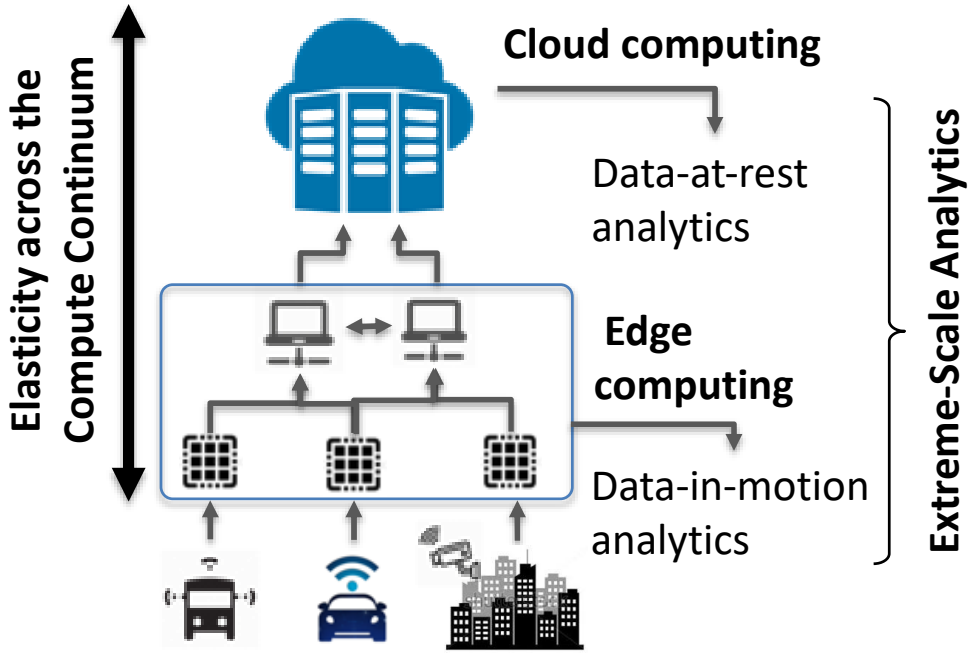
H2020 ELASTIC Quick Facts

- ELASTIC: A Software Architecture for Extreme-ScaLe Big-Data AnalyticS in Fog CompuTing ECosystems
- H2020 RIA project (Dec-2018, May-2022)
- Website: https://elastic-project.eu/
- Coordinator: BSC, Spain
- Partners

"The ELASTIC project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 825473"

Motivation

- Extreme-scale analytics are more and more a key enabling application for smart systems
  - process huge amounts of heterogenous data, geographically dispersed, both on the fly and at rest
  - necessity to fulfil non-functional properties inherited from the system (real-time, energy efficiency, communication quality or security)
- Providing the required computing capacity for extreme-scale analytics is of paramount importance
  - dynamically manage resources as needed, guaranteeing required levels of service
  - consider the full architecture of the system, from the Edge devices to cloud infrastructures



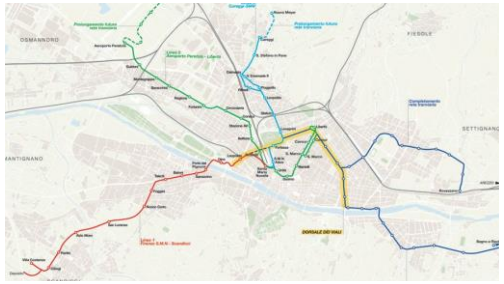
Motivation

- **Challenge**: fulfil non-functional properties
  - including real-time, energy-efficiency, quality of communications, security
  - need to consider these in a holistic way, as they are interdependent
- **Challenge**: limits of the existent elasticity concept
  - in which cloud computing resources are orchestrated to provide maximum throughput
  - does not take into account the computing resources located on the edge
  - elasticity mainly focuses on system throughput, without taking into account the non-functional properties
- Need to **manage resources to address these two challenges along the compute continuum**, i.e., from the edge to the cloud
  - paramount importance to take full benefit of extreme-scale analytics in smart systems, in industrial and societal environments
  - there are no known end-to-end solutions applied along the complete compute continuum

ELASTIC Use-Case

- Smart City use case
  - Test and highlight the benefits of the ELASTIC Software architecture
    1. Next Generation Autonomous Positioning (NGAP) and Advanced Driving Assistant System (ADAS)
    2. Predictive maintenance
    3. Interaction between the public and the private transport in the City of Florence
  - Deployed on the Florence tramway network (Italy) with tram vehicles equipped with a variety of sensors/computing/network capabilities

