TRANSPARENT APPLICATION DEPLOYMENT IN A SECURE, ACCELERATED AND COGNITIVE CLOUD CONTINUUM

USTUTT Audit – Oct 2023

High Performance Computing Center, Stuttgart-HLRS

Presentation Outline

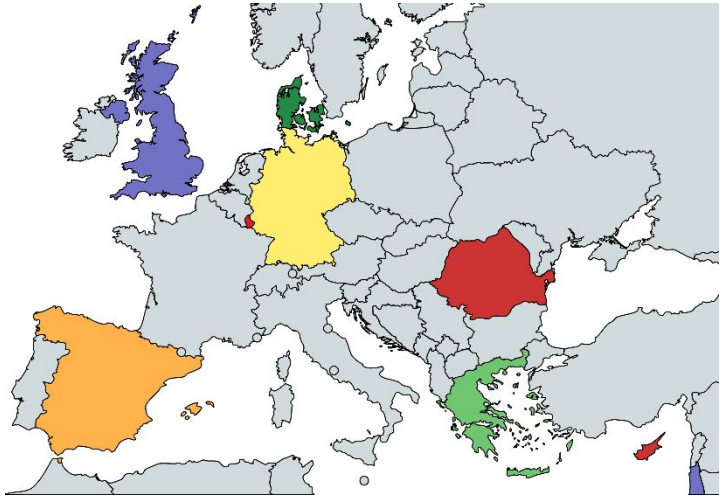
- Project Administrative Information
- Motivation
- Vision and Use Cases
- Objectives
- Work Packages, Workplan, Timeline
- Major Achievements
- Towards the realization of the Objectives
- Conclusions

Project Administrative Information (1/2)

- **Project Name**: Transparent application development in a secure, accelerated and cognitive cloud continuum
- **Call identifier**: ICT-40-20 on “Cloud Computing: towards a smart cloud computing continuum”
- **Project Type**: Research & Innovation Action (RIA)
- **Grant Agreement Number**: 101017168
- **Project Coordinator**: Institute of Communication and Computer Systems – ICCS
- **Duration**: 36 months (01/01/2021 – 31/12/2023)
- **Funding from the EC**: 4,343,180 €
- **Project Officer**: Javier Mata

Project Administrative Information (2/2)

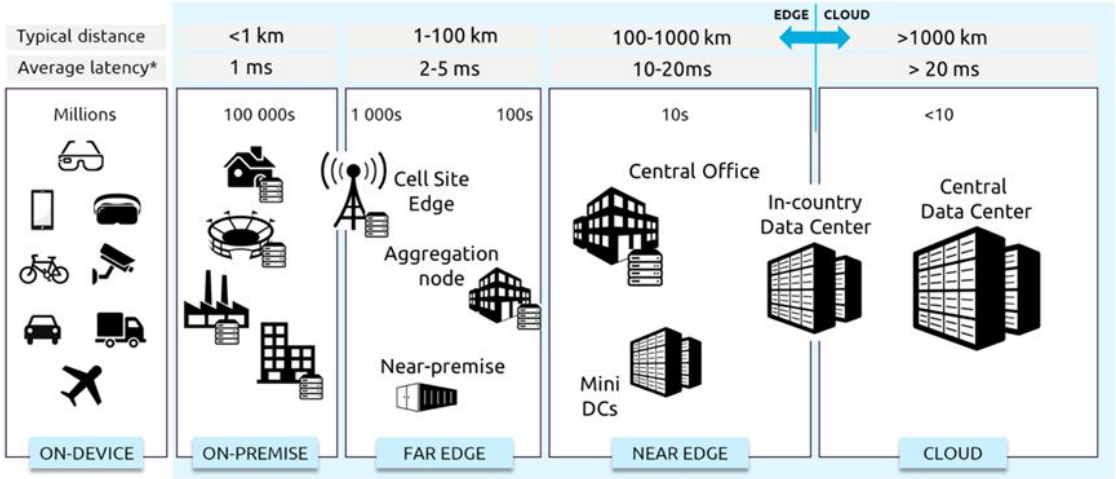
- 11 Partners
  - 5 research institutes: **ICCS, USTUTT, AUTH, IDEKO, UVT**
  - 4 SMEs: – **CC, INB, INNOV, NBFC**
  - 2 industrial partners: **MLNX, INTRA**



- Partner Roles
  - Edge, acceleration and security: **MLNX, CC, AUTH, NBFC**
  - Application profiling, HPC, resource orchestration and telemetry: **ICSS, USTUTT, INNOV, UVT**
  - Integration and use cases: **INTRA, CC, INB, IDEKO**

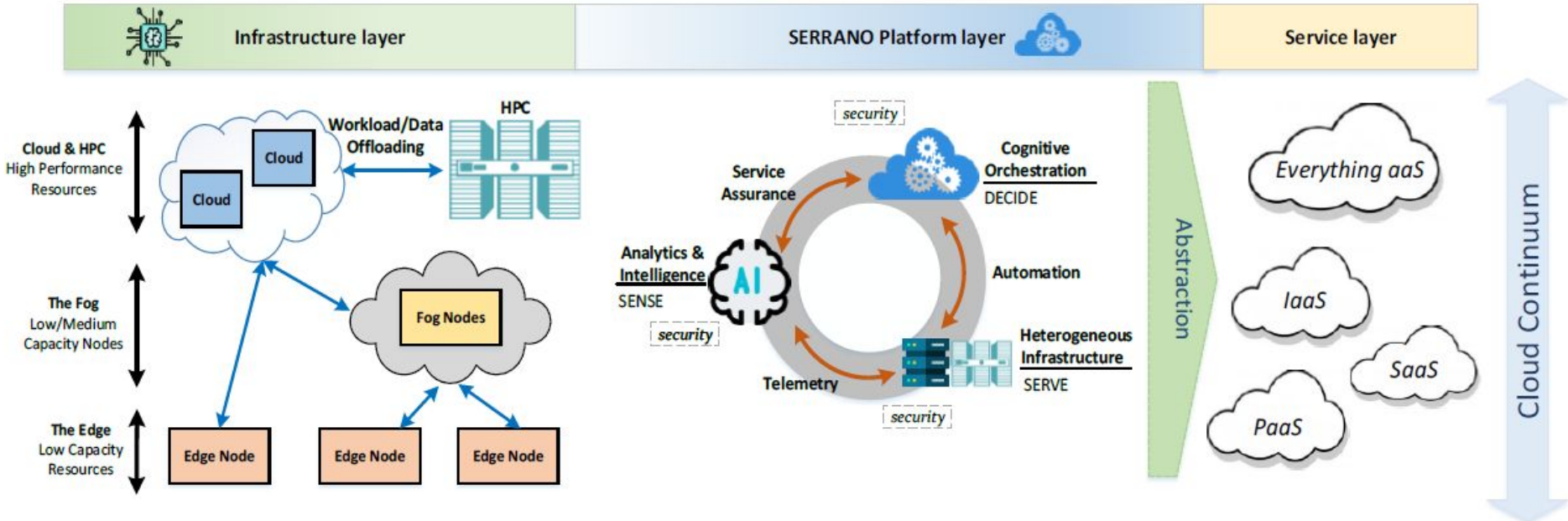
Motivation

- Traditionally the cloud computing resources handle the processing and storage workload of monolithic applications
  - This model cannot adequately address the strict and dynamic requirements of emerging new applications
- Edge computing is a computing paradigm that brings computation and data storage closer to the sources of data
  - Multiple resources at edge with heterogeneous characteristics (CPUs, GPUs, Mini DC, FPGA, ASICs, HPC, Storage nodes)



