ICOS

Contents

Artificial Intelligence at the Continuum: the ICOS Intelligence Layer

Introduction

The evolution towards a compute continuum, where processing and data are distributed between the Edge and the Cloud, presents both immense opportunities and significant challenges. European edge spending will grow to $75 billion in 2026, following a five-year compound [annual growth rate (CAGR) of 14.6% since 2021](#). However, realising the full potential of this continuum is hampered by the inherent complexity of managing heterogeneous, volatile, and distributed resources. Ensuring data security, optimising resource usage, and providing contextual intelligence remain daunting challenges.

To address these issues, the ICOS project is developing a meta-operating system for the continuum, built on principles of openness, adaptability, and data sharing. At the heart of this system lies the ICOS Intelligence Layer, a sophisticated suite of AI tools designed to bring intelligent automation, predictive capabilities, and enhanced efficiency to the Edge-to-Cloud continuum[1].

This whitepaper provides an overview of the ICOS Intelligence Layer. It details its architecture, its core modules, and the twofold approach to AI that serves both the underlying ICOS MetaOS and the specific needs of diverse applications. The objective is to illustrate how ICOS leverages AI to create an intelligent, adaptive, and efficient compute continuum, paving the way for a new generation of data-intensive applications.

The ICOS Intelligence Layer Architecture

The ICOS Intelligence Layer works as an added layer in the ICOS architecture, providing the cognitive capabilities that enable the system to adapt and optimise its operations in near real-time. It is designed as a modular system, comprising five key modules that work in concert to deliver a diverse suite of AI-driven functionalities.

To ensure flexibility, scalability, and the ability to evolve as new AI techniques and requirements emerge, the Intelligence Layer is not a monolithic entity but a collection of interconnected services that interact with other parts of the ICOS ecosystem, including the Data Management Layer and the Meta-Kernel Layer.

A key architectural principle of the Intelligence Layer is its twofold approach to AI:

- **Intelligence for the ICOS MetaOS**: The layer provides a set of default or "static" metrics, such as forecasts for CPU, RAM, and GPU utilisation. These predictions are crucial for the ICOS MetaOS's matchmaking and resource allocation decisions, enabling it to proactively manage workloads and prevent resource contention.

- **Intelligence as a Service (IaaS) for ICOS Use Cases**: The Intelligence Layer also offers its capabilities as a service to applications running on the ICOS platform. This allows use cases to define and request their own dynamic, application-specific metrics. For example, an application could request a forecast for an application-specific metric, which can then be used to tweak ICOS's matchmaking and policy enforcement to better suit the application's unique needs.

Due to this, ICOS is not only intelligent in its own operations but also empowers applications to leverage the platform's AI

1. Garcia, J., Masip-Bruin, X., Giannopoulos, A., Trakadas, P., Ordoñez, S. A. C., Samanta, J., … & D'Andria, F. (2025, March). ICOS: An Intelligent MetaOS for the Continuum. In Proceedings of the 2nd International Workshop on MetaOS for the Cloud-Edge-IoT Continuum (pp. 53-59).

capabilities for their own specific optimisation goals.

The five core modules of the Intelligence Layer are:

1. **Intelligence Layer Coordination Module**
2. **Data Processing Module**
3. **AI Analytics Module**
4. **Trustworthy AI Module**
5. **AI Models and Data Repository**

The following sections will delve into each of these modules in detail.

Intelligence Layer Coordination Module

The **Intelligence Layer Coordination Module** works as the main gateway of the Intelligence Layer. It's the primary interface for other ICOS components to interact with the AI services and orchestrates the workflows within the Intelligence Layer itself, and is responsible for:

- **AIOps across the continuum**: It provides the mechanisms for sharing, predicting, training, updating, and deploying AI models throughout the Edge-to-Cloud continuum.

- **API for the MetaOS**: It allows the ICOS Meta-Kernel Layer to request AI-driven metric predictions. This API is also accessible via the ICOS Shell, enabling administrators and users to interact with the Intelligence Layer directly.

- **Internal Orchestration**: It manages the internal workings of the Intelligence Layer. It receives requests from the frontend API and coordinates the other modules (AI Analytics, Trustworthy AI, etc.) to fulfil them.

