https://www.ai-sprint-project.eu/

Danilo Ardagna

Project Coordinator

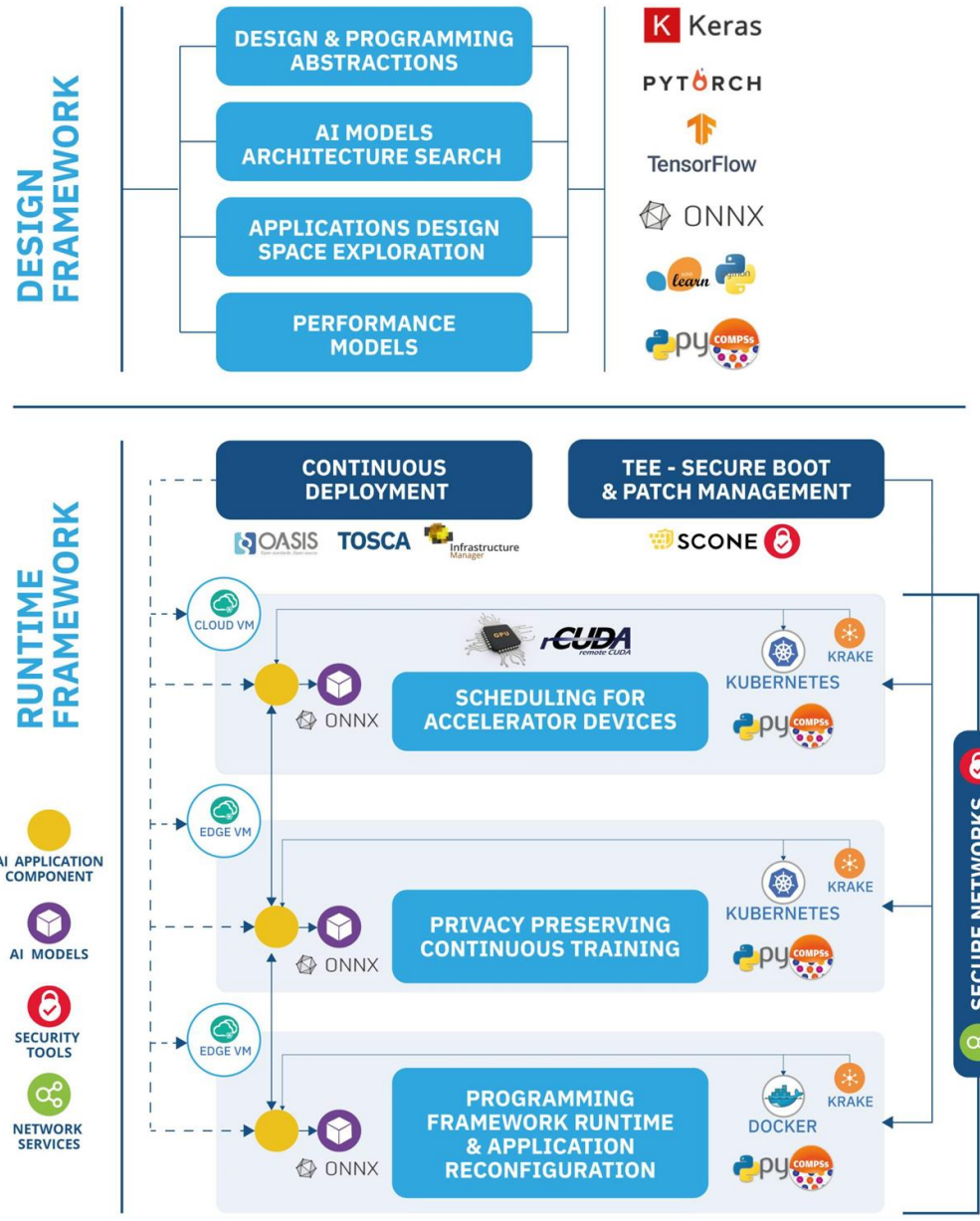
Politecnico di Milano, Italy

danilo.ardagna@polimi.it

AI-SPRINT: challenges & objectives

- AI is hungry
  - Expertise, trial & error
- Continuum heterogeneity
  - Deployment complexity
  - Performance & costs
- Need to trust the cloud and storage provider
- Privacy of data and models