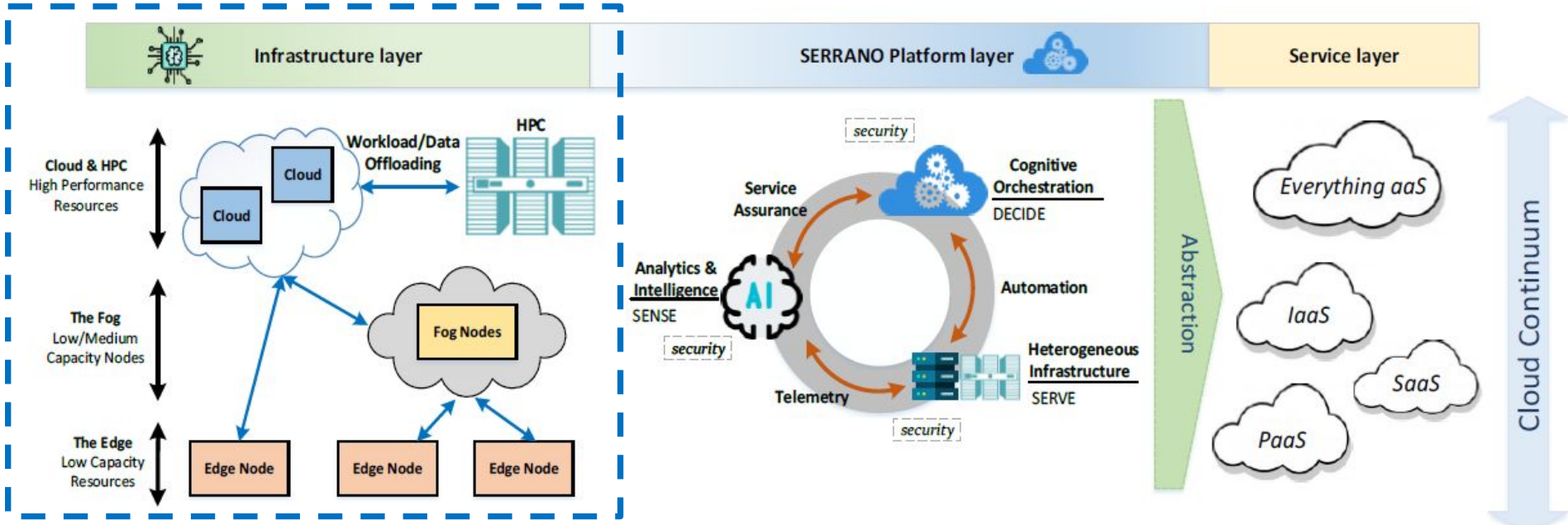
SERRANO vision

- SERRANO targets disaggregated and federated infrastructures consisting of edge, cloud and HPC resources through the development of the SERRANO Platform
- An abstraction layer will automate the operation and full exploitation of the available diverse resources, enabling transparent application deployment
- SERRANO platform will be a self-optimizing system that continuously adapts, over an infinite time horizon control loop
- SERRANO platform will enable the creation of a plethora of IaaS, PaaS, SaaS services targeting today's and future's cloud/edge/HPC computing markets



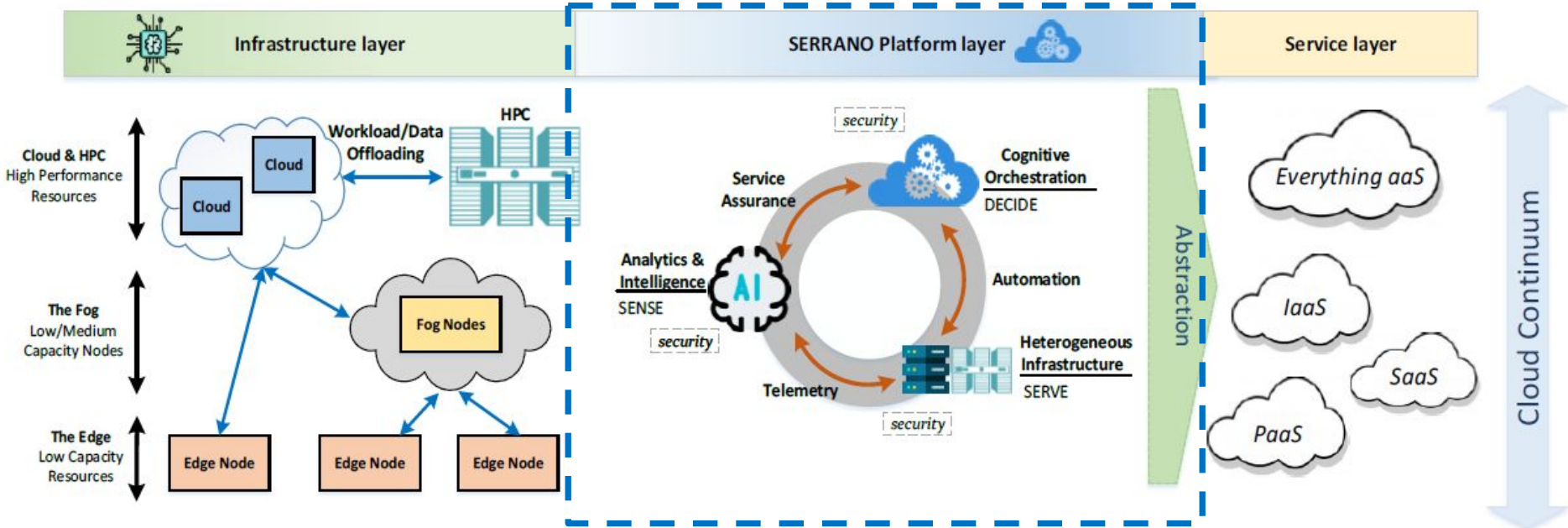
SERRANO vision

- SERRANO aims to address disaggregated and federated infrastructures that include edge, cloud, and HPC resources by developing the SERRANO Platform
- An abstraction layer will automate operations and maximize the utilization of the diverse resources, facilitating seamless application deployment.
- SERRANO platform will be a self-optimizing system that continuously adapts, over an infinite time horizon control loop
- SERRANO platform will enable the creation of a plethora of IaaS, PaaS, SaaS services targeting today's and future's cloud/edge/HPC computing markets



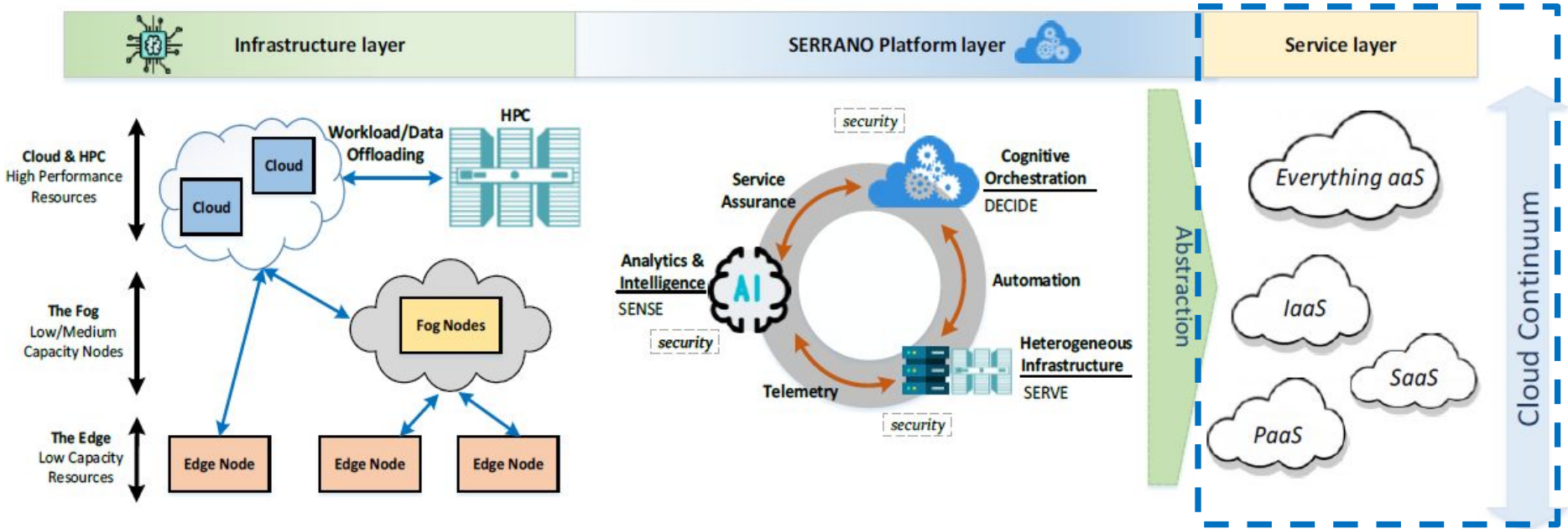
SERRANO vision

- SERRANO targets disaggregated and federated infrastructures consisting of edge, cloud and HPC resources through the development of the SERRANO Platform
- An abstraction layer will automate the operation and full exploitation of the available diverse resources, enabling transparent application deployment
- The SERRANO platform will function as a self-optimizing system, continuously adapting through an infinite time horizon control loop
- SERRANO platform will enable the creation of a plethora of IaaS, PaaS, SaaS services targeting today's and future's cloud/edge/HPC computing markets



