Cognitive Cloud Continuum

Dana PETCU
West University of Timisoara
Research Institute e-Austria Timisoara
Romania
https://staff.fmi.uvt.ro/~dana.petcu/

Content

- Generalities

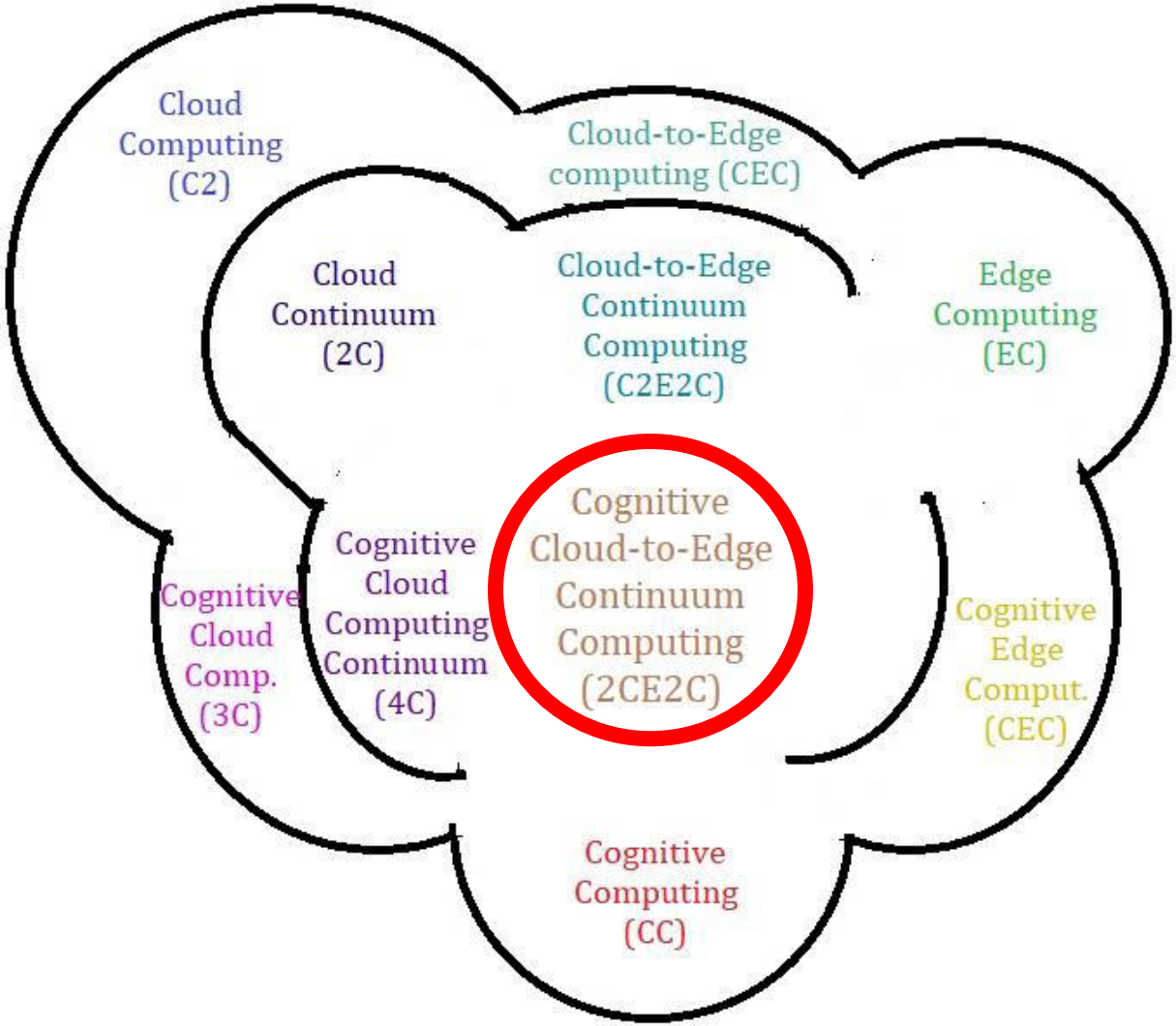
  Cloud + Edge + Cognitive

  ———————

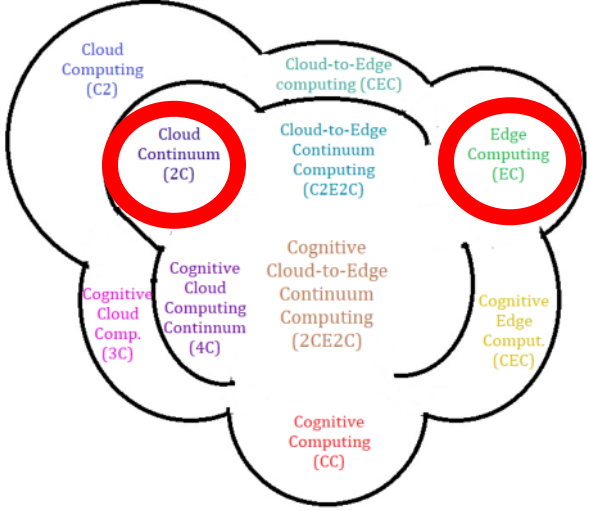
  continuum + computing

- Work in progress
  - solutions of SERRANO

General conceptual schema



Cloud vs. Edge computing



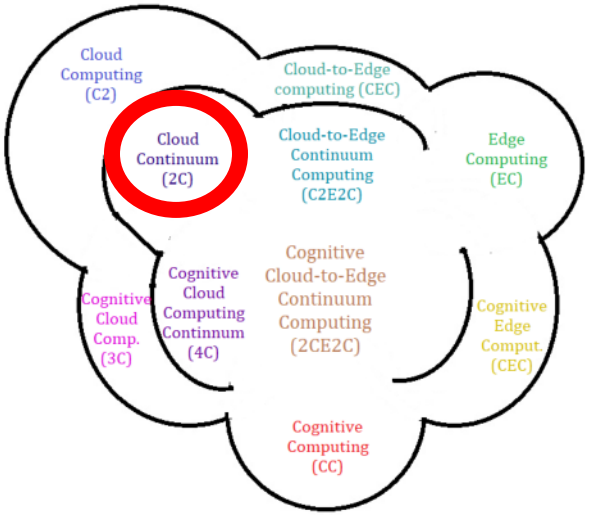
Cloud Computing (C2)

- *Based on:* centralized data centers that provide vast storage & computational capabilities
- *Used for:* data storage, resource-intensive processing, or long-term analytics.
- *Offers:* scalability, high-speed connectivity, AI & ML capabilities for complex data analysis tasks