- Simplified programming models
- Automated deployment and dynamic reconfiguration
- Secure execution of AI applications
- Specialized building blocks for privacy preservation, distributed training, and NAS



https://gitlab.polimi.it/ai-sprint/

AI-SPRINT: lessons learned & success stories

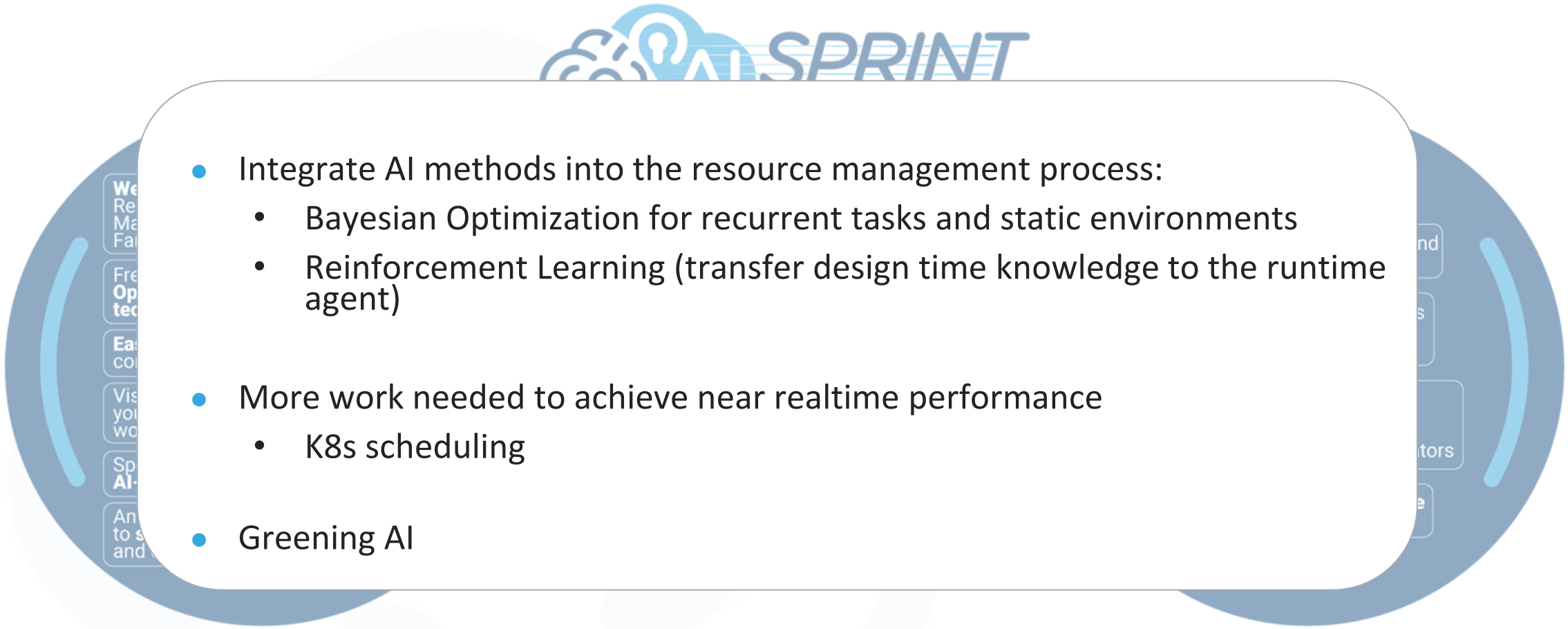
- Rely on standards when possible even you pay in performance
- Moving to FaaS is beneficial, issues with cold start and K8s scheduler
- The edge is the bottleneck, size the system for peak load, resource scaling is crucial
- NAS is hungry, trade-off accuracy and performance

- Cost savings wrt first principle heuristics about 30%
- Deployment tools and core monitoring available at the European Open Science Cloud (EOSC)
- Trusted Execution environment protects against root-level attacks with about 25% overhead
- POPNAS training time speedup about 3.5x, SGDE improves performance 3-10%

Developing an automated system for personalised stroke risk assessment and prevention.