SERRANO vision

- SERRANO targets disaggregated and federated infrastructures consisting of edge, cloud and HPC resources through the development of the SERRANO Platform
- An abstraction layer will automate the operation and full exploitation of the available diverse resources, enabling transparent application deployment
- SERRANO platform will be a self-optimizing system that continuously adapts, over an infinite time horizon control loop
- SERRANO platform will enable the creation of a plethora of IaaS, PaaS, SaaS services targeting today's and future's cloud/edge/HPC computing markets



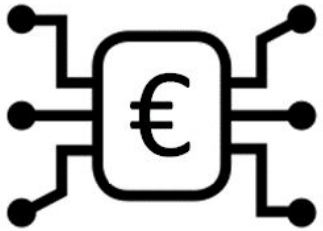
SERRANO objectives

- **Objective 1**: Define an intent-driven paradigm of federated infrastructures consisting of edge, cloud and HPC resources
- **Objective 2**: Develop security and privacy mechanisms for accelerated encrypted storage over heterogeneous and federated infrastructures
- **Objective 3**: Provide acceleration and energy efficiency at the edge and cloud
- **Objective 4**: Cognitive resource orchestration and transparent application deployment over edge/fog-cloud/HPC infrastructures
- **Objective 5**: Demonstrate the capabilities of the secure, disaggregated and accelerated SERRANO platform in supporting highly-demanding, dynamic and safety-critical applications.

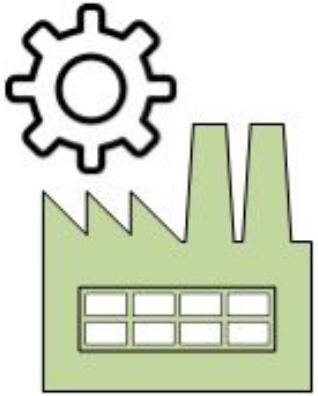
SERRANO (3) use cases

- **Secure Storage**
  - Provide secure and high-performance storage at the edge
  - Integrate SERRANO with a multi-cloud storage service

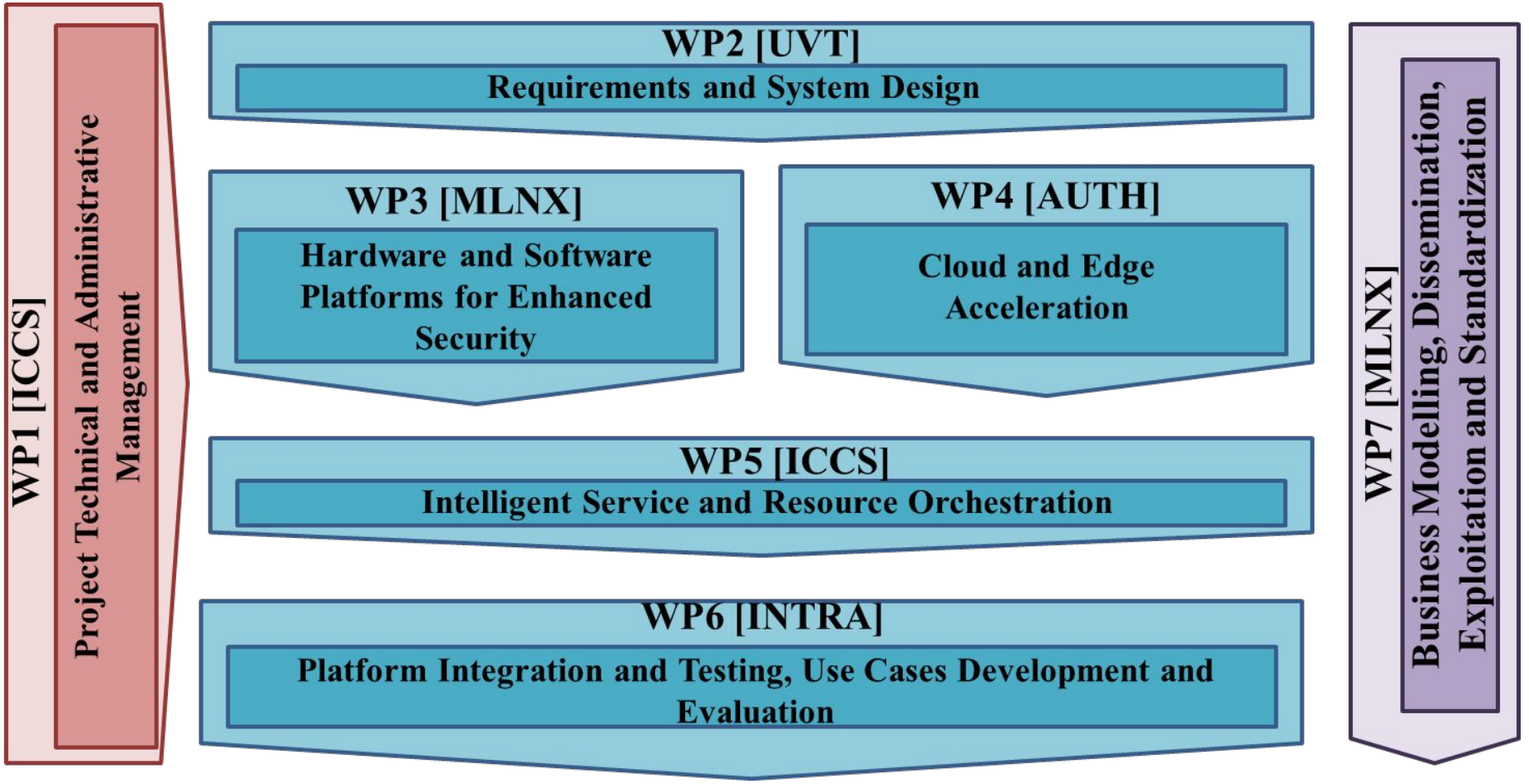
- **High-performance Fintech Analysis**
  - Apply AI and ML algorithms in financial operations
  - SERRANO will provide security and intelligent fintech app deployment