Edge Computing (EC)

- *Based on:* the network edge or devices located closer to the data sources or end-users
  - IoT devices, gateways, local servers.
- *Used for:* capable of processing & analyzing data in real-time, with low-latency requirements.
- *Offers:* computation closer to the data source, reducing the transfer of large volumes of data to the cloud

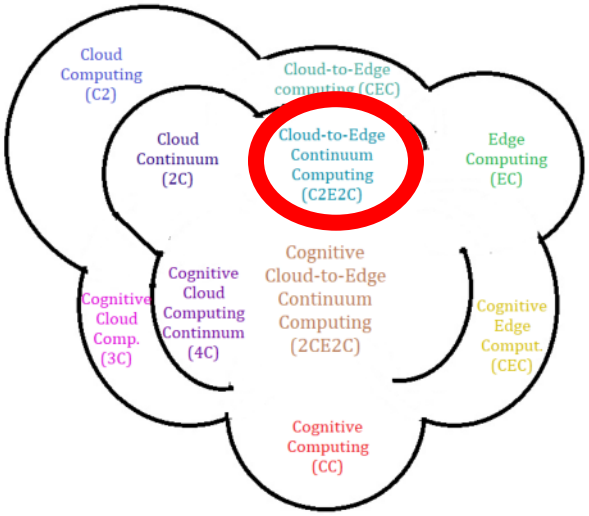
Cloud continuum (2C)



- *What is:* seamless integration & interconnectivity of various cloud computing services & deployment mode
- *Based on:* range of cloud deployment models, for public/private/hybrid/ federated/multi-clouds
  - Each model with different degrees of flexibility, cost-effectiveness & control over infrastructure and resources
- *Aim:* integration & migration of workloads/data/services across different cloud environments & service models
- *Challenges:* portability & interoperability

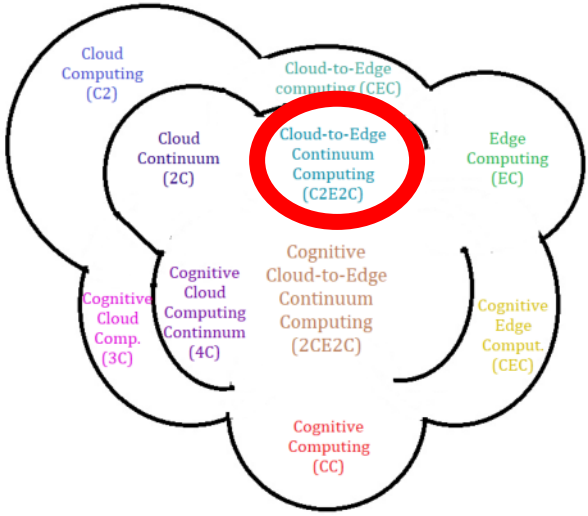
D. Petcu et al, Portable Cloud Applications - from Theory to Practice, FGCS 29 (6), 2013, doi: 10.1016/j.future.2012.01.009

Cloud-to-Edge continuum computing (C2E2C)



- *Based on:* the integration & coordination of computational resources & capabilities across the cloud & edge computing environments
- *What is:* an architectural approach for optimizing the processing & data analysis by distributing workloads between centralized cloud servers & edge devices located closer to the data source or end-users
- *Aim:* balance between the benefits of centralized cloud computing and the advantages of edge computing

Functionality & advantages of C2E2C



Example of functionality

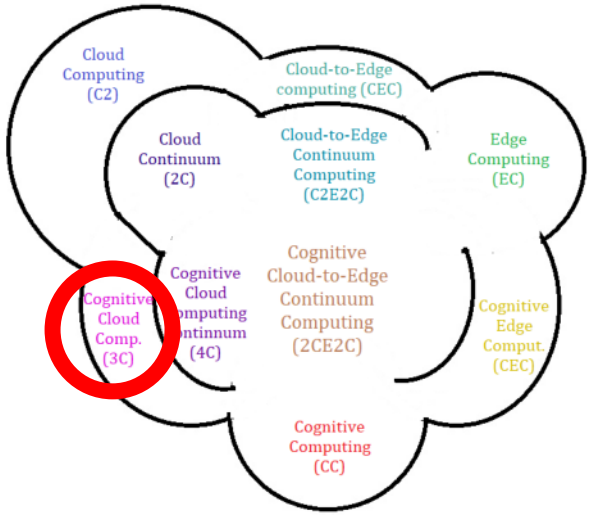
- computational tasks / data processing performed@Cloud
  - for e.g. long-term analytics, model training, or resource-intensive computations
- time-sensitive or latency-critical tasks are offloaded to edge devices
  - for real-time processing,
  - enabling faster decision-making & immediate actions

D. Petcu, Service Deployment Challenges in Cloud-to-Edge Continuum, SCPE 22 (3), 2021, doi: 10.12694/scpe.v22i3.1941

Advantages vs. Cloud

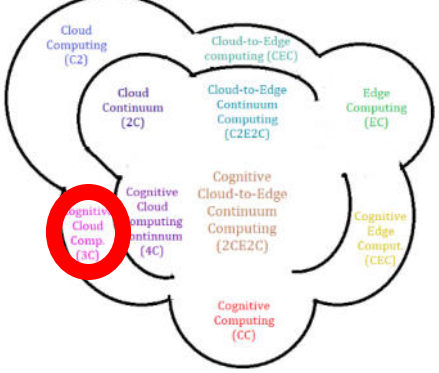
- Reduced network latency
- Bandwidth optimization
- Real-time insights and actions
- Improved privacy & security
- Scalability & flexibility:
  - allows dynamic resource allocation + scaling processing capabilities between cloud & edge based on demand & availability

Cognitive Cloud Computing (3C)



- *What is:* the integration between cognitive computing and cloud computing technologies
- *Based on:* a conceptual framework that combines the power of AI & ML with the scalability & flexibility of cloud computing infrastructure
- *Cognitive technologies:* natural language processing, computer vision, pattern recognition & knowledge representation

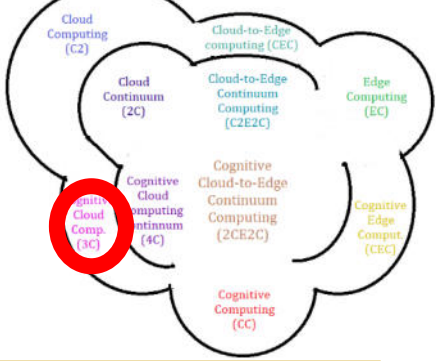
Application of 3C



Harness the power of AI & ML on a scalable and flexible cloud infrastructure, enabling intelligent & data-driven decision-making