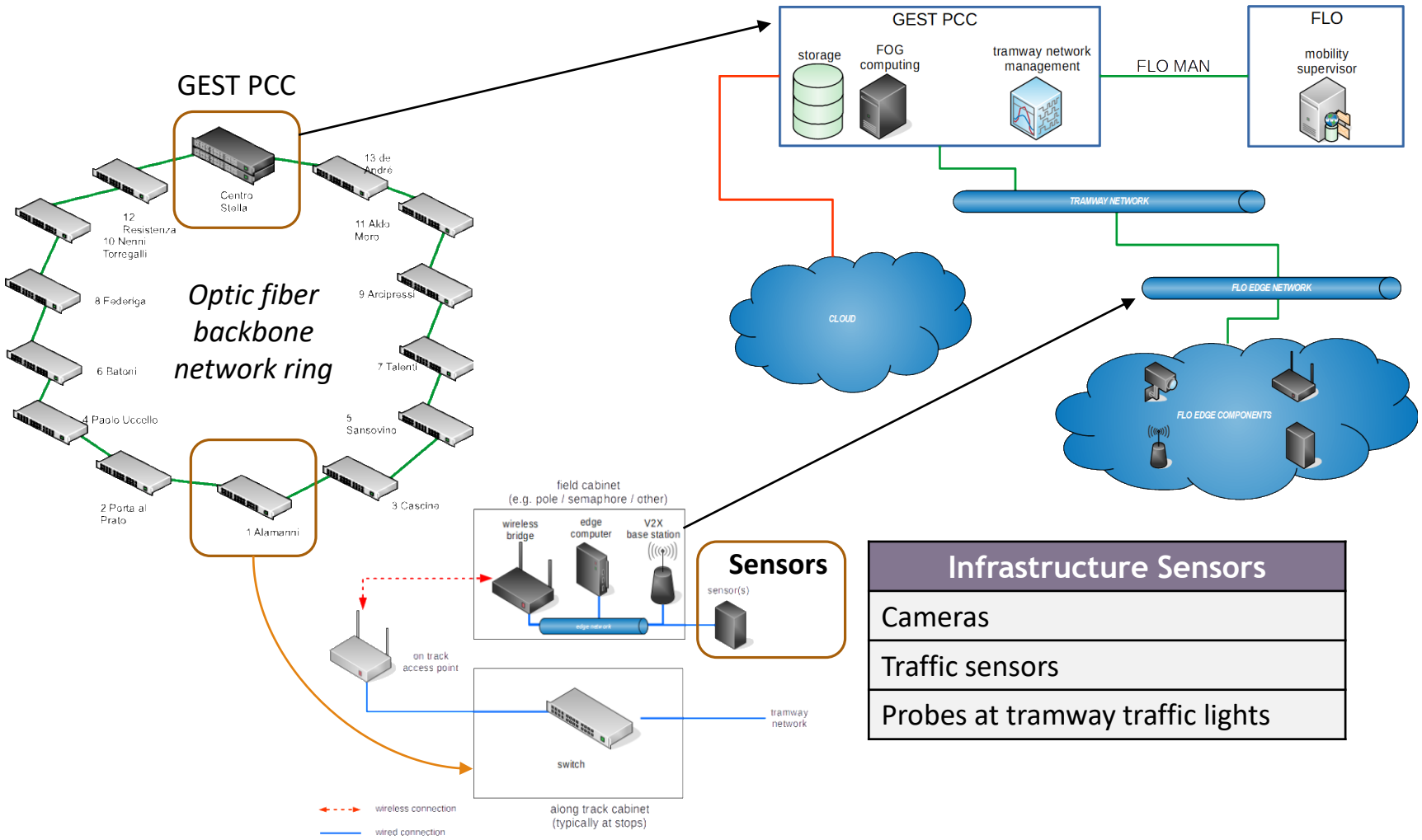
City of Florence



Florence Tramway network

Tram vehicle from the Florence Tramway network

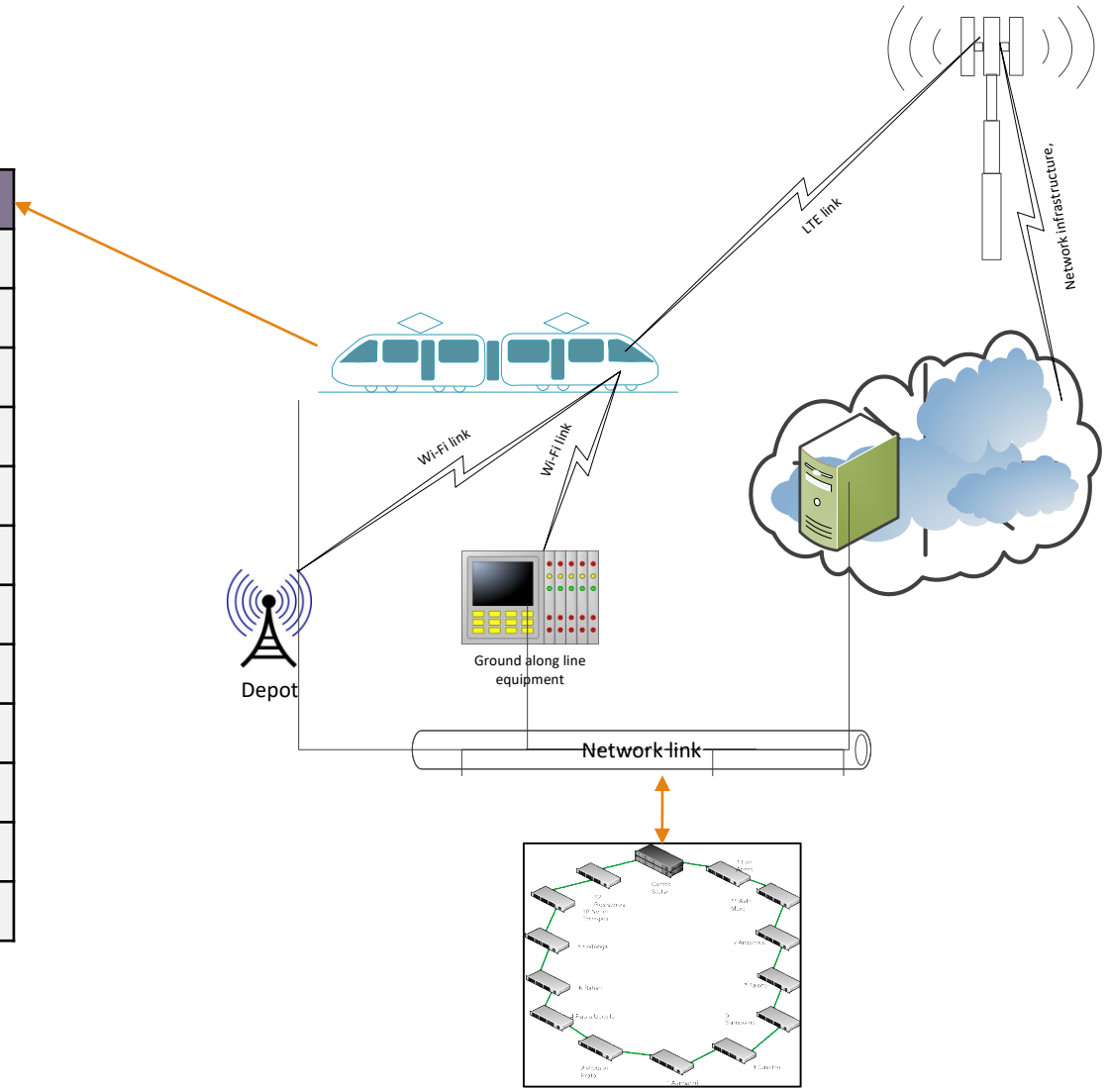
ELASTIC Use-Case



Infrastructure Sensors
Cameras
Traffic sensors
Probes at tramway traffic lights

ELASTIC Use-Case

Tram Vehicle Sensors
Laser-based tram rail track
Inertial Platform
Triaxial wheels' accelerometers
Odometer
Electric power meter
Tram vehicle weight device
IMU
GPS
Radar/LiDAR
Infrared Camera
Tram Position
Several Cameras