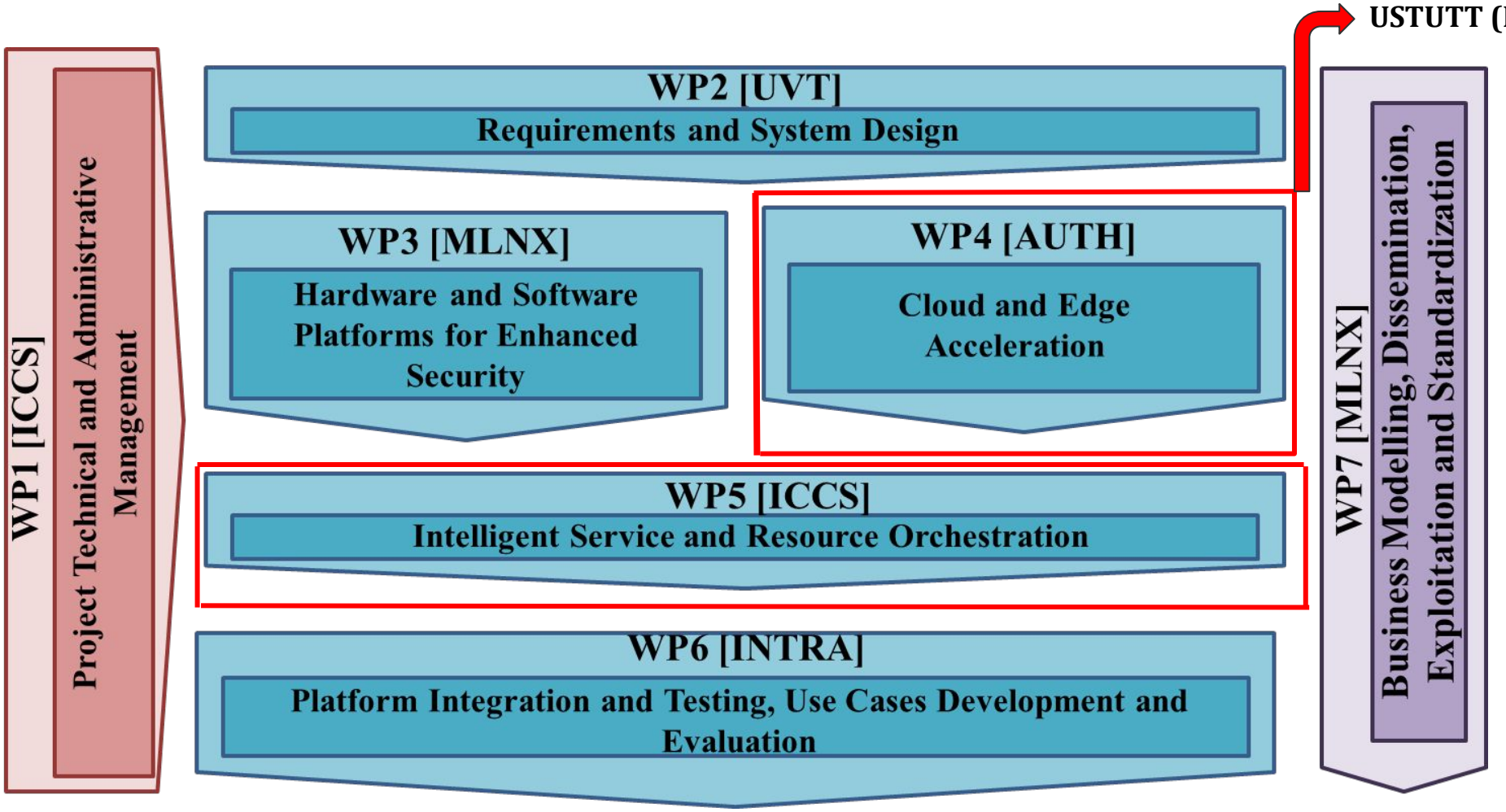
- **Machine Anomaly Detection in Manufacturing Environments**
  - Detect machine anomalies in real-time
  - SERRANO will orchestrate computations and data from high-frequency machine sensors



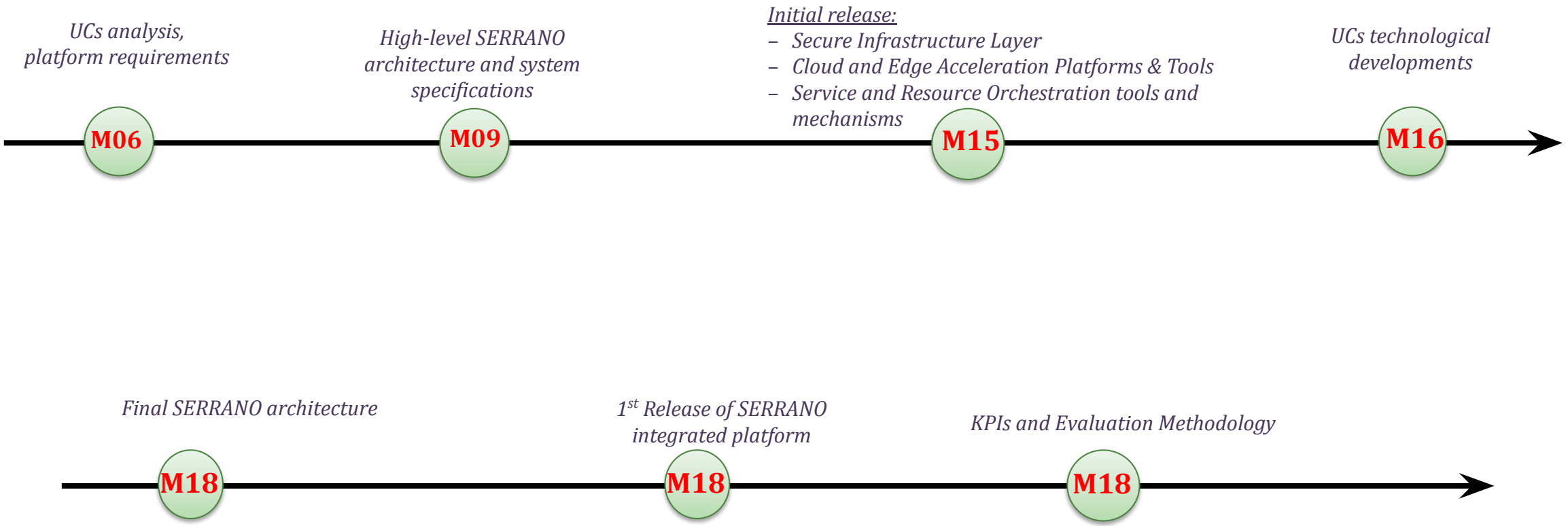
Work Packages and Workplan



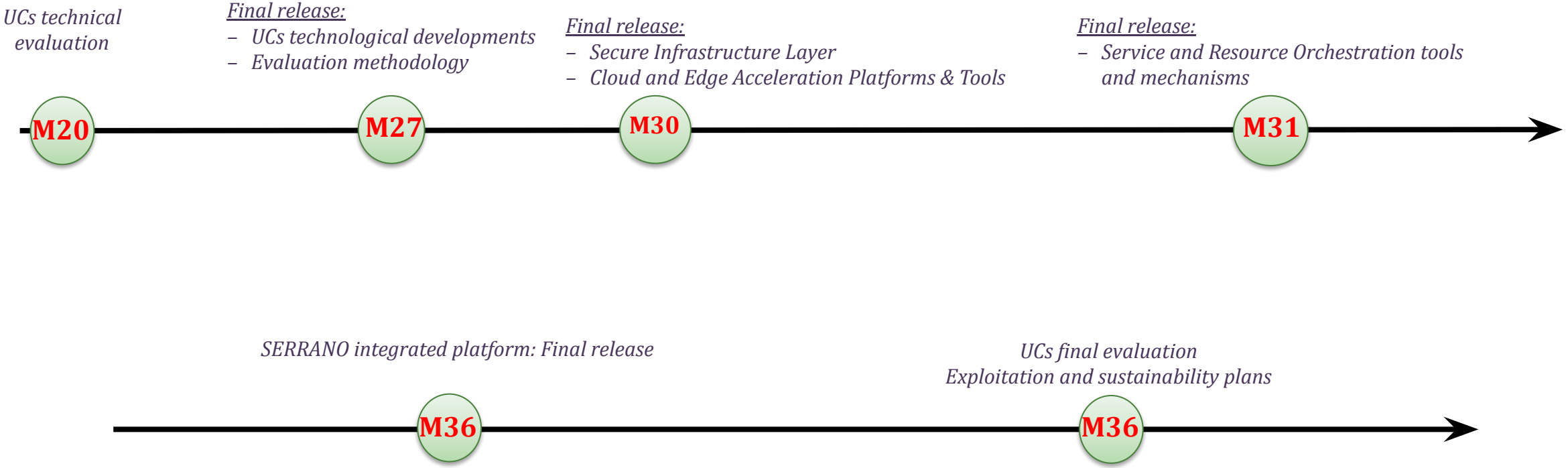
Work Packages and Workplan



Major Achievements so far (M1-M18)