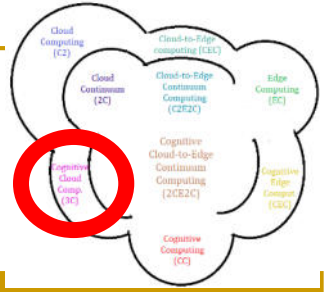
Domain	Example app	Power of AI and ML	Power of Cloud
Healthcare diagnostic & treatment	analyze huge medical data sets	apply AI & ML, to gain valuable insights for accurate diagnosis & personalized treatment plans	enables the storage, processing, and secure sharing of sensitive medical data, ensuring scalability & availability
Fraud detection	detection of fraudulent activities in real-time	analyzing historical transaction data, user behavior patterns, and external factors like market trend	computational power to handle large-scale data processing and analytics
Intelligent customer services	virtual assistants that can understand and respond to customer inquiries	natural language powered by AI & ML providing personalized recommendations, placing orders, scheduling appointments	enables the scalability & availability of the virtual assistant service
Autonomous vehicles	support the development of autonomous vehicles	AI interpret sensor data from cameras/LiDAR, to recognize objects, predict behavior & make informed decisions	collect & process data from multiple vehicles, enable collaborative learning & enhancing the intelligence of autonomous systs.
Smart manufactu-ring	optimize manufacturing processes by analyzing data from sensors & IoT devices deployed on factory	AI algorithms can detect anomalies, predict equipment failures, and optimize production schedules	centralize data collection, analysis, & decision-making, enable real-time insights & proactive maintenance

3C technologies support



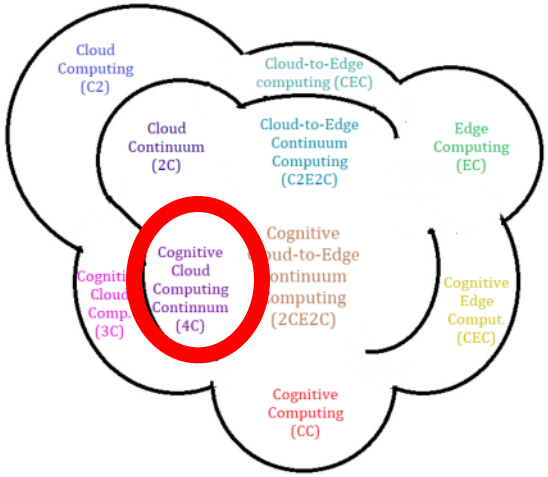
Topic	How supports 3C
AI	ML, NLP, CV, DL
ML	analyze and process large volumes of data, recognize patterns, and generate insights
NLP	text analysis, sentiment analysis, and language translation, particularly in areas like chatbots, virtual assistants, and customer service
CV	image recognition, object detection, and facial recognition used in autonomous vehicles, surveillance systems, and medical imaging analysis
DL	algorithms like CNNs or RNNs are effective in processing complex data, such as images, audio, & natural language for tasks like image and speech recognition, recommendation systems & predictive analytics
C2	offer scalable computing power, storage, data management & deployment options for cognitive applications provide APIs and tools for integrating cognitive services, such as AI & ML frameworks, into cloud-based applications
Tools	libraries, APIs, and development environments for building and deploying cognitive applications, training ML models & processing large-scale data e.g. Python, R, TensorFlow, PyTorch, Keras, scikit-learn, Apache Spark.

3C software architecture



Layer	Responsabilities	Techs
Data Ingestion	responsible for ingesting &collecting data from various sources, eg. IoT devices, databases, external APIs, streaming platforms; data integration, transformation, and normalization processes to ensure that the data is suitable for cognitive analysis	Apache Kafka, Apache NiFi, or cloud-based data ingestion services
Data Storage	data collected in the previous layer is stored in a scalable and resilient manner in the data storage	Amazon S3, Google Cloud Storage, MongoDB, Apache Hadoop
Data Processing	data is processed and analyzed using cognitive computing techs and ML involving involves data preprocessing, feature extraction, model training, and inference	TensorFlow, PyTorch, Apache Spark, scikit-learn, cloud-based ML platforms like Amazon SageMaker / Google Cloud ML Engine
Cognitive Services	pre-built cognitive services and APIs provided by cloud service providers or third-party vendors. These services include natural language processing, computer vision, sentiment analysis, recommendation engines, or speech recognition	AWS Rekognition, Azure Cognitive Services, or Google Cloud Natural Language API
Application	provides the necessary computational resources, scalability, and reliability to support cognitive applications	web applications, mobile applications, chatbots, or virtual assistants; Django, Flask, React, Angular, native mobile app dev frameworks
Cloud Infrastructure	provides the necessary computational resources, scalability, and reliability to support cognitive applications	AWS, Azure, or GCP offer a range of services, including VMs, containers, serverless computing, networking, and storage solutions
Monitoring and Management	monitoring, logging, and management tools to ensure the performance, availability, and security of the cognitive applications and the cloud infrastructure	cloud monitoring platforms, APM tools, log management systems, security monitoring tools, orchestration frameworks like Kubernetes

Cognitive Cloud Computing Continuum (4C) - applications

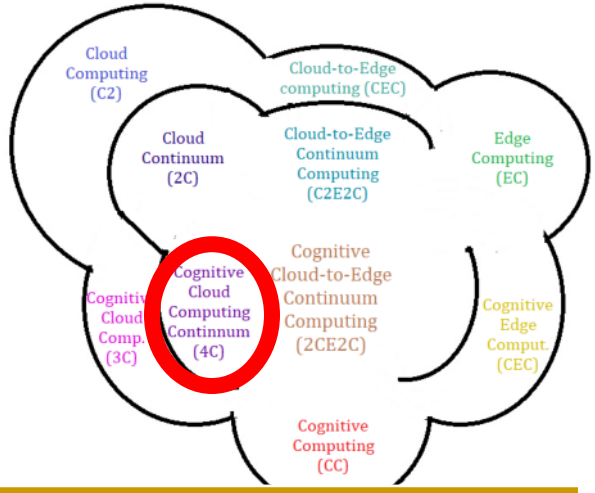


harness the power of AI, ML, and real-time data analysis to enhance decision-making, optimize processes, and deliver personalized experiences

Improve 3C appls with real-time analysis + new:

Domain	Example app	Power
Smart Energy Management	efficient energy distribution, demand response programs, and cost savings	optimize energy management by analyzing data from smart meters, IoT devices, and other sources applying cognitive capabilities and ML, energy consumption patterns can be analyzed, anomalies detected & predictive models developed for energy demand forecasting
Marketing	personalized marketing and customer engagement	analyze customer data, social media interactions, and demographic information, organizations to deliver targeted advertisements, personalized recommendations & tailored customer experiences. enable real-time data analysis and dynamic content delivery

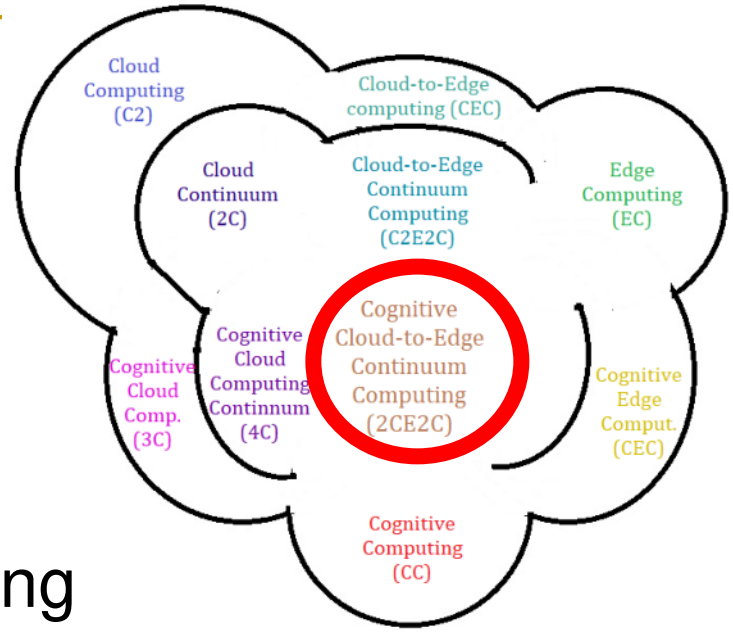
Cognitive Cloud Computing Continuum (4C)



- Supported by 3 C techs +

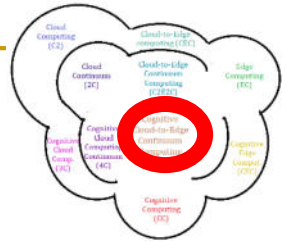
Topic	Tech	How
Big Data	Apache Hadoop, Apache Spark, distributed databases	enable efficient storage, processing & analysis of big data sets
Workflow Orchestration	Apache Airflow, Luigi, AWS Step Functions	used to automate & manage the workflow of cognitive applications. coordinate & schedule various tasks, data processing steps, and model training and deployment processes
Container-zation	Docker, Kubernetes	enable the efficient deployment & scaling of cognitive applications in cloud environments provide portability, scalability & resource management capabilities
Data Visualisation	Tableau, Power BI, Matplotlib	enable the visual representation of data & insights generated by cognitive applications facilitate the interpretation and communication of results to end-users

Cognitive Cloud-to-Edge continuum computing (2C2E2C)



- *What is*:
  - integration of cognitive computing capabilities, cloud computing, and edge computing
- *Based on:*
  - power of cognitive technologies
  - computational resources of Cloud
  - proximity of edge devices

2C2E2C key aspects



Key aspect	Particularity	Functionality
Cloud-based Cognitive Services	Cloud offer pre-built cognitive services, such as language understanding, image recognition, sentiment analysis & recommendation engines	Use ML algorithms and AI models trained on vast amounts of data to deliver cognitive capabilities
Edge-based Cognitive Processing	Edge devices are equipped with computational capabilities to perform cognitive processing tasks locally	Running AI models, executing ML algorithms, analyze data in real-time at the edge
Data Fusion and Insights	Data collected from various edge devices and sensors are fused and analyzed to extract meaningful insights	Cognitive techniques are applied to process, analyze, and derive actionable intelligence from the data, enabling real-time decision-making and proactive actions at the edge
Hybrid Data Processing	Data processing is distributed across the Cloud and Edge based on factors like latency requirements, data volume, privacy concerns & the need for real-time responsiveness	Data may be processed in the cloud for long-term analytics and training of AI models, while time-sensitive or latency-critical tasks are handled at the edge
Dynamic Resource Allocation	Computational tasks are distributed between cloud servers and edge devices based on resource availability, network conditions, and application requirements	Enables efficient utilization of computational resources and optimizes the overall system performance

2C2E2C appls – enhancement vs. 3C/4C

Domain	Edge	Cloud
Smart Manufacturing	Edge devices can perform local cognitive processing to detect anomalies, predict failures, and optimize production processes	Cloud-based cognitive services can provide higher-level analysis and decision support for overall process improvement and optimization
Autonomous Vehicles	Edge devices like onboard sensors, cameras & processors analyze data locally for real-time object detection, road condition assessment & collision avoidance	Cloud-based cognitive services enhance decision-making by incorporating higher-level context, traffic patterns, and learning from other vehicles

2C2E2C applications – new vs. 3C/4C

Domain	Edge	Cognitive services @ Edge	Cognitive services @ Cloud
Telemedicine	edge devices like wearable health monitors or IoT-enabled medical devices can continuously monitor patient health data	enable real-time analysis of life signs, anomaly detection & personalized health recommendations	provide deeper analysis, access to medical knowledge bases, and assist professionals in making accurate diagnoses
Smart Cities	edge devices & sensors deployed in city collect & process data in real-time to enable intelligent traffic signal optimization, energy load balancing & surveillance for detecting suspicious activities	applied to traffic management, energy optimization & public safety	enhance the analysis with city-wide insights and predictive modeling
Retail and Customer Service	edge devices like smart shelves or cameras analyze customer behavior, inventory levels, and product placements	enable real-time inventory tracking, personalized recommendations & targeted advertising	provide deeper customer insights, sentiment analysis, and customer service support
Environmental Monitoring	Edge devices deployed in environmental monitoring systems can capture data related to air quality, weather conditions, and ecological parameters	perform real-time analysis, anomaly detection, and early warning systems for natural disasters	aggregate data from multiple edge devices, provide advanced climate modeling, and support decision-making for environmental conservation and disaster management