ELASTIC Concept

- ELASTIC software architecture takes into consideration a number of trade-offs
  - performance, precision/accuracy, non-functional system properties
  - dynamic management of computation
- Edge devices may deliver the time-predictability needed to implement real-time functionalities
  - but do not provide sufficient computational power to run analytics
  - fast and time-predictable, but limited, precision algorithms will be deployed on the edge-side for data-in-motion
- Cloud computing resources provide the computation capabilities to support complex analytics
  - but communication delays may make systems unstable
  - cloud resources will be used to run only accurate but costly models for the long-term refinement and global modelling

ELASTIC Concept

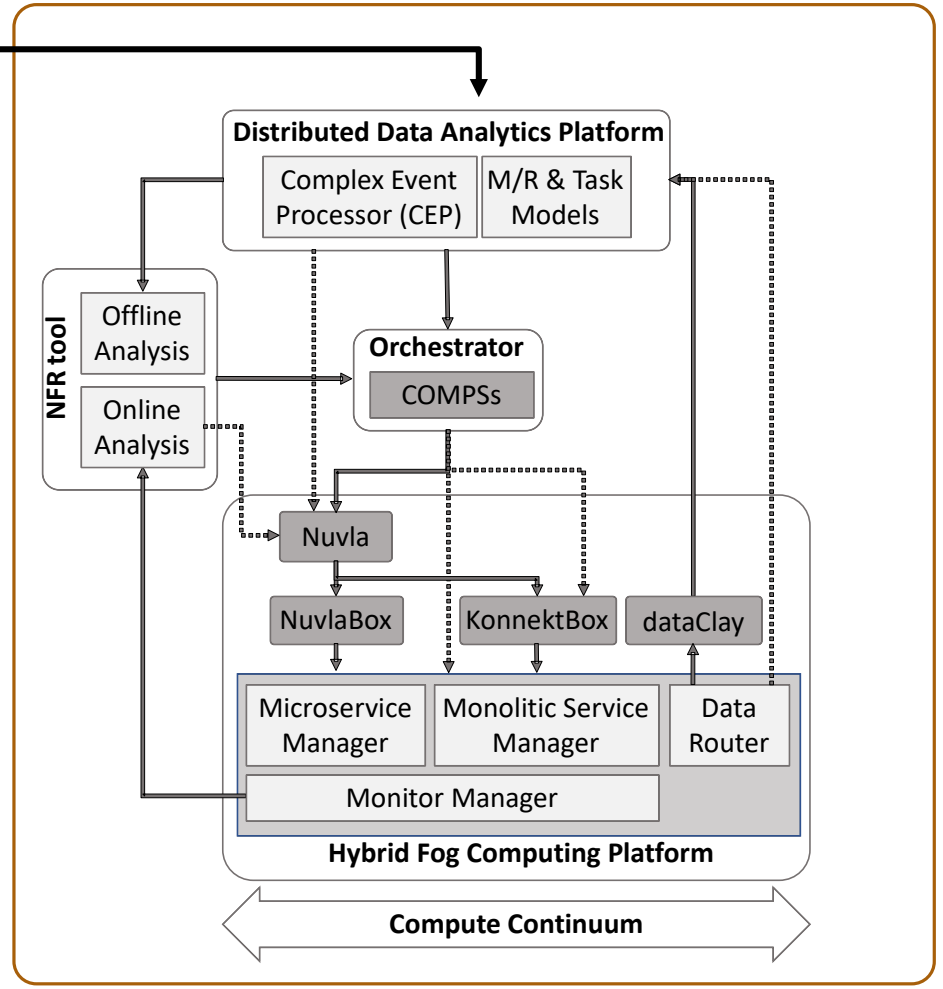
Sensors

Data collected from the city and the tramway vehicles

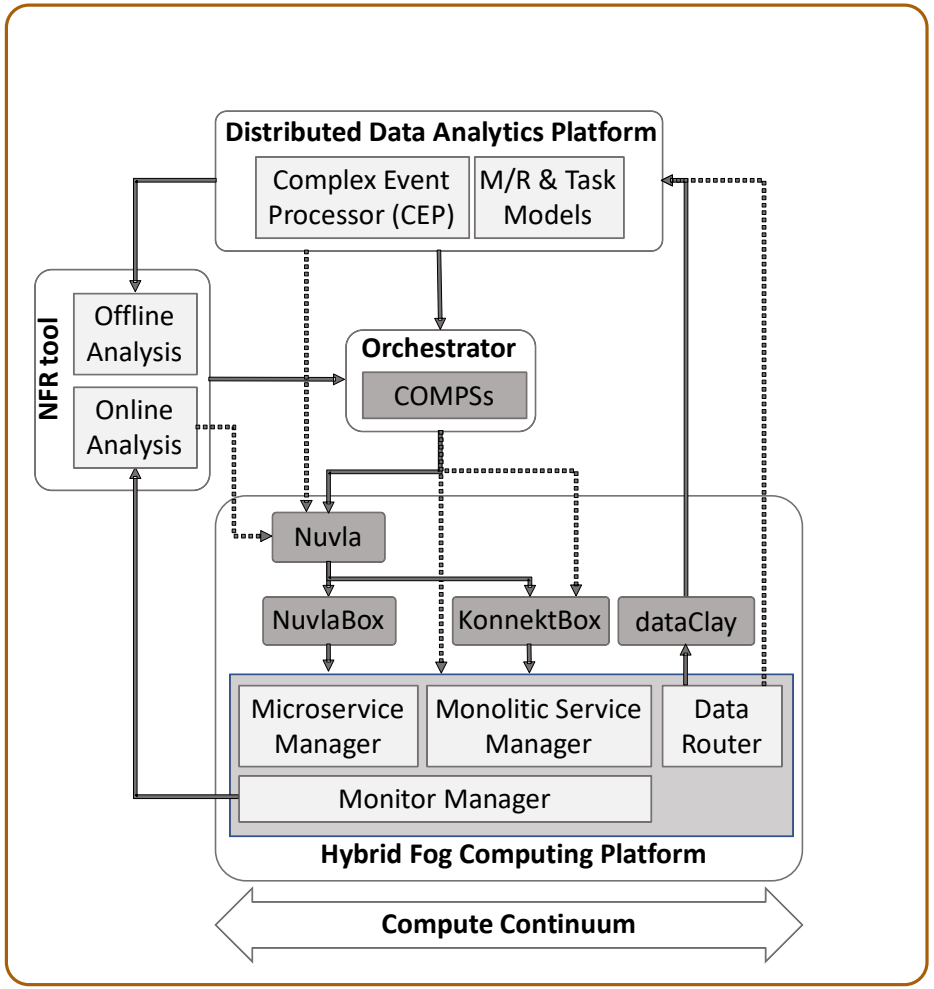
Communication Infrastructure

Data-Sets + Model Construction
(*Knowledge included in the datasets*)

Algorithms and Analytics Methods
(*Extract knowledge from the model*)