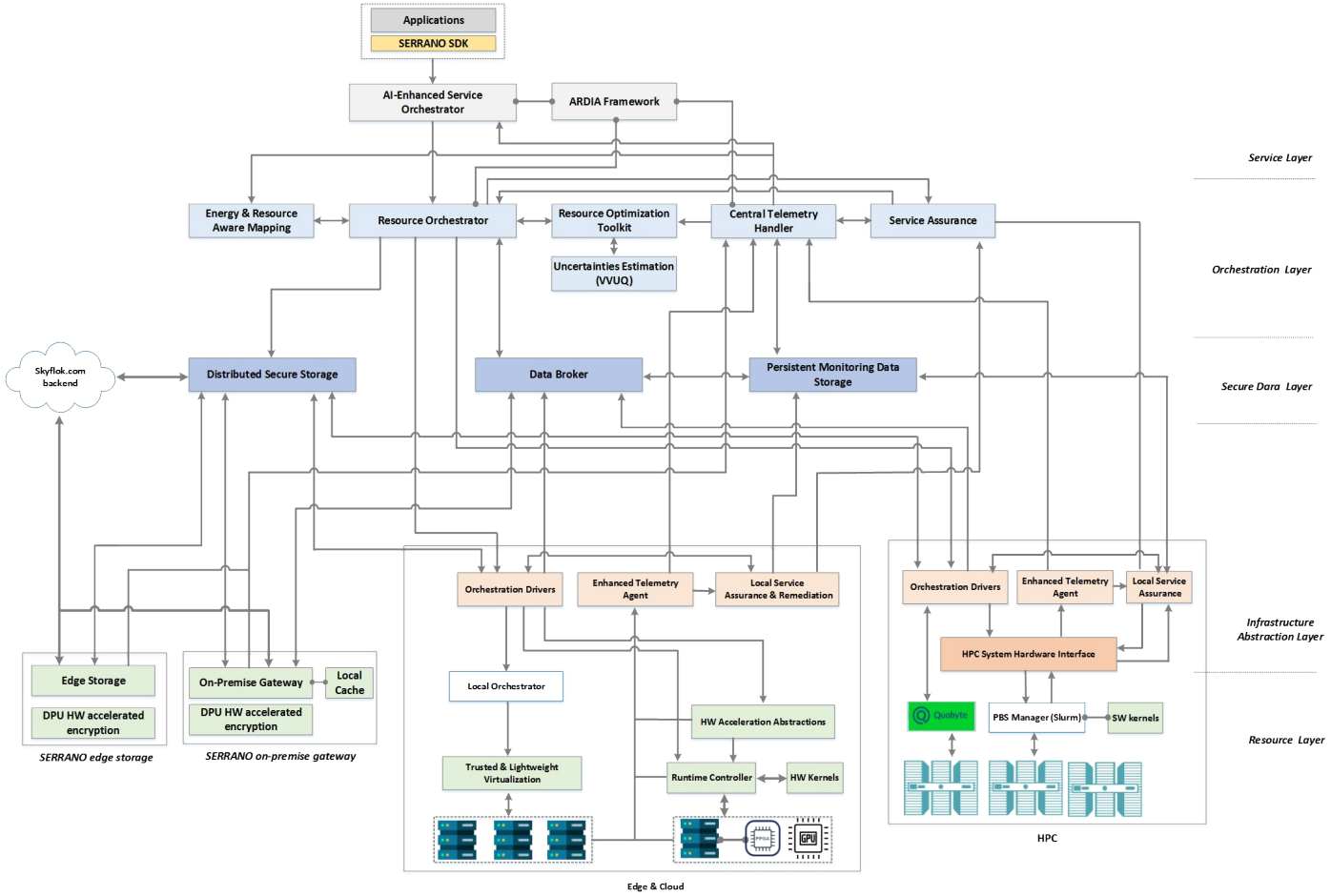


Major Achievements so far (M19-M33)



Towards the achievement of Objective - 1

Define an intent-driven paradigm of federated infrastructures consisting of edge, cloud and HPC resources

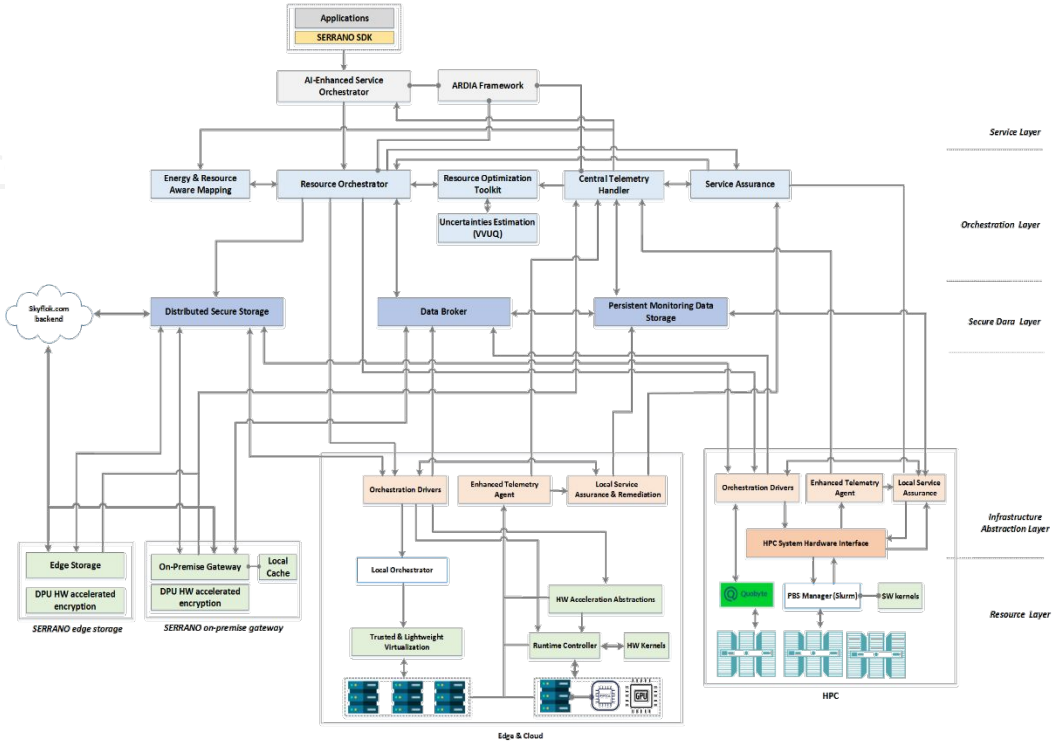


Towards the achievement of Objective - 1

Define an intent-driven paradigm of federated infrastructures consisting of edge, cloud and HPC resources

- Performed an in-depth analysis of the current advances in SERRANO topics of interest
- Identified the SERRANO platform's high-level requirements
- Specified Key Performance Indicators (KPIs)
- Defined SERRANO multi-layer overall architecture
- These guide all the technical developments towards the r of the SERRANO use cases (UCs)

- **Related WP**: WP2 – M18
- **Partners**: All
- **Deliverables submitted**: D2.1, D2.2, D2.3, D2.4, D2.5