Creating an infrastructure that reduces downtime and revenue losses caused by degenerative asset performance.

AI-SPRINT: Recommendations for the Future



- Integrate AI methods into the resource management process:
  - Bayesian Optimization for recurrent tasks and static environments
  - Reinforcement Learning (transfer design time knowledge to the runtime agent)
- More work needed to achieve near realtime performance
  - K8s scheduling
- Greening AI

https://ai-sprint-project.eu/ai-sprint-alliance

CHARITY

Cloud for Holography and Augmented RealITY

Concertation meeting

Alessandro Bassi

Eurescom GmbH

Project Presentation and Main Challenges

- Advanced media applications enabling immersive communication are becoming ubiquitous in our lives, and there is a global trend to adopt virtual solutions to support day-to-day business operations, social events, and general lifestyle.
- A subset of highly attractive innovative immersive communication applications includes Virtual Reality, Augmented Reality, and Holography (known as XR), but they do not come without their share of challenges and requirements.

Success Stories

- Achieve autonomicity in orchestrating network, compute, and storage resources in hybrid edge/cloud infrastructures.
- Provide holistic support for the orchestration of advanced media solutions.
- Develop highly interactive and collaborative services and applications.

Recommendation for the future

- There is great interest from application developers and service providers in XR and the holography domain to take benefit of 5G network capabilities
- Benefits are seen in the outsourcing of resource consuming service provider and end user device operations to the (edge) cloud, but also for enabling to reach a wider (global) audience
- A clear and for a non-telecom expert easy to understand template (or interface) must be defined in which the service developer can describe the service requirements on the networks.

CHARITY

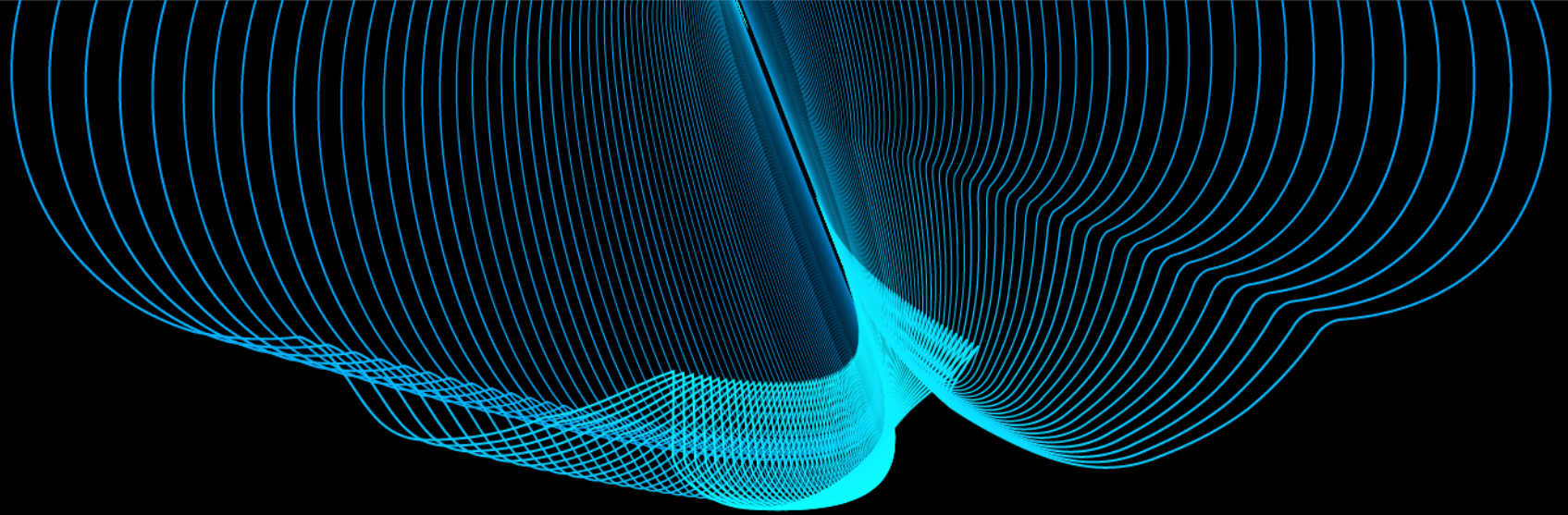
Cloud for Holography and Augmented RealITY