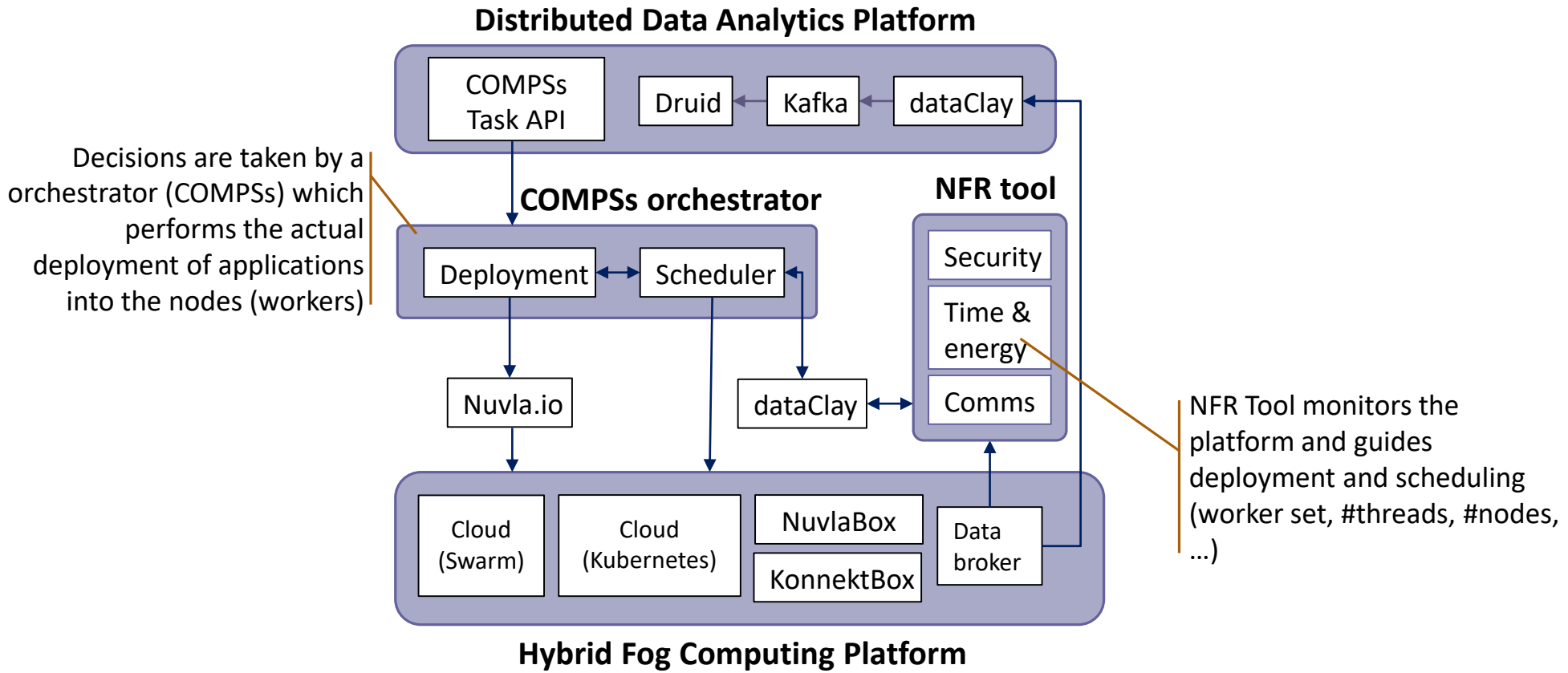


ELASTIC Concept

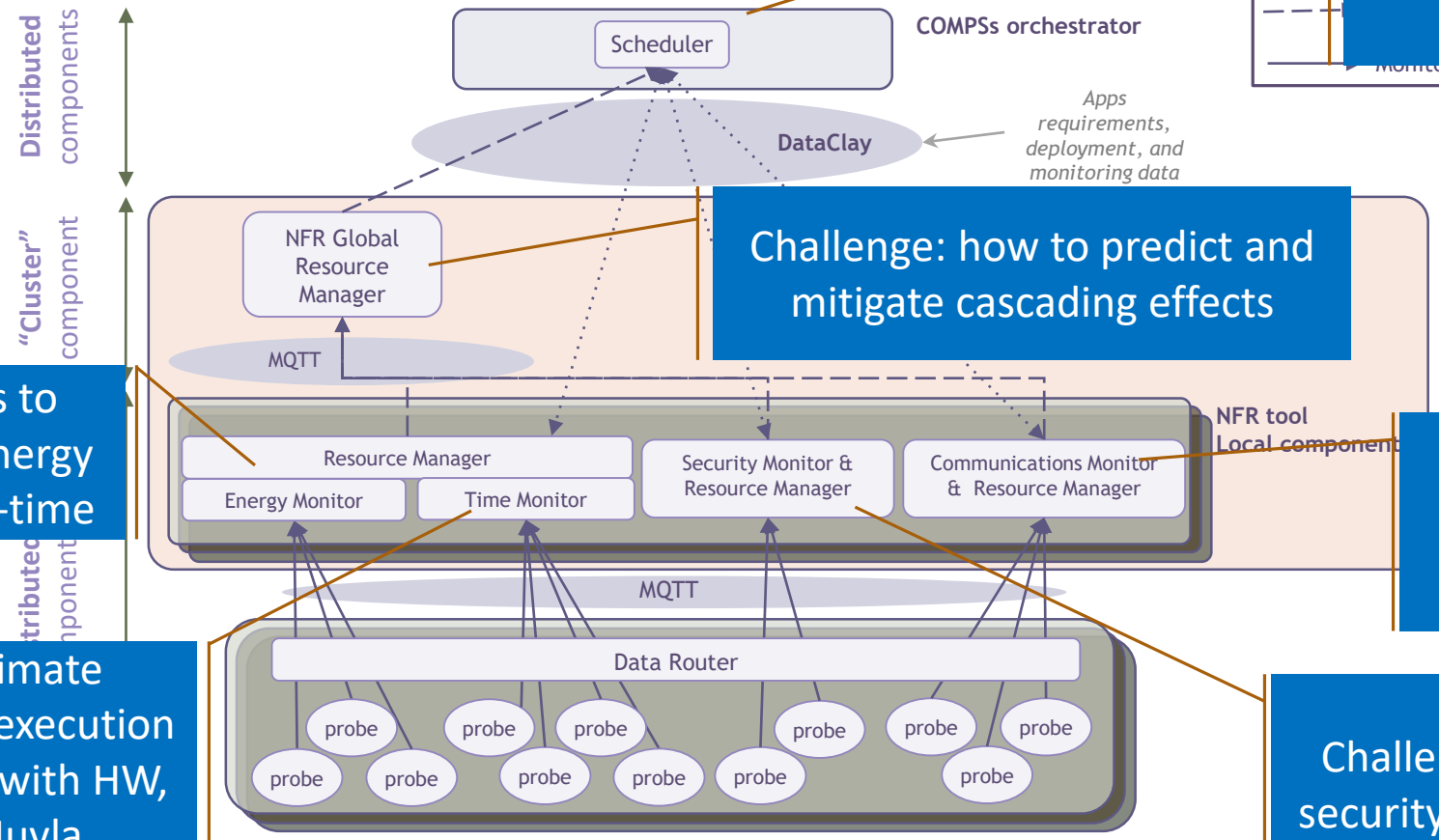


Software Component	
Distributed Data Analytics Platform	COMPSs
	Flink
	Spark
	Kafka
Orchestrator	COMPSs
NFR tool	Static Analysis tools
	Run-time analysis tools
Hybrid Fog Computing Platform	Nuvla/NuvlaBox
	KonnektBox
	dataClay
	Kubernetes
	Docker

ELASTIC Concept



ELASTIC Resource Challenges



Challenge: how to "optimize" the different dimensions in a holistic way

Challenge: how to predict and mitigate cascading effects

Challenge: models to predict high-level energy consumption at run-time

Challenge: model for communication quality of links and application flows

Challenge: estimate applications time execution in a vertical stack with HW, OS, Docker, Nuvla