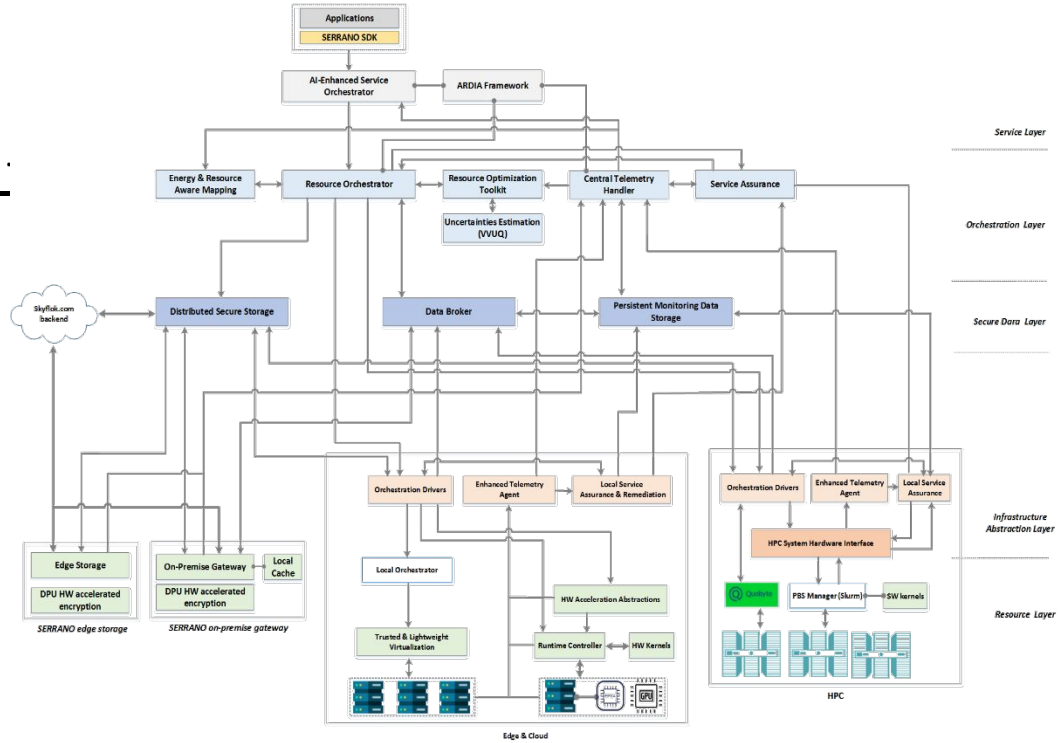


Towards the achievement of Objective - 1

Define an intent-driven paradigm of federated infrastructures consisting of edge, cloud and HPC resources

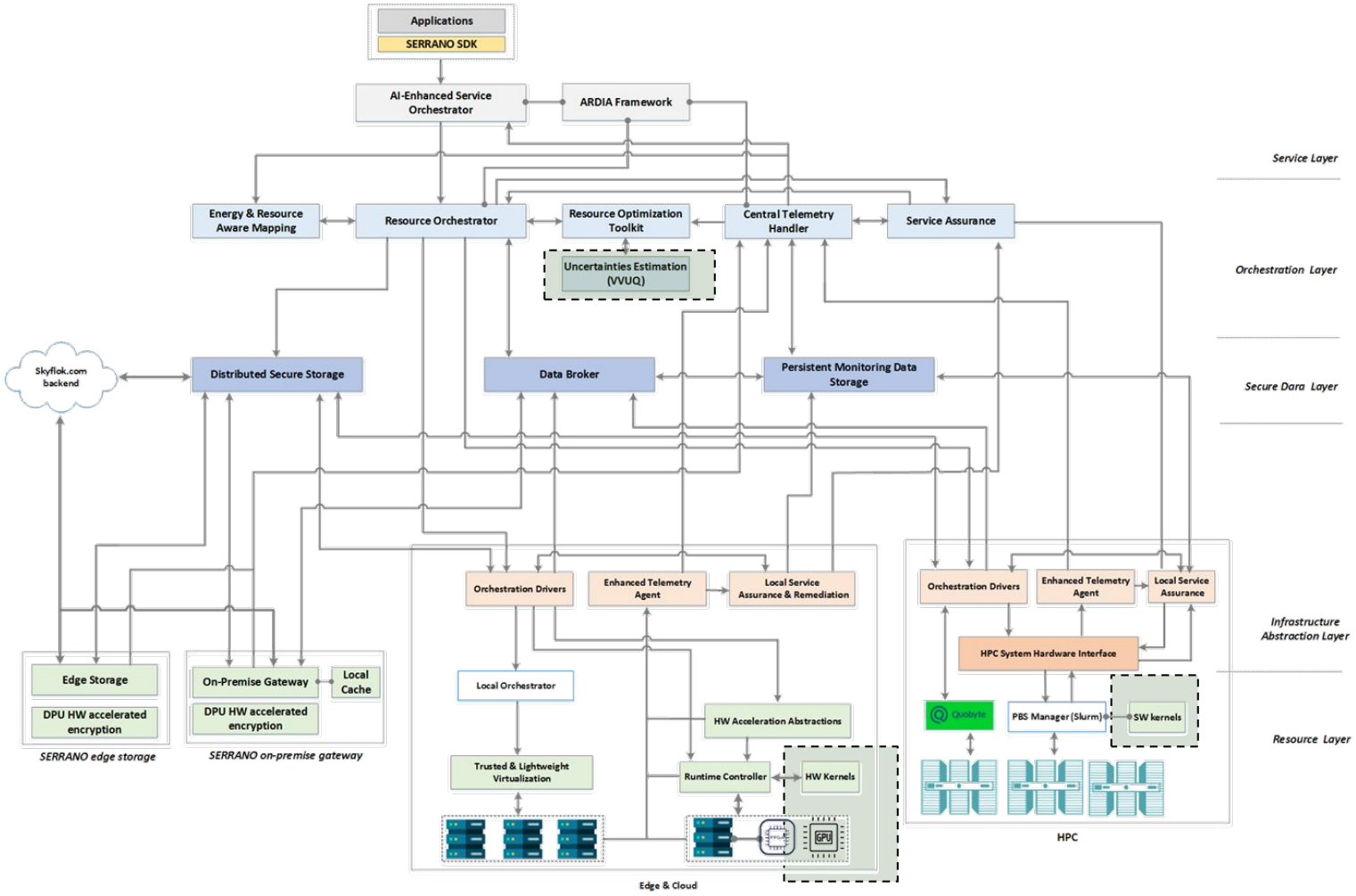
- Performed an in-depth analysis of the current advances in SERRANO topics of interest
- Identified the SERRANO platform's high-level requirements
- Specified Key Performance Indicators (KPIs)
- Defined SERRANO multi-layer overall architecture
- These guide all the technical developments that will lead to the realization of the SERRANO use cases (UCs)

- **Related WP**: WP2 – M18
- **Partners**: All
- **Deliverables submitted**: D2.1, D2.2, D2.3, D2.4, D2.5