2C2E2C technologies

Category	Examples	Functionality
Cloud Computing Platforms	Amazon Web Services, Microsoft Azure, Google Cloud Platform	provide the infrastructure and services necessary for deploying and managing cloud resources, offer services such as VMs, storage, databases, and AI/ML
Edge Computing Frameworks	AWS IoT Greengrass, Azure IoT Edge, Google Cloud IoT Edge	enable the deployment & management of edge devices and applications, provide tools for containerization, local data processing & connectivity between edge devices and the cloud
Machine Learning and AI Frameworks	TensorFlow, PyTorch, and scikit-learn	used for developing/deploying ML/AI models, provide APIs, libraries, and tools for training, optimizing & deploying models to run in cloud & and on edge devices
Containerization Technologies	Docker, Kubernetes	used to package applications and their dependencies into containers, enable consistent deployment across different environments, including cloud and edge devices, allowing for portability/scalability
Edge Analytics Libraries	Apache NiFi, Apache Flink, Apache Kafka	for performing data processing, analysis, and inference at the edge, e.g. for stream processing, data routing, and real-time analytics on edge devices
IoT Protocols /Technologies	MQTT, CoAP, AMQP / Zigbee, Bluetooth Low Energy (BLE), LoRaWAN	used for data transmission between edge devices and cloud platforms. Additionally / facilitate connectivity and data exchange in IoT deployments
Programming Languages	Python, Java, C++	rich ecosystem of libraries and frameworks for data analysis, ML and AI.
Data Management and Integration Tools	Apache Airflow	integrate data flows between edge devices and cloud platforms. They provide features for data ingestion, transformation, and routing, ensuring seamless data exchange

Software architecture of 2C2E2C

Layer	Built on	To do	Why
Edge Devices	sensors, actuators, cameras, or IoT devices located at the edge of the network	capture data, perform local data preprocessing & transmit relevant data to the cloud for further analysis	local computational capabilities, storage, and connectivity to communicate with other edge devices or the cloud
Edge Computing	edge servers or gateways responsible for processing data locally	real-time data analysis, inference, and decision-making based on predefined rules or machine learning models	reduces latency, minimizes data transmission to the cloud, and enables faster response times
Cloud Computing	cloud infrastructure, such as servers, storage, and virtualization resources	hosts cloud services, including data storage, AI/ML platforms, cognitive APIs, and analytics tools	provides scalability, high computational power, and access to advanced cognitive services
Data Flow and Communication	use protocols like MQTT, HTTP, or custom APIs	flows between edge devices to cloud computing nodes for processing and analysis	processed data is then transmitted to the cloud for further analysis, long-term storage, and advanced cognitive processing
Data Proces-sing and Analysis	cognitive services and AI/ML platforms	handle analytics, ML inference, and local decision-making	perform analytics, training of models, and higher-level cognitive processing
Orchestrating and Managing the Continuum	monitoring and management tools	coordinates the interaction between edge devices/ computing nodes & cloud resources; manages the allocation of computational tasks, data routing &dynamic scaling of resources based on workload and availability	track the health, performance, and security of the entire continuum

Scientific challenges of 2C2E2C

- Distributed Data Management
- Latency and Real-Time Decision-Making
- Resource Allocation and Optimization
- Security and Privacy
- Cognitive Model Distribution and Training
- Context Awareness and Adaptability
- Scalability and Heterogeneity
- Quality of Service and User Experience

Technological challenges of 2C2E2C

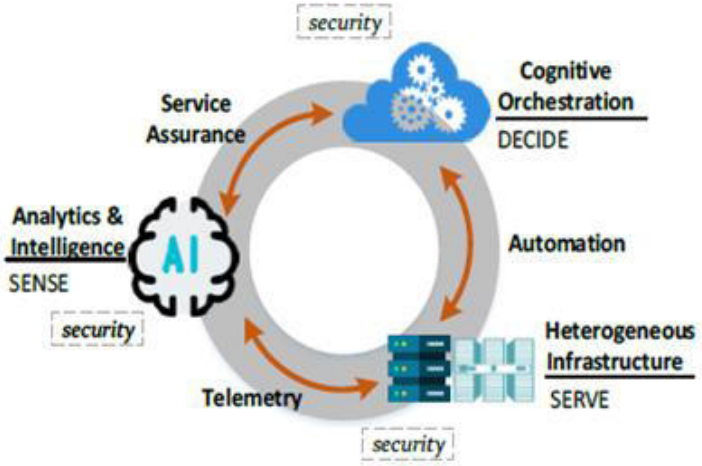
- Connectivity and Network Infrastructure
- Edge Computing Capabilities
- Data Preprocessing and Filtering
- Edge-Cloud Synchronization and Consistency
- Cognitive Model Distribution and Deployment
- Security and Privacy
- Scalability and Interoperability
- Resource Management and Optimization

Research groups/academic institutions/industry organizations/ 2C2E2C

Organization	What
IBM Research	projects related to edge intelligence, distributed ML, and C2E integration for cognitive applications
Microsoft Research	intelligent edge devices, distributed ML, real-time analytics, enable intelligent & responsive applications
Intel Labs	optimize the performance/efficiency of edge devices & enable seamless connectivity & data processing across the continuum
Google Research	edge intelligence, federated learning & scalable ML for the C2E2C
Carnegie Mellon University	CyLab Edge Comp. Res. Group: edge intelligence, data analytics & security in 2C2E2C
University of California, Berkeley	RISELab: efficient data processing, distributed ML, and real-time decision-making in the C2E2C
National Institute of Standards and Technology	standardization for EC & C2E2C: reference architectures, interoperability frameworks, and best practices for integrating cloud and edge components

SERRANO project

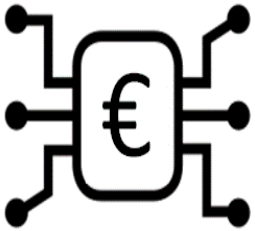
- **Title:** Transparent application development in a secure, accelerated and cognitive cloud continuum
- **Grant Agreement Number**: H2020 101017168
- **Duration**: 01/01/2021 – 31/12/2023
- **Web site:** https://ict-serrano.eu/
- **Objective 5**: Cognitive resource orchestration and transparent application deployment over edge/ cloud/ HPC infrastructures.
  - Resource orchestration: continuous adaptations for the deployed applications, based on the observe, decide, act approach
  - July 2023: final release



Mellanox TECHNOLOGIES
Chocolate Cloud
Universität Stuttgart
ARISTOTLE UNIVERSITY OF THESSALONIKI
INTRASOFT INTERNATIONAL
inbestME
InnovActs
IDEKO
Universitatea de Vest din Timișoara
nubificus