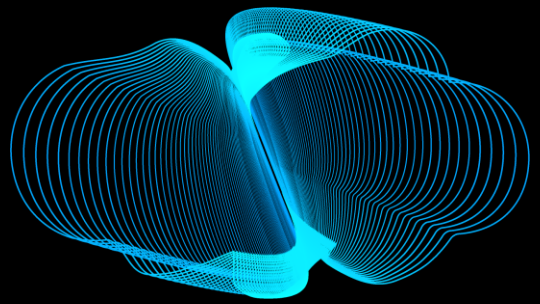
Thank you for your attention!

www.charity-project.eu

@CharityProj

Acknowledgment:
The research conducted by CHARITY receives funding from the European Commission H2020 programme under Grant Agreement Nº 101016509. The European Commission has no responsibility for the content of this presentation.





DataCloud

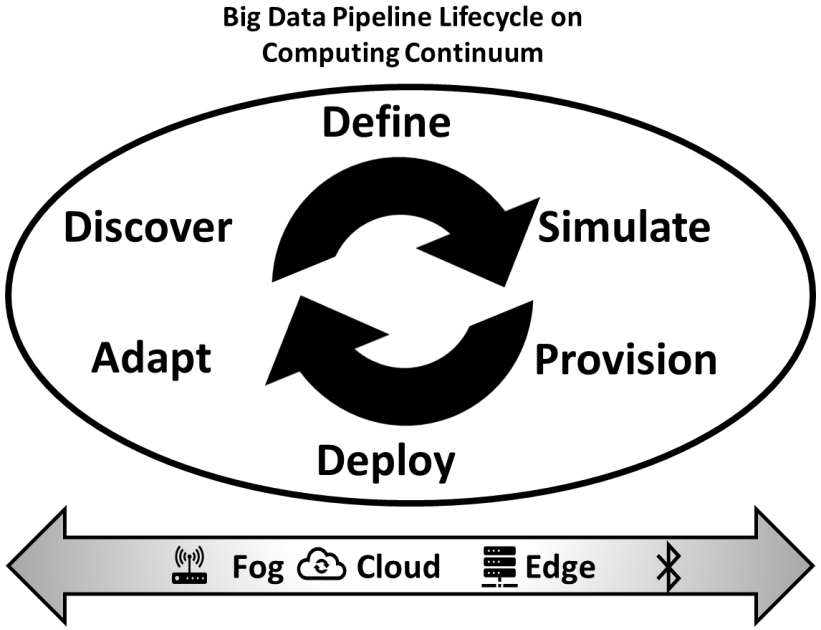
DataCloud: Enabling the big data pipeline lifecycle on the computing continuum