With a well-defined interface for the rest of the ICOS system, while maintaining the flexibility to evolve its internal workings independently in mind, the Intelligence Layer Coordination Module is designed as a two-part system:

- **Frontend (namely 'Metrics Export API')**: This component acts as a middle layer between the WP3 components (the Meta-Kernel Layer) and the Intelligence Layer. It is responsible for tasks such as interacting with ICOS telemetry to determine when new metric predictions are needed and then triggering the appropriate actions in the backend.

- **Backend (the 'Intelligence API')**: This is the backbone of the Intelligence Layer Coordination Module. It integrates with the other Intelligence Layer modules to manage the entire AI pipeline, from data processing and model training to model deployment and serving. It also interacts with the Intelligence Layer model registry, which stores metadata about the available AI models, including their descriptions, algorithms, and performance metrics from their evaluation stage, and allows pulling and pushing models to the online ICOS AI Models and Data Repository.
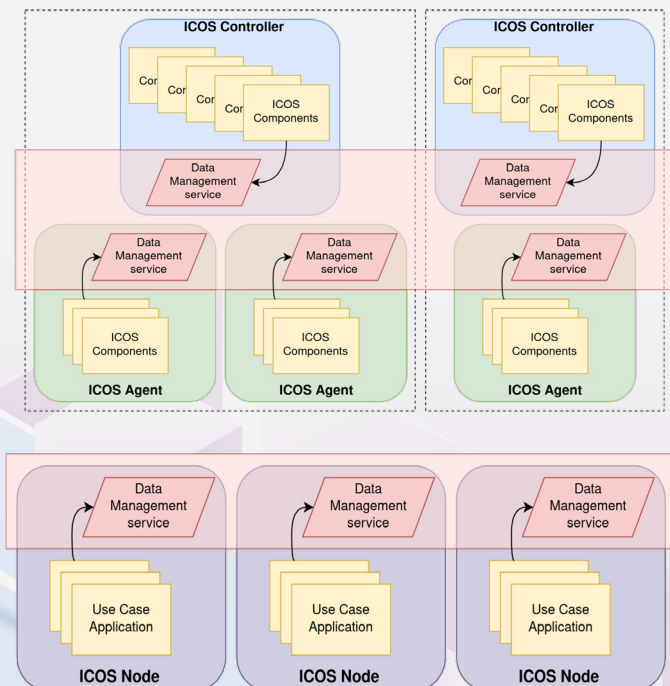
Data Processing Module

The quality and suitability of the data used to train AI models are critical to their performance. The Data Processing Module is a dedicated component within the Intelligence Layer that handles all aspects of data preparation, transformation, and access. It serves three primary purposes:

- **Data Access**: It is responsible for extracting raw training data from the ICOS Data Management Layer, which uses dataClay for distributed storage . The module connects to the dataClay client and reads datasets using unique identifiers (aliases).

- **Data Transformation and Pipelining**: Once the data is accessed, the module performs a series of operations to clean, transform, and prepare it for model training. This includes handling missing values, normalising data, and creating data pipelines that can be used to feed the AI Analytics Module.

- **Model Training Offloading**: The module facilitates the offloading of computationally intensive model training tasks to other nodes in the ICOS continuum. This is achieved through its integration with the Data Management service (through the dataClay client) and the AI Analytics Module.

AI Analytics Module

The AI Analytics Module is the powerhouse of the Intelligence Layer, responsible for building, training, and deploying the machine learning models that drive ICOS's intelligent features. It is designed to be an efficient and scalable toolkit for a wide range of AI tasks. It provides the following key functionalities:

- **AI Model Training and Inferencing**: It supports the training of a variety of AI models, including those for time-series forecasting, anomaly detection, and resource utilisation prediction. It supports multivariate and multi-output models, allowing it to predict multiple metrics (e.g., CPU and memory utilisation) simultaneously.

- **MLOps Capabilities**: The module integrates with the MLFlow platform for experiment tracking, model management, and visualisation of model evaluation results. This provides developers with a robust environment for developing, comparing, and managing their AI models.

- **Model Compression**: To enable the deployment of AI workloads on resource-constrained edge devices, the AI Analytics Module incorporates model compression techniques such as quantisation and knowledge distillation. These techniques can significantly reduce the size of AI models without a substantial loss in accuracy.