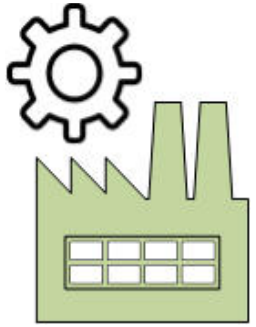
SERRANO's use cases

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



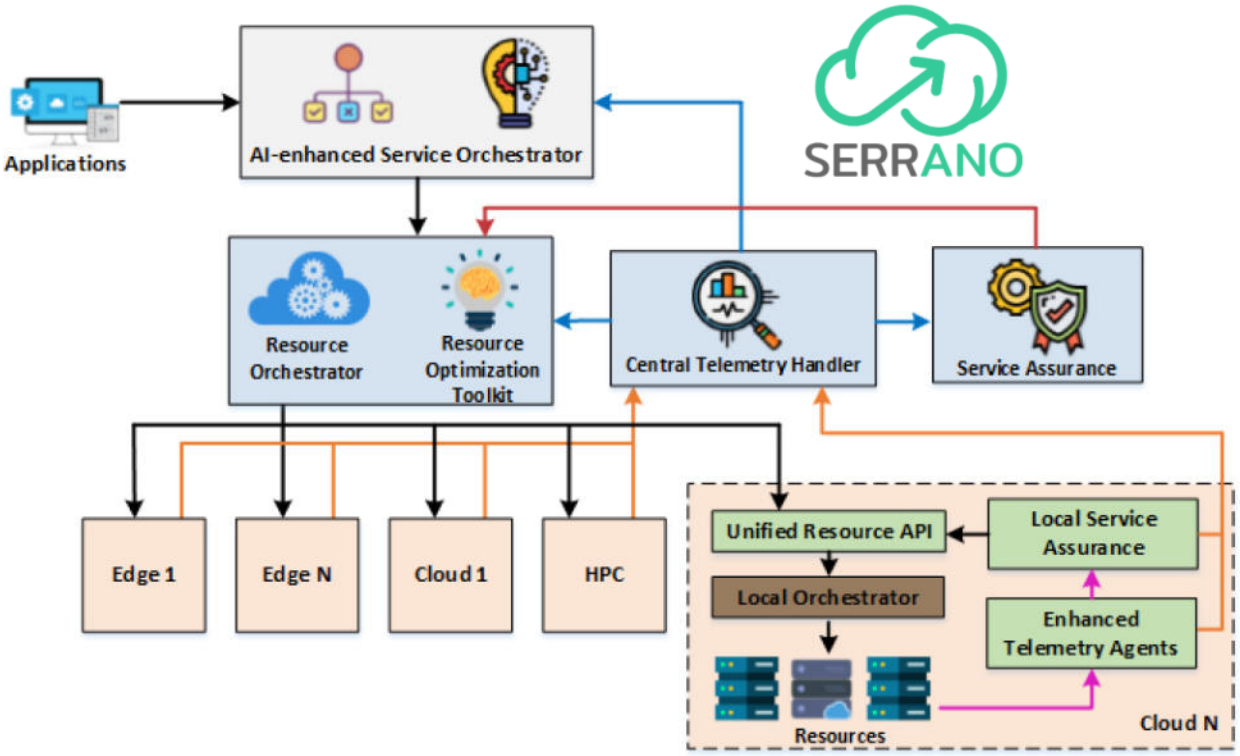
- **High-performance Fintech Analysis**
  - Apply AI and ML algorithms in financial operations
  - SERRANO will provide security and intelligent fintech app deployment

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



SERRANO's orchestration building blocks

- AI-enhanced Service Orchestrator
- Central & Local Telemetry Handler
- Central & Local Service Assurance
- Resource Optimization Toolkit
- Resource & Local Orchestrator



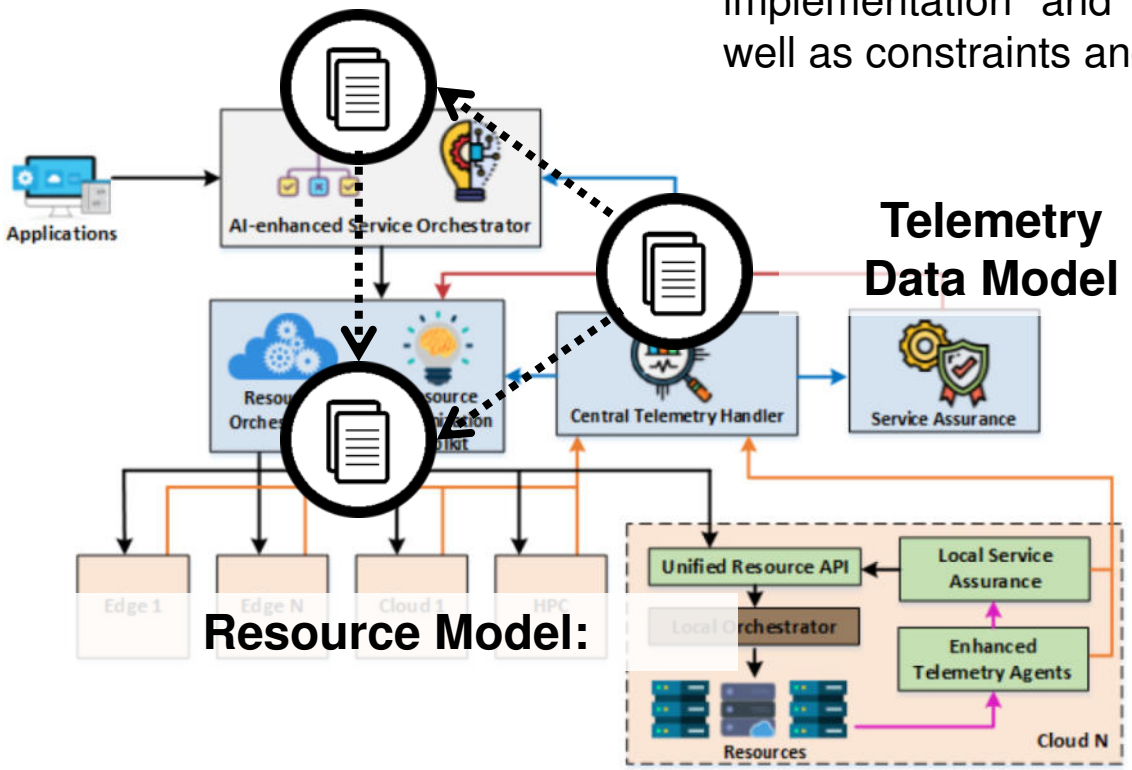
SERRANO's orchestration tools

- **Available:**
  - **ARDIA** Framework: a series of models representing the requirements and the metadata for the overall description of resource, service and application in the context of data-intensive security-critical applications and possible mappings among them
  - **AI-SO,** a service orchestrator with AI mechanisms enriching decisions with forecasting scenarios on application and overall infrastructure needs
  - **SAR,** the Service Assurance and Remediation which Trigger pro-actively and re-actively dynamic adjustments
    - Includes **EDE**, Event Detection Engine, with connectors for several common monitoring technologies and storage solutions
- **Under construction**:
  - **ROT,** the Resource Optimization Toolkit
  - AI/ML-assisted Network and Cloud Telemetry
  - Energy and Resource Aware Flow Mapping
  - Lightweight Virtualization Mechanisms