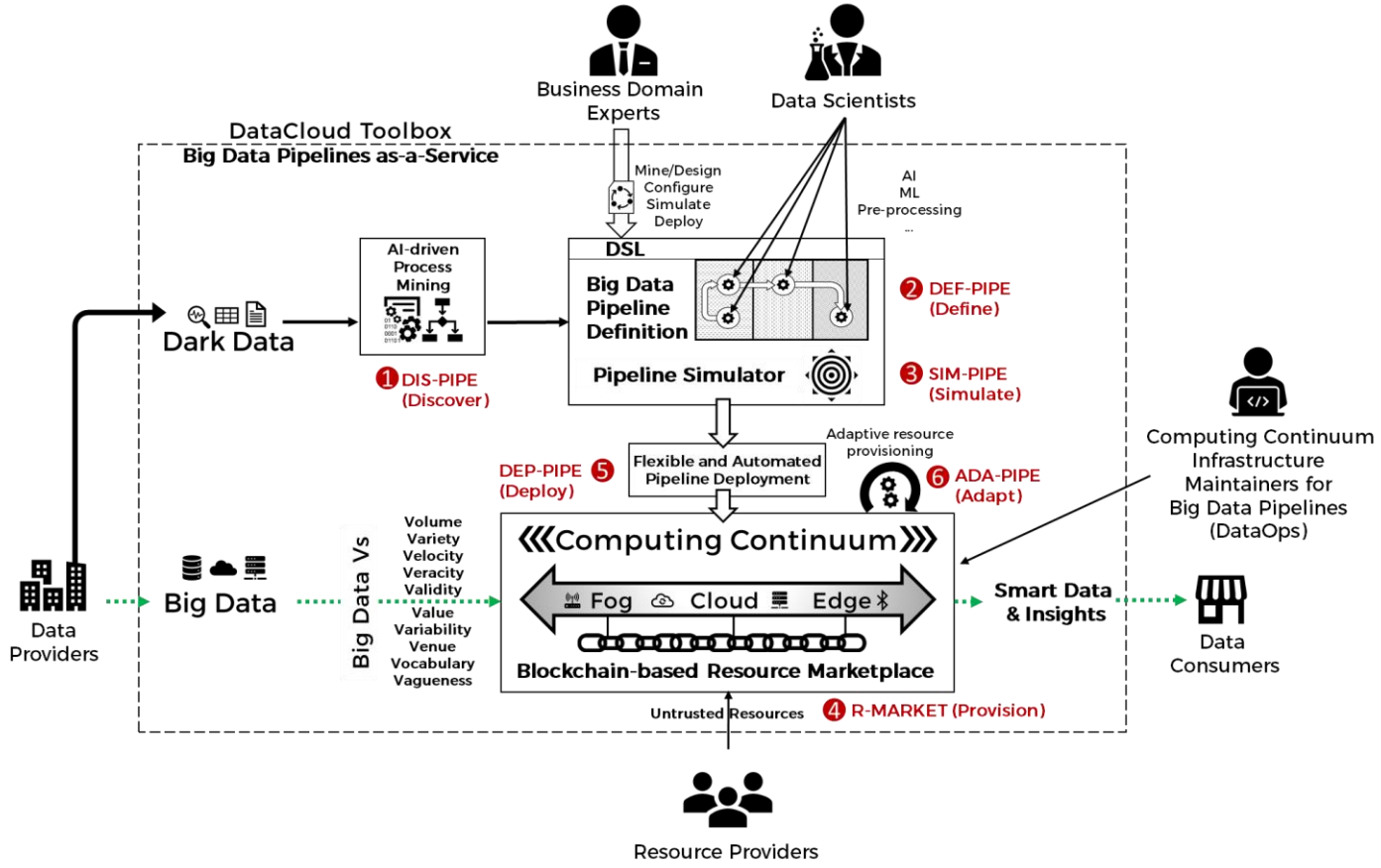
Dragi Kimovski

Klagenfurt University

Main challenges, scope, and a brief description

Big Data pipelines are composite pipelines for processing data with non-trivial properties and characteristics, commonly referred to as the *Vs of Big Data*.