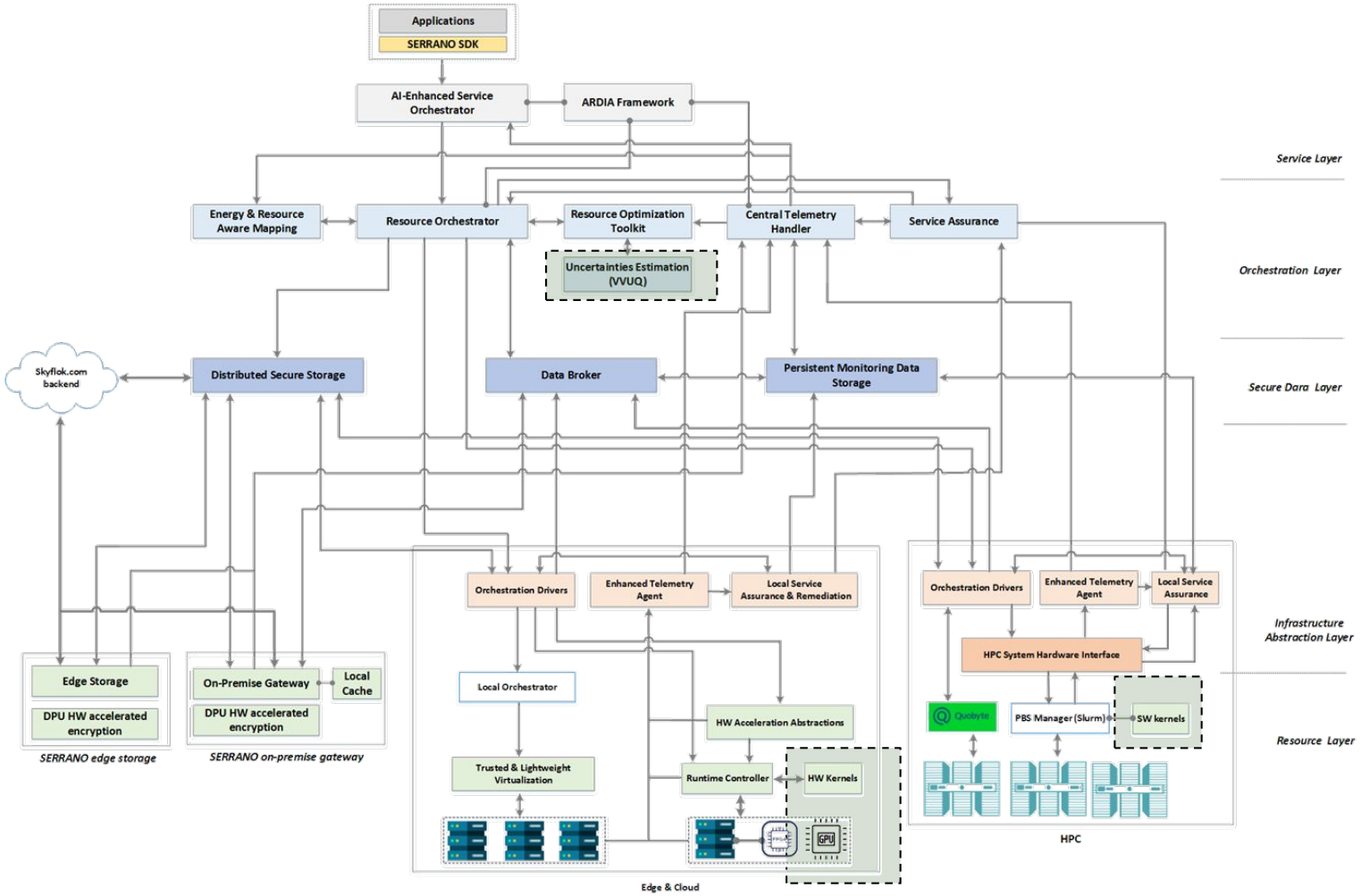
Towards the achievement of Objective - 4

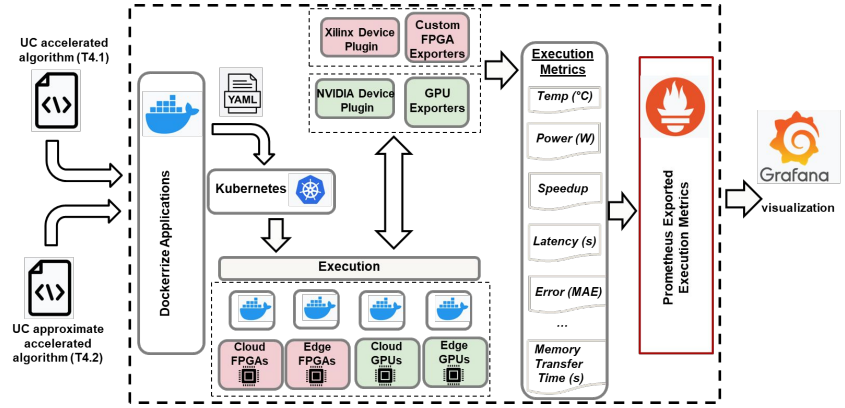
Provide acceleration and energy efficiency at the edge and cloud

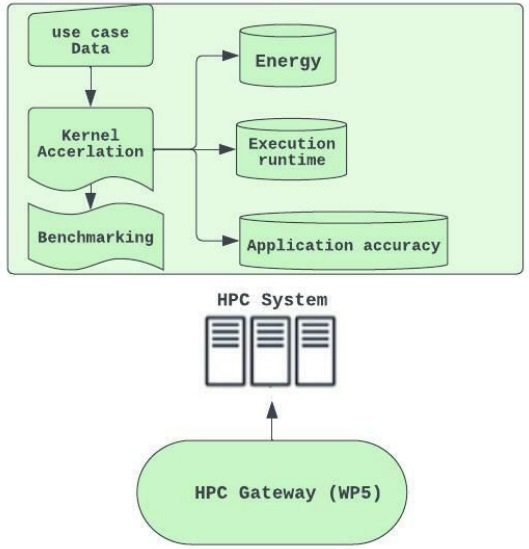


Towards the achievement of Objective - 4

Provide acceleration and energy efficiency at the edge and cloud



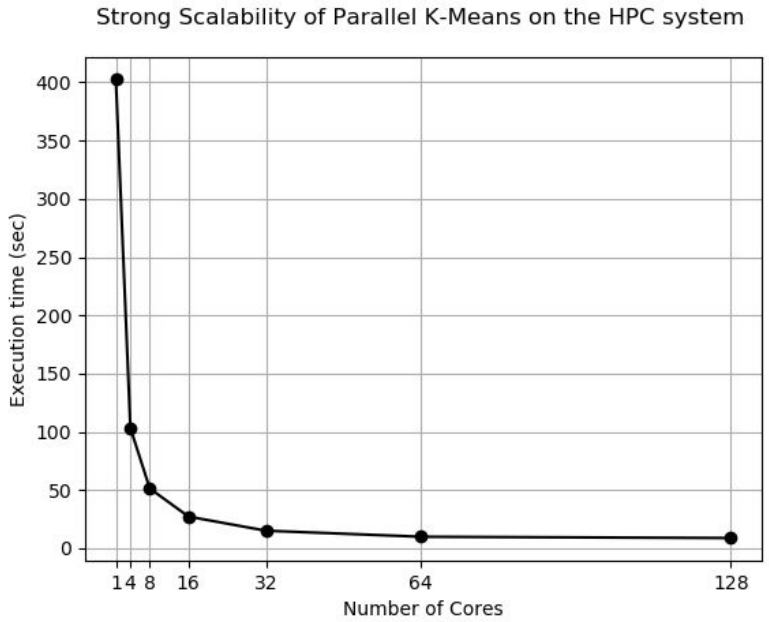




Towards the achievement of Objective - 4

Provide acceleration and energy efficiency at the edge and cloud

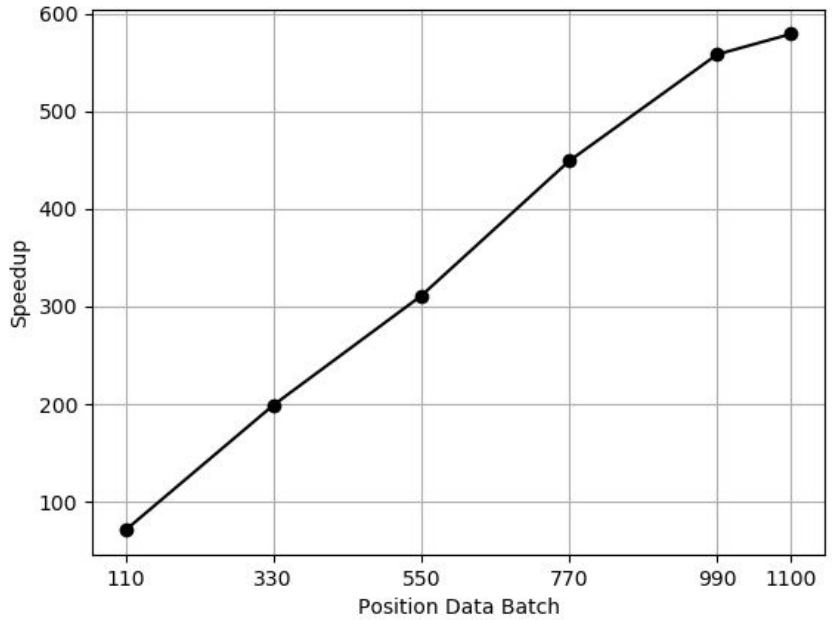
- Developed accelerated kernels that utilize approximation techniques, and transprecision techniques
- 12 kernels accelerated in parallel shared and distributed memory for HPC system using MPI/OpenMP framework
  - **Speedup of parallel K-Means with 110 time series signals achieves 40X improvement**