Challenge: how to monitor security and enforce changes

H2020 AMPERE Quick Facts

- AMPERE: A Model-driven development framework for highly Parallel and EneRgy-Efficient computation supporting multi-criteria optimisation
- H2020 RIA project (Jan-2020, Dec-2022)
- Website: https://www.ampere-euproject.eu/
- Coordinator: BSC, Spain
- Partners

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Instituto Superior de Engenharia do Porto

ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Scuola Superiore Sant'Anna
di Studi Universitari e di Perfezionamento

EVIDENCE® EMBEDDING TECHNOLOGY

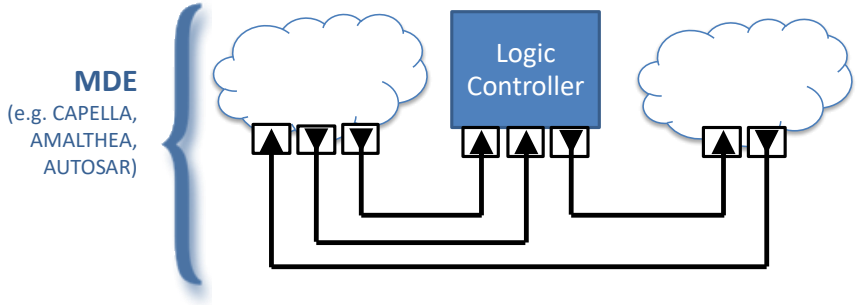
BOSCH

THALES

SYSGO EMBEDDING INNOVATIONS

"The AMPERE project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 871669"