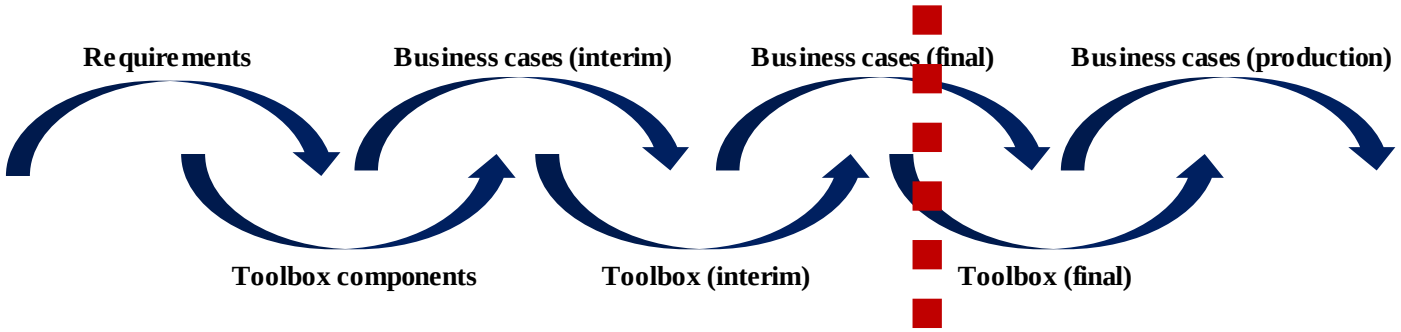
Lessons learned and Success Stories

Lessons learned:

- Decentralisation complexity
- Pipeline discovery accuracy
- Pipeline definition ambiguities
- Simulation and its relation with the real world
- Scheduling, adaptation and deployment scalability, elasticity and events identification

Core achievements and successes:

- Software for all technical components and their first integration in the toolbox
- Novel approaches, algorithms and tools for: pipeline discovery, definition, simulation, scheduling, adaptation, and deployment
- Open, decentralise resource marketplace
- Business cases prototypes
- Validation of the toolbox components in the business cases



Recommendations for the Future

- Big Data pipeline discovery approaches are essential for identifying dark data and extracting knowledge from it
- Proper and powerful DSL language is prerequisite for optimizing data pipelines
- Integration of public cloud instances with private edge resources using decentralized marketplace is essential for supporting the future computing continuum
- Efficient simulation tools can improve both pipeline definition and deployment approaches
- Scheduling, deployment, and adaptation to identified anomalies is necessity to enable efficient runtime over decentralized resources over the computing continuum

PHYSICS

oPTIMIZED HYBRID SPACE-TIME SERVICE CONTINUUM IN FAAS

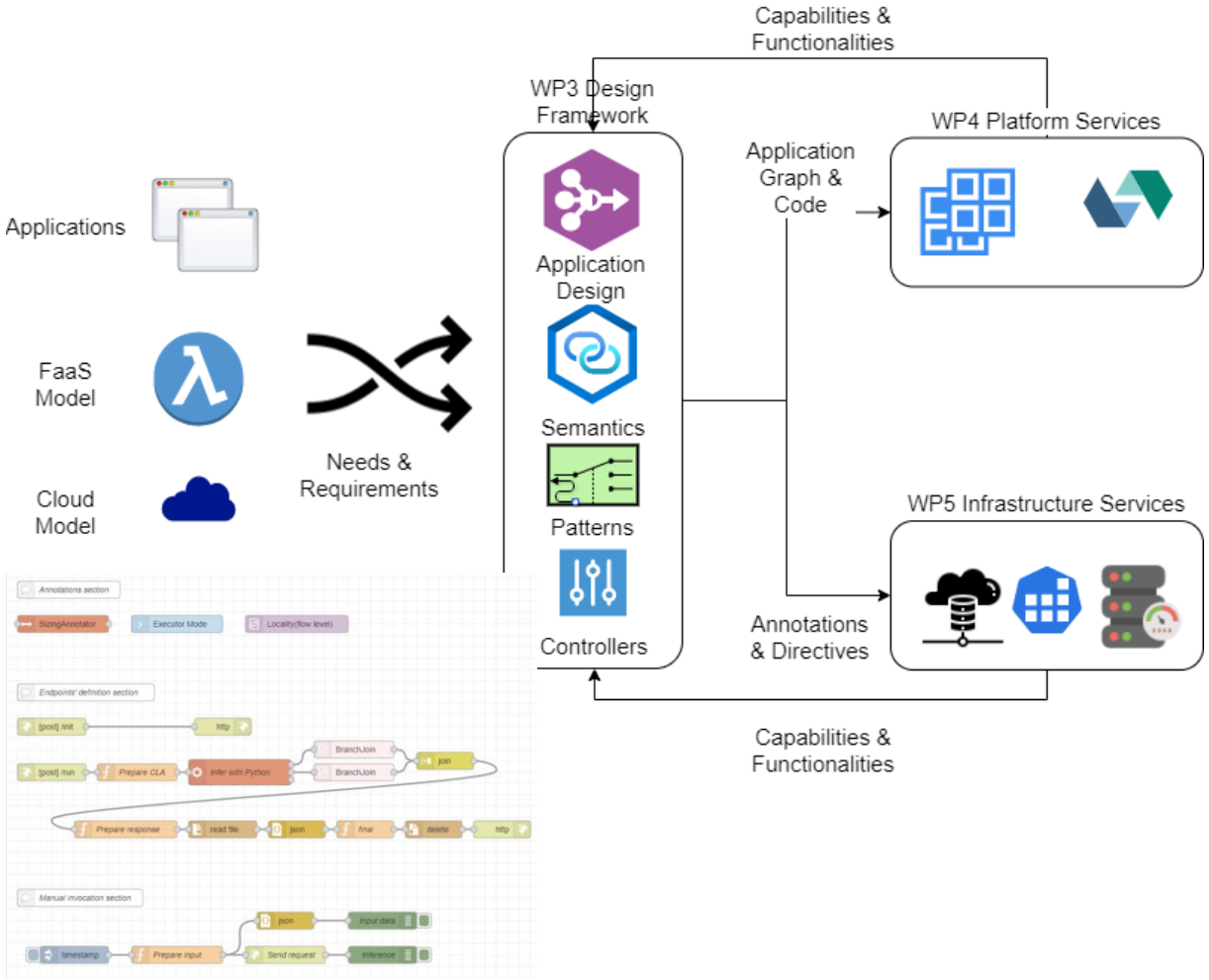
This project has received funding from the European Union's horizon 2020
research and innovation programme under grant agreement no 101017047