ARDIA

Application Model:

Represents the **application** in terms of its high level requirements, including intent, design, implementation and deployment information as well as constraints and capabilities



Telemetry Data Model :

Represents information collected from the **deployment** and **execution** of applications in the SERRANO platform including their overall status& communication data

Resource Model:

Represents SERRANO (physical and software) **resources**, each in terms of its capabilities (incl. performance, security), utilization and deployment details, and overall as part of an infrastructure

AI-SO

Functionality:

- Identify possible deployment scenarios based on the particular intent and application needs
- Translate high-level Application Requirements to intermediate-level Resource Requirements

Input:

- **Application Metadata** & **Requirements** specified by the app. provider based on the Application Model
- **Telemetry & Service Assurance Data** provided by the relevant SERRANO component expressed based on the Telemetry Data Model

Output:

- **Application Resource Requirements** based on the Resource Model
- Produces zero or more **Deployment Scenarios**

AI-SO step ½ - **Incorporate Experts Knowledge**

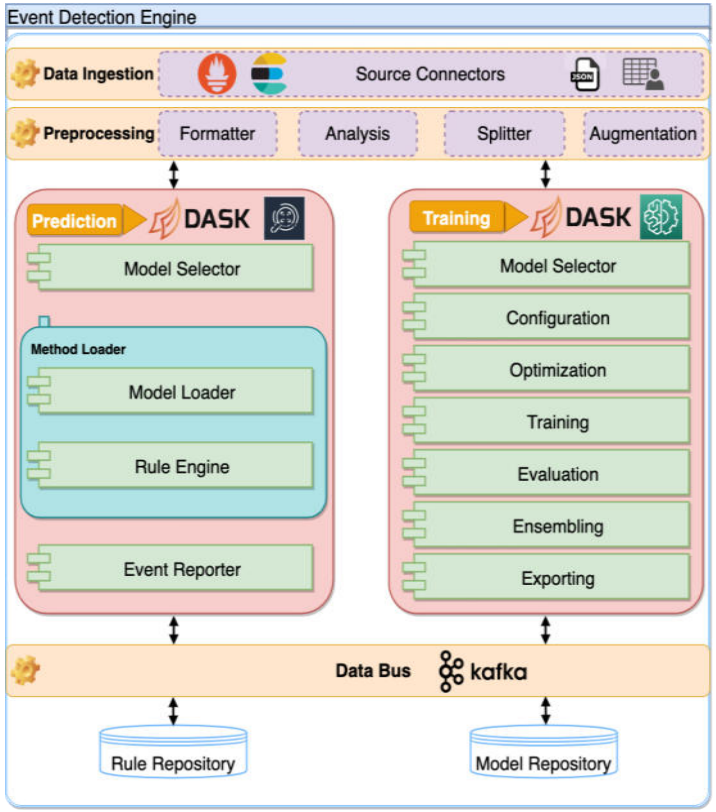
- Incorporation of Experts Knowledge in the form of Mapping Rules
  - Declarative Part – Source & Target Elements
    e.g., Security is linked with Data Encryption / Erasure Coding
  - Procedural Part – Data Transformation Process
    e.g., How Security Value affects all the others
- Deployment Scenarios Specification
  e.g., with or without hardware acceleration
- Translation of Application's Requirements to Resource Constraints
  - Application Parameters & User's Intent translation
    e.g., Translation of specific Application Security Level to
    the appropriate Data Encryption and Erasure Coding schemas

AI-SO step 2/2

Learn from Past Experience through Telemetry Data

- Application of Unsupervised ML Techniques for Data Analysis
  - Organise Telemetry Data in Groups
    e.g., detect Clusters of Telemetry Data for a particular Application
  - Study Relation among Application/Resource Constraints and Telemetry Data
    e.g., study impact of CPU constraints on Energy Consumption
  - Algorithms/Techniques: K-Means, HAC, DBSCAN, GMM, SOM
- Application of Supervise ML Techniques for Models Development
  - Utilise Service Assurance Outcome and/or Users Evaluation
  - Models Development for Decision Making
    e.g., further automatic translation of application constraints to resource constraints
  - Algorithms/Techniques: GNB, DT, RF, LR, SVM, MLP, FFNN, CNN, RNN, LSTM

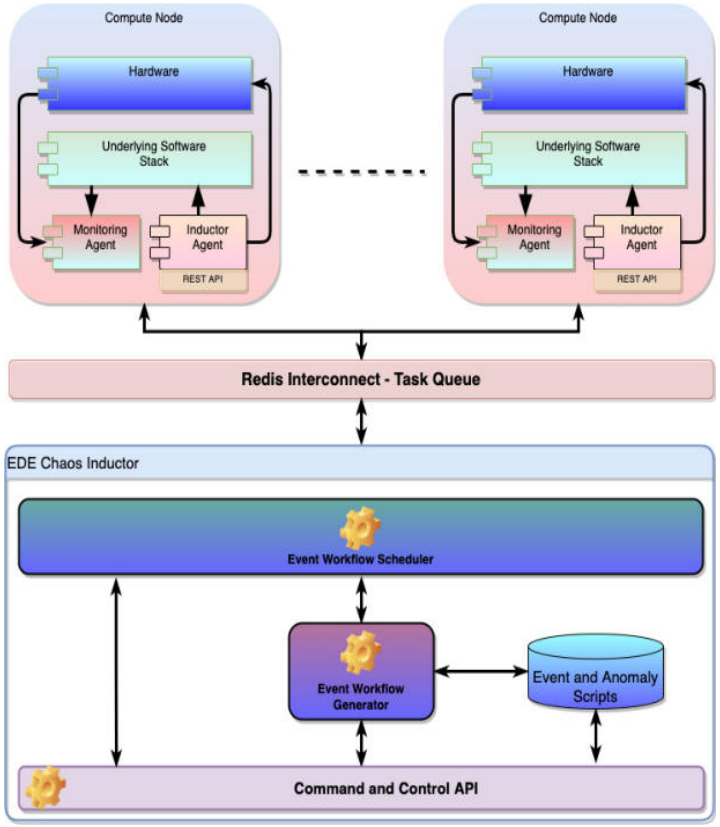
EDE - Event Detection Engine (EDE)