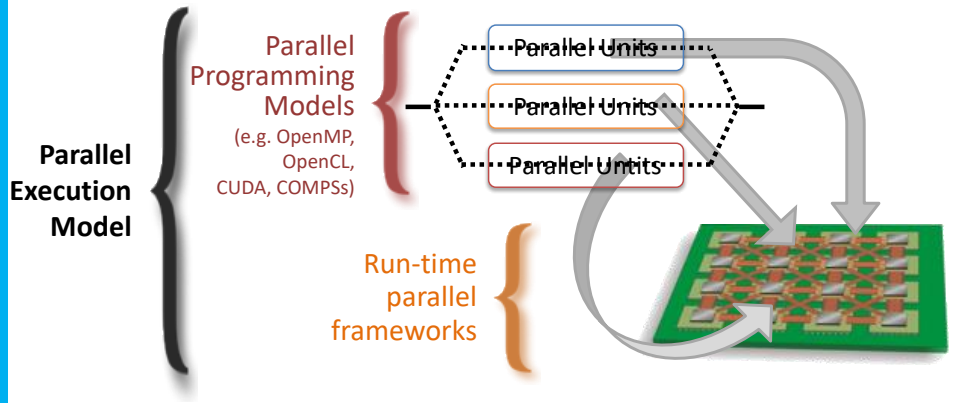
Motivation



Model Driven Engineering (MDE)

1. Construction of complex systems
2. **Formal verification** of functional and non-functional requirements with **composability** features
3. **Correct-by-construction paradigm** by means of code generation
   - Suitable only for single-core execution or with very limited multi-core support

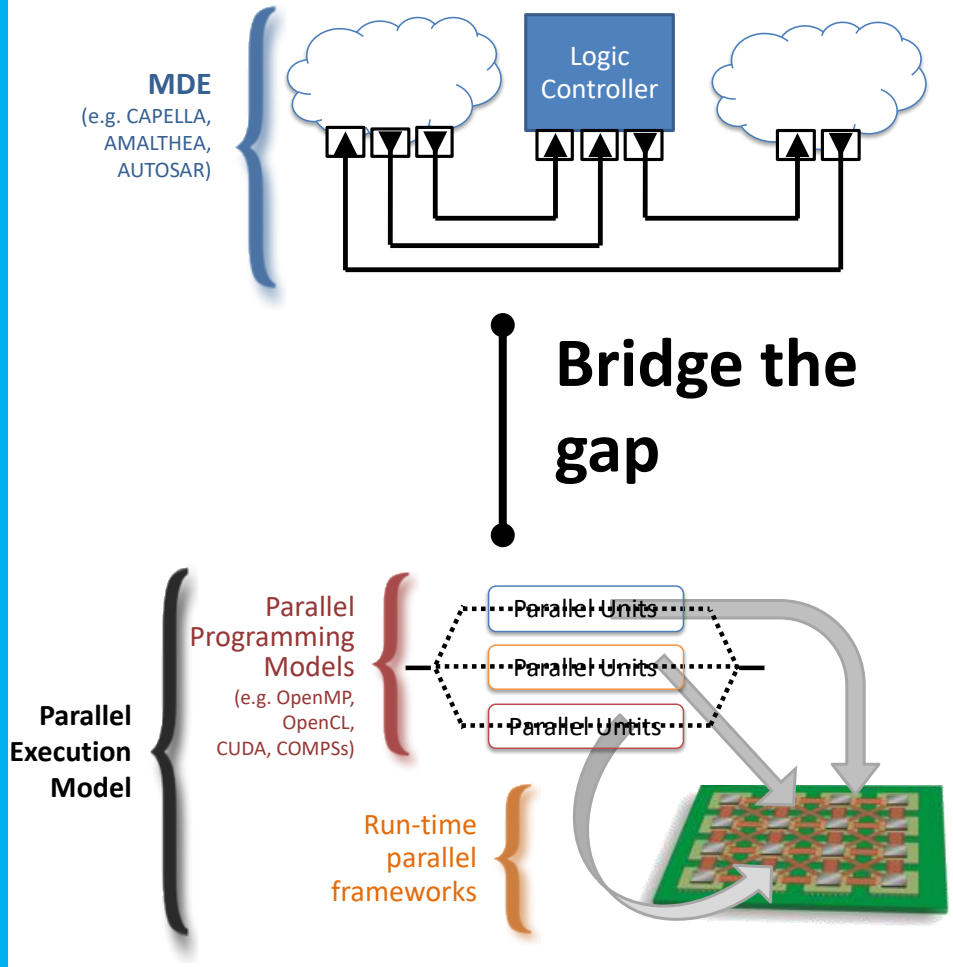
Gap between the MDE used for CPS and the PPM supported by parallel platforms



Parallel Programming Models (PPM)

1. Mandatory for **SW productivity** in terms of
   - Programmability: Parallel abstraction while hiding HW complexities
   - Portability: Compatibility multiple HW platforms
   - Performance: Exploiting parallel capabilities of underlying HW
2. **Efficiet offloading** to HW acceleration devices for an energy-efficient parallel execution

Motivation



1. **Synthesis methods** for an efficient generation of parallel source code, while keeping non-functional and composability guarantees
2. **Run-time parallel frameworks** that guarantee system correctness and exploit the performance capabilities of parallel architectures
3. **Integration** of parallel frameworks into MDE frameworks

AMPERE Use-cases

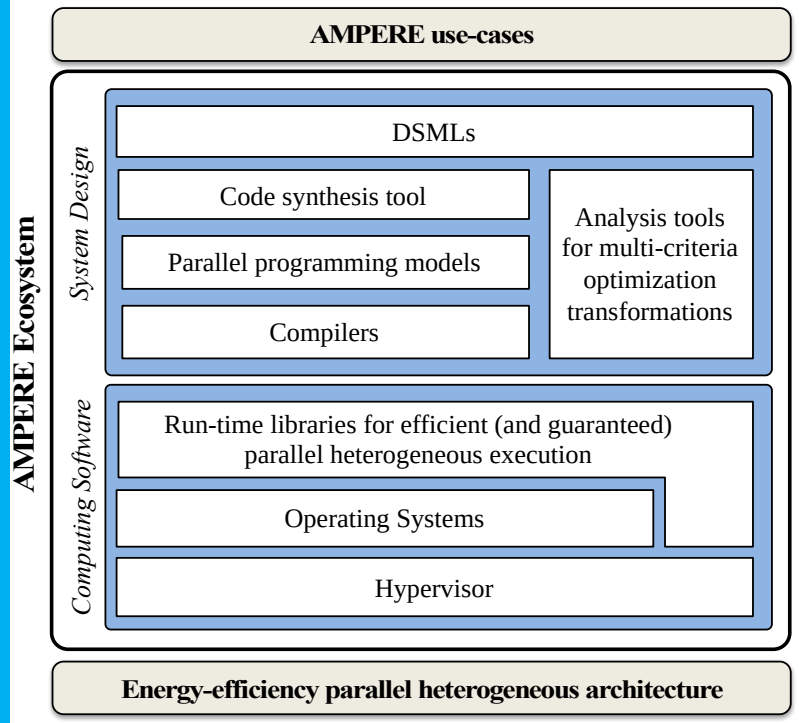
Obstacle Detection and Avoidance System (ODAS)