Project Presentation

Visual programming environment to create **arbitrary** serverless functions and workflows with **reusable** patterns

- Abstractions, Automation, Customization

Platform-level functionalities to orchestrate FaaS workflows and optimize cloud/edge interplay with parameterized selection, placement and dynamic reconfiguration



Provider-local resource management mechanisms to offer competitive and optimized services execution with improved local scheduling and internal placement ablities

This project has received funding from the European Union's horizon 2020 research and innovation programme under grant agreement no 101017047

Lessons Learned & Success Stories

- Early engagement with the Use cases
- Enable agility for the environment
  - Define placeholders or plugin points for different technologies/stacks
  - Enable user customization to the largest extent
- Less is More
  - Integration Perspective: Multiple combinations of subsystems without dependencies
    - Ability to cherrypick
  - Functionality Perspective: simple advancements can have high impact
- Asset Based Approach
  - Multiple standalone artefacts-> multiple exploitation packages-> multiple audiences
- Bootstrapped development
- Pattern based approach paid off
  - Boosts up productivity, reuse, speed of development, abstraction especially in low code environments

This project has received funding from the European Union's horizon 2020 research and innovation programme under grant agreement no 101017047

Recommendations for the future

- Applications struggle to meet advancements and be cloud-native
  - Need to provide better support for software engineering
- In the future, IT personnel will probably be less skilled
  - Gap between IT professionals and current/future needs
  - Conversion courses
  - Abstractions and low learning curve/background are needed
- Convergence, integration and automation
  - Platform engineering should play a more decisive role

This project has received funding from the European Union's horizon 2020 research and innovation programme under grant agreement no 101017047

SERRANO - Transparent application development in a secure, accelerated and cognitive cloud continuum, No. 101017168

ICT-40-20 on “Cloud Computing: towards a smart cloud computing continuum”, Research & Innovation Action (RIA)