Towards the achievement of Objective - 4

Provide acceleration and energy efficiency at the edge and cloud

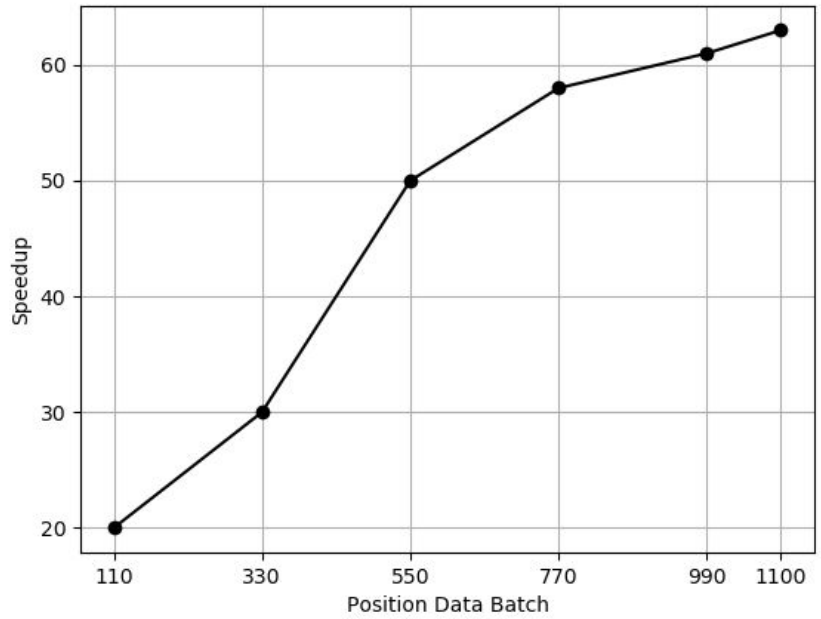
- Developed accelerated kernels that utilize approximation techniques, and transprecision techniques
- 12 kernels accelerated in parallel shared and distributed memory for HPC system using MPI/OpenMP framework
  - **A maximum speed up from 71x up to 579x**



Towards the achievement of Objective - 4

Provide acceleration and energy efficiency at the edge and cloud

- Developed accelerated kernels that utilize approximation techniques, and transprecision techniques
- 12 kernels accelerated in HPC system using MPI/OpenMP framework
  - **A maximum speed up from 71x up to 579x**
  - **A Energy gains ranging from 10x up to 128x**



Towards the achievement of Objective - 4

Provide acceleration and energy efficiency at the edge and cloud

- Developed accelerated kernels that utilize approximation techniques, and transprecision technques
- 12 kernels accelerated in HPC system using MPI/OpenMP framework
  - **Energy gains ranging from 10x up to 128x**
  - **A minimum speed up of 7.4x in the acceleration of the kernels**

- A framework was developed to adapt dynamically the precision level and apply, in a coordinated manner, approximate computations and approximate data transfers

- **Related WP**: WP4 – M30
- **Partners**: AUTH, USTUTT
- **Deliverables submitted**: D4.1, D4.2, D4.3, D4.4

Towards the achievement of Objective - 5

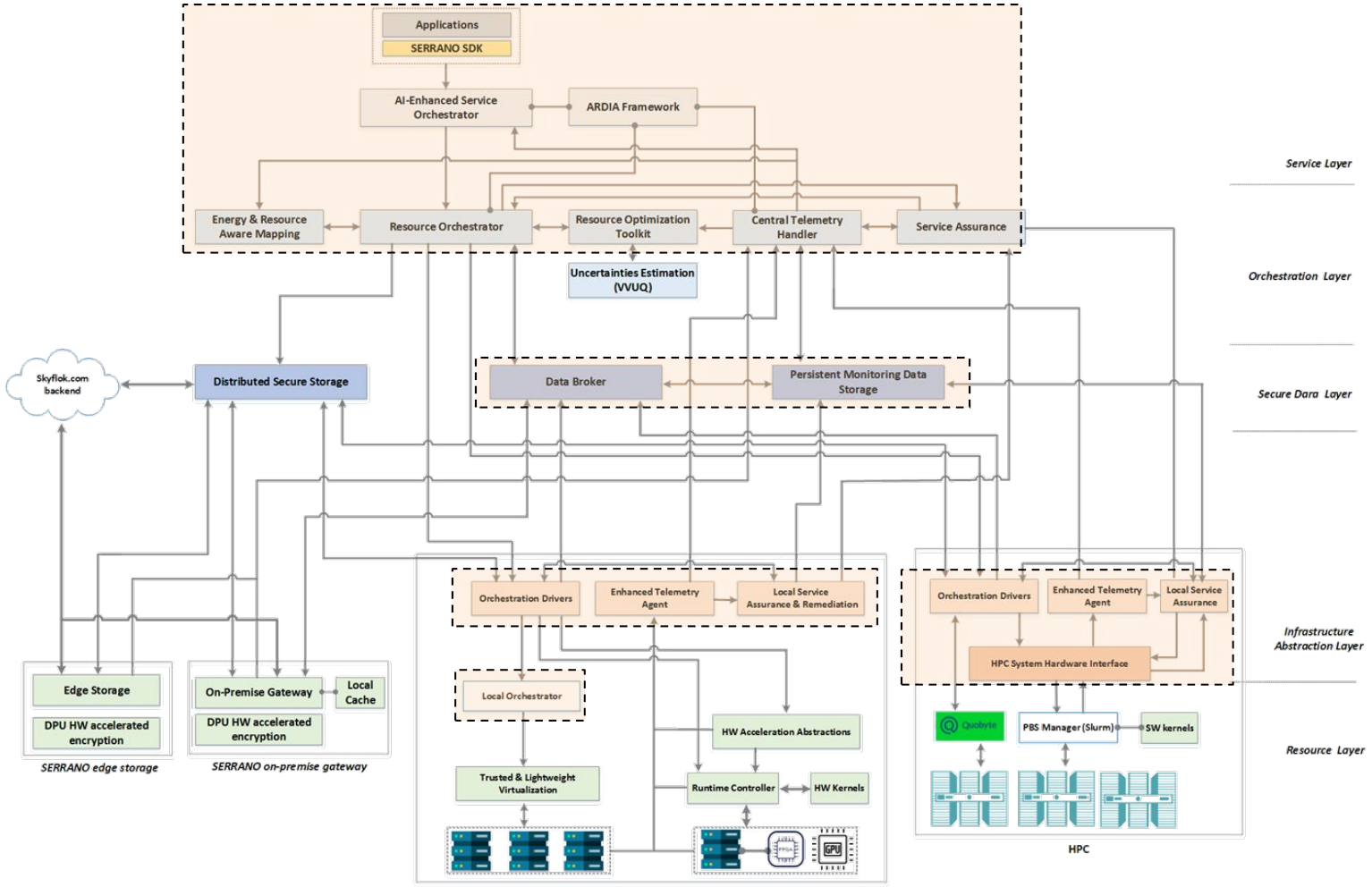
Cognitive resource orchestration and transparent application deployment over edge/fog-cloud/HPC infrastructures

- Developed Verification, Validation and Uncertainty Quantification (VVUQ)
- Integrate the automated Benchmarking and providing the metadata
- Applying AI method to approximate performance, and energy gain

- Algorithms for resource allocation and for analyzing monitoring data were developed
- A power measurement system was integrated into the EXCESS test HPC cluster

Towards the achievement of Objective - 5

Cognitive resource orchestration and transparent application deployment over edge/fog-cloud/HPC infrastructures



Thank you!

Javad Fadaie Ghotbi

High Performance Computing Center, Stuttgart – HLRS

Visit our website!
https://ict-serrano.eu

The research leading to these results has received funding from the EC HORIZON 2020 SERRANO project, under grant agreement number 101017168