- ADAS functionalities based on data fusion coming from tram vehicle sensors

Predictive Cruise Control (PCC)

- Extends Adaptive Cruise Control (ACC) functionality by calculating the vehicle's future velocity curve using the data from the *electronic horizon*
- Improve fuel efficiency (in cooperation with the powertrain control) by configuring the driving strategy based on data analytics and AI

AMPERE Concept

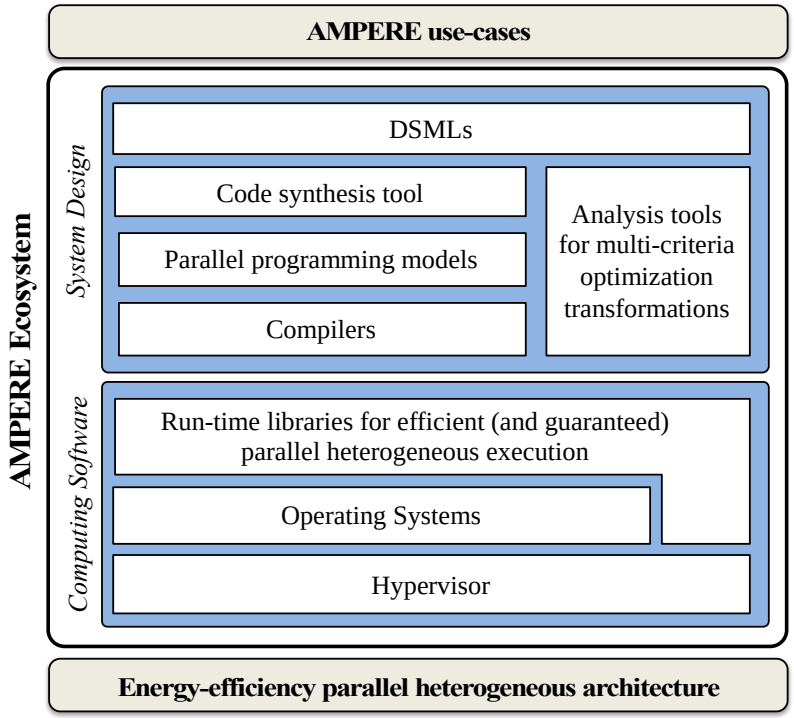


Develop a novel software architecture capable of:

1. Capturing the component definition and non-functional requirements for the system model and transform it to parallel constructs
2. Fulfillment of non-functional properties described in the CPSoS description
   - Energy-efficiency, safety and cyber-security, real-time response, resiliency and fault-tolerance
3. Efficient usage of advance parallel and heterogeneous embedded architectures

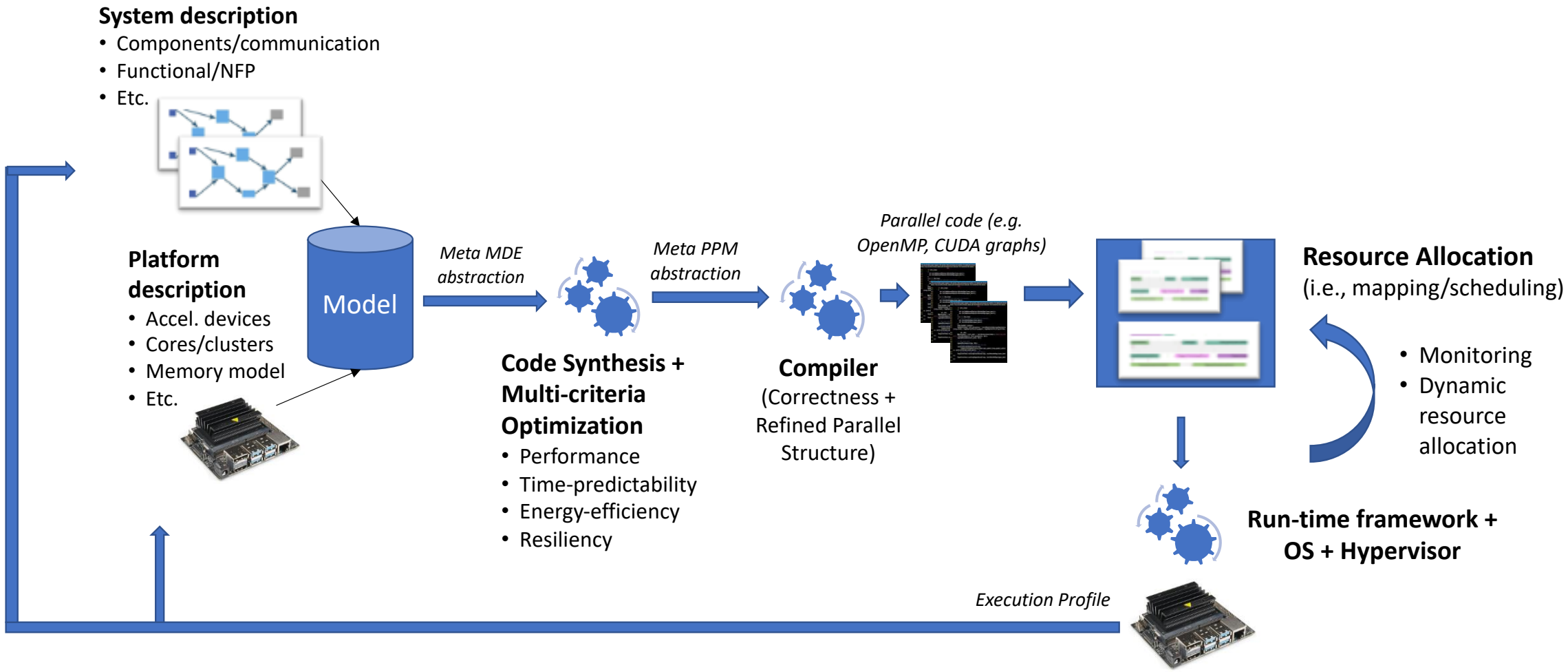
Productivity = + **Programmability** + **Portability/Scalability** + **Performance**