Concertation and Consultation on Computing Continuum: From Cloud to Edge to IoT

10-11 May 2023

Prof. Emmanouel (Manos) Varvarigos

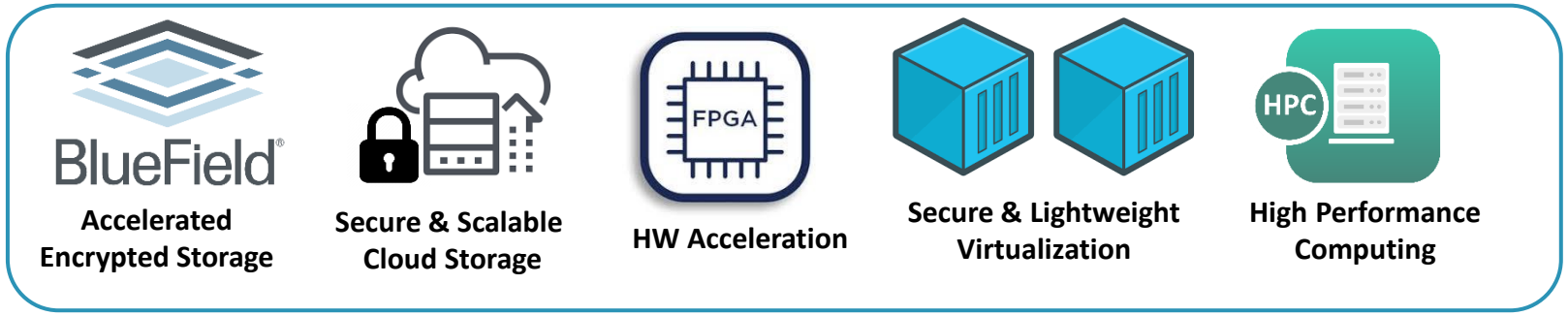
Dr. Aristotelis Kretsis

Dr. Panagiotis Kokkinos

Institute of Communication and Computer Systems – ICCS

Brief Project Presentation and Main Challenges

- SERRANO targets disaggregated and federated infrastructures consisting of edge, cloud and HPC resources
- SERRANO introduces and integrates in a single platform a number of key innovations:
  - Intent-driven operations for infrastructure-agnostic deployments
  - HW/SW platforms for enhanced security and privacy
  - Hardware acceleration and approximation computing
  - Workload isolation and execution trust on untrusted resources
  - Cognitive and intelligent service & resource orchestration



Lessons Learned and Success Stories

- Platform architecture, technologies and components developments were based from the start on the use cases
  - This enabled the close collaboration of partners in the development & integration activities towards SERRANO's objectives
- **Success Stories**
- SERRANO-enhanced Storage Service for secure storage at the edge-cloud continuum
  - Security and privacy mechanisms based on a number of hardware and software innovations
  - Achieve less CPU utilization, improvements on the transfer speed and reduction on latency
- Develop common kernels that utilize approximation techniques: on edge and cloud HW accelerators, HPC SW kernels
  - Achieve energy gains and speed ups
- Facilitate Application Development/Deployment
  - Utilize unikernels to isolate instances and to reduce the attack surface of the running applications
  - Enable programming access to hardware accelerators by applications that run as containerized functions
  - Implement tools to ease development on HW accelerators (FPGAs, GPUs)
    - Achieve lower kernel latency and increase in the accelerator density
- SERRANO Platform realization – integration of technologies and components
  - Cognitive service and resource orchestration
  - Event and Anomaly Detection Engine for complex & contextual anomalies in large scale distributed systems
- Use Cases Realization: i) UC related applications were developed, ii) platform components and integration points identified and validated

Recommendations for the Future

- There are several hardware and software technologies developed in IoT, edge, cloud, HPC
  - The validation of the developed technologies in the various domains should be continuous
- The realization of any edge infrastructure depends highly on its deployability
  - Enable more affordable deployments of novel edge platforms, through HW/SW disaggregation, open APIs and open-source platforms
- The business aspects of operating numerous edge resources on the field should be considered through resource sharing and federations of resources
- Future edge infrastructures should be able to accommodate any kind of processing and storage tasks/workload that today are served by cloud resources as if there is no cloud
- Interoperability between data producers and data consumers should be enhanced enabling the development of innovative application in various domains by third-party developers