- Connectors for several common monitoring technologies and storage solutions.
- Several core features can be user-defined:
  - Formatter, Analysis, Augmentation, Training
- Support for large collection of detection methods (supervised and unsupervised):
  - Scikit-learn API conventions
  - Standard model export formats
- Custom HPO and evaluation metrics support:
  - Including model ensembles
- Root cause analysis:
  - Via computation of Shapely values
- Reporting via Kafka topics (i.e. Serrano Message Broker)

D.Kimovski, D.Petcu et al, Autotuning of exascale applications with anomalies detection, FBD 4, 2021, doi: 10.3389/fdata.2021.657218, I.Dragan,G.Iuhasz,D.Petcu, A Scalable Platform for Monitoring Data Intensive Applications, JoGC 17 (3), 2019 doi: 10.1007/s10723-019-09483-1

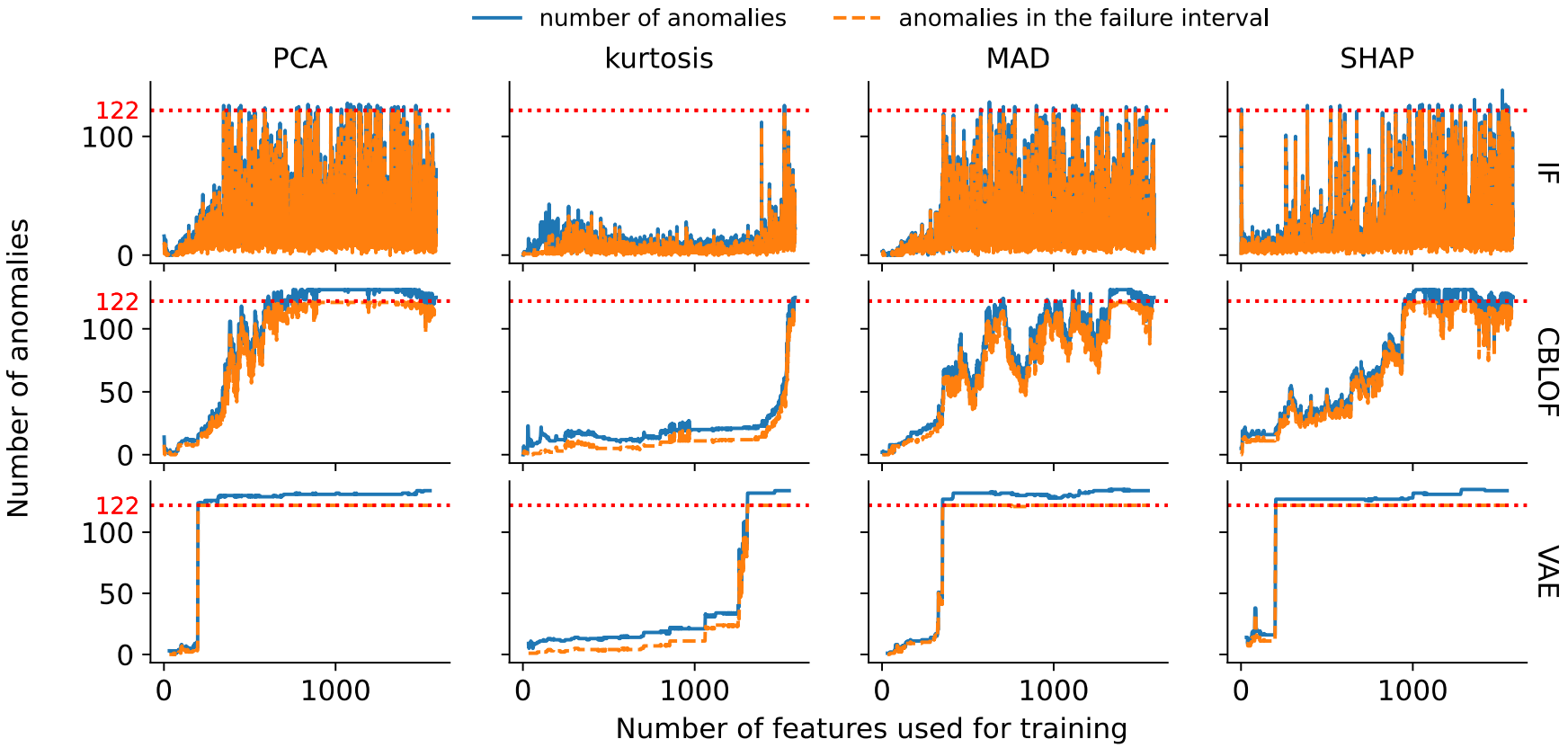
EDE - Anomaly Induction Tool



- Capable of inducing hardware anomalies in distributed systems:
  - Kubernetes support (Docker containers)
- Used on toy application*
- Anomalies:
  - CPU overload
  - Memory Eater
  - Memory Leak
  - Dial (ALU interference)
  - DDot (CPU cache fault)
  - Copy (I/O interference)
  - Page fail (page allocation fail)
  - IOError

*A. Spătaru et al, "TUFA: A TOSCA extension for the specification of accelerator-aware applications in the Cloud Continuum," *2022 IEEE 46th COMPSAC*, 2022, doi: 10.1109/COMPSAC54236.2022.00185.

Experiments



L.Cerdà-Alabern, G. Iuhasz et al, Anomaly detection for fault detection in wireless community networks using machine learning, Computer Communications 202, 2023, doi: 10.1016/j.comcom.2023.02.019

Other on-going projects

- **DiPET - Distributed Stream processing on Fog and Edge Systems via Transprecise Computing**, European CHIST-ERA, https://www.chistera.eu/projects/dipet
  - The Queen's University of Belfast - United Kingdom
  - Universitat Politècnica de Catalunya - Spain
  - Institut de Recherche en Informatique et Systèmes Aléatoires - France
  - Foundation for Research and Technology - Hellas - Greece
  - West University of Timisoara – Romania

- **COCO- Adaptivity in Cloud-to-Edge Continuum Computing**, Romanian PNIII-P4-PCE, https://coco.hpc.uvt.ro/en/home/
  - West University of Timisoara - Romania

Conclusions

- Combining Cognitive Computing with Cloud-to-Edge Computing opens the doors for the implementation of challenging & exciting applications in many fields
- Multiple approaches are available, an optimum one/standardization has not yet emerge
- Current research & development have not solve multiple problems – remain open to be investigated

QUESTIONS?