AMPERE Concept



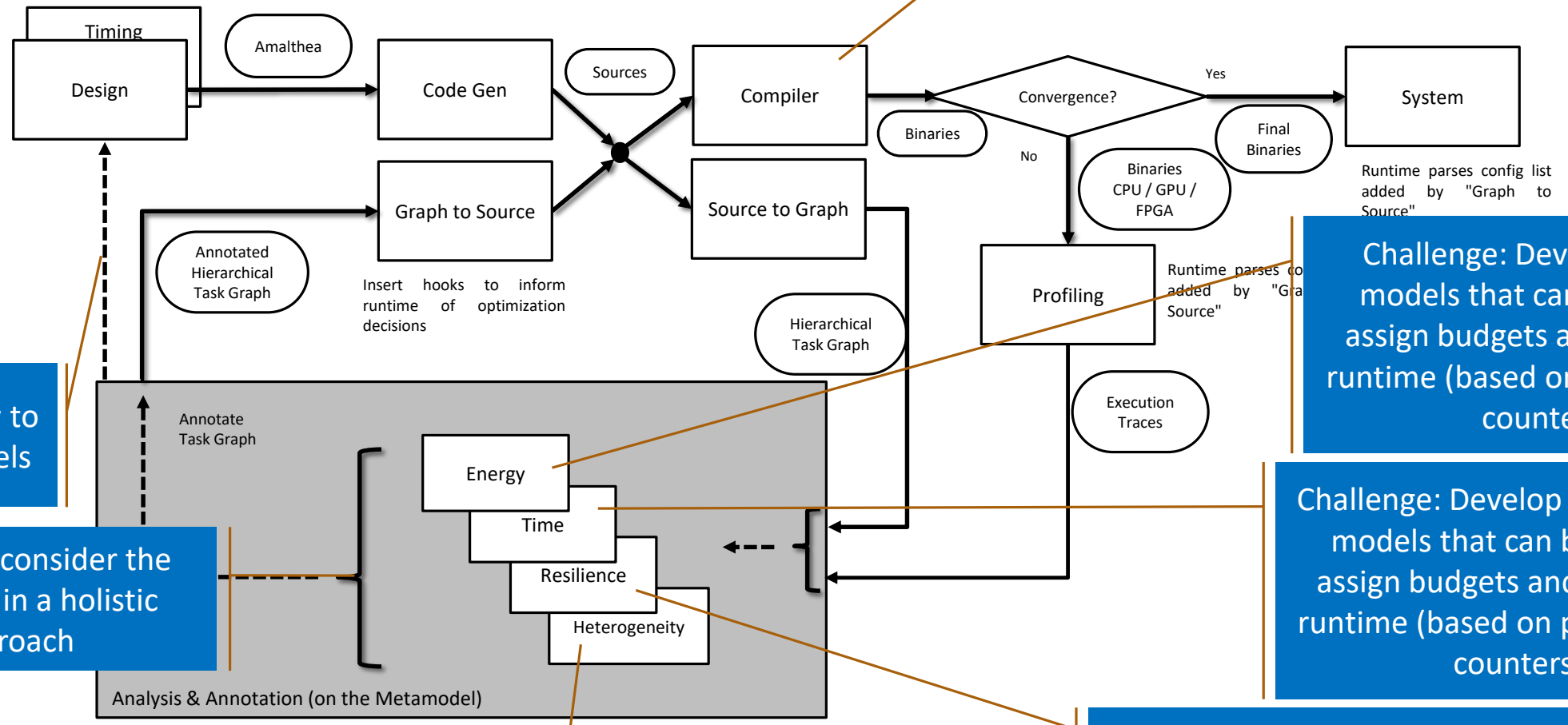
Software Layer	Tool
DSMLs	AUTOSAR
	AMALTHEA
	CAPELLA
Parallel programming models	OpenMP
	CUDA
	OpenCL
	COMPSs
Artificial Intelligence	TensorFlow
Code synthesis tools	Synthesis tools
Analysis and testing tools	NFP analysis
Compilers and hardware synthesis tools	Mercurium
	GCC/LLVM
	Vivado
Run-time libraries	GOMP
	KMP
	Vivado
Operating systems	Linux
	ERIKA Enterp.
Hypervisors	PikeOS

AMPERE Workflow Overview



AMPERE Resource Challenges

Challenge: how to select the “best” resource mapping?



Challenge: how to annotate models

Challenge: Develop energy models that can be used to assign budgets and predict at runtime (based on performance counters)

Challenge: consider the resources in a holistic approach

Challenge: Develop interference models that can be used to assign budgets and predict at runtime (based on performance counters)

Challenge: Consider CPU, FPGA, GPU

Challenge: Software redundancy to cope with hw faults

Summary of research challenges

- Predictability and Performance
  - Compute continuum
  - Heterogenous platforms
- Multi-criteria optimization
  - Time
  - Energy
  - Communication
  - Security
  - Reliability
- Vertical stack
  - Model-driven development
  - Parallel programming abstractions
  - Resource allocation/reservation
  - Scheduling
  - Monitoring
  - Platforms

Challenges in Resource Management in the ELASTIC and AMPERE European Projects

CERCIRAS COST Action Workshop

September 2-3, 2021

isep Instituto Superior de Engenharia do Porto

P.PORTO