Pre-trained Models

The AI Analytics Module comes with a set of pre-trained models that are available in the Intelligence Layer's model registry. These models focus on key tasks for intelligent resource allocation, including:

- Anomaly detection
- CPU utilisation forecasting
- Intelligent, energy-aware task offloading

These pre-trained models provide a baseline of intelligence for the ICOS MetaOS and can also be used as a starting point for developing more specialised models for specific use cases.

Trustworthy AI Module

As the importance of AI-powered tech is growing in Software development and operations, ensuring that AI models used and trained in ICOS are trustworthy and adhere to European standards and laws is of paramount importance. The Trustworthy AI (TAI) Module is dedicated to providing the features and functionalities necessary to build and maintain trust in the Intelligence Layer's AI models; it focuses on three key areas to ensure

transparent, private, and reliable AI performance:

- **Model Explainability**: To enhance the transparency of its AI models, the module integrates with the SHAP (SHapley Additive exPlanations) framework. SHAP provides insights into how a model makes its predictions, helping developers and users to understand and trust the model's decisions.

- **Privacy-Aware Model Training**: The Intelligence Layer Coordination module supports model training through federated learning workloads, developed as a separate component and supported as part of the TAI module. This technique allows AI models to be trained on decentralised data without the data ever leaving the device where it is stored. This is crucial for privacy-sensitive applications and also helps to reduce data movement across the continuum.

- **Confidence and Robustness**: This module provides the Intelligence Layer API in the Intelligence Layer Coordination module with confidence scores and confidence intervals for each model prediction, giving users an indication of how confident the model is in its output. This helps in decision-making processes, particularly in situations where a high degree of certainty is required. The module also integrates with model monitoring to enable ICOS administrators to automate monitoring processes through the Intelligence API to track and ensure the robustness of the models over time [2] [2].

2. Garcia, J., Masip-Bruin, X., Giannopoulos, A., Trakadas, P., Ordoñez, S. A. C., Samanta, J., ... & D'Andria, F. (2025, March). ICOS: An Intelligent MetaOS for the Continuum. In Proceedings of the 2nd International Workshop on MetaOS for the Cloud-Edge-IoT Continuum (pp. 53-59).

AI Models and Data Repository

To promote the reuse of AI models and data, and to foster a collaborative ecosystem, the Intelligence Layer integrates with ICOS AI Models and Data Repository. This repository serves as a cold storage of models and data, as well as a central catalogue for all AI-related assets within the ICOS project, being one of the key exploitable assets of the project as well as the Intelligence Layer. It is designed to be an online catalogue that is external to the core ICOS elements. The Intelligence Layer provides a user interface allowing the possibility to download and upload AI models and datasets.

This repository not only stores models and data themselves but also rich metadata about them, including:

- **Model descriptions and documentation**

- **Performance metrics and evaluation results**

- **Information about the data used to train the models**

- **Licensing information**

Conclusions

The ICOS Intelligence Layer represents a significant step forward in the quest to build a truly intelligent and adaptive compute continuum. It enables ICOS to optimise its own operations, empower applications with "Intelligence as a Service," and foster a collaborative ecosystem around AI for the Edge and the Cloud.

The five core modules of the Intelligence Layer – the Intelligence Layer Coordination Module, the Data Processing Module, the AI Analytics Module, the Trustworthy AI Module, and the AI Models and Data Repository – work together to deliver a powerful and flexible platform for AI-driven innovation.

As the ICOS project finalises, the Intelligence Layer will continue to evolve as part of its open source ecosystem, incorporating new AI techniques and expanding its capabilities to meet the ever-growing demands of data-intensive applications in the Edge-to-Cloud continuum. The ultimate goal is to create a future where the compute continuum is not just a collection of distributed resources but a truly intelligent, self-aware, and adaptive system that can seamlessly and efficiently serve the needs of our increasingly digital world.

Models 9 ⊕

◈ ICOS-AI/ICOS-AI_icos_models Updated 13 days ago	◈ ICOS-AI/consumption_recommendation_uc4_r1 private Updated 21 days ago
◈ ICOS-AI/electricityconsumption_forecast_uc_v1 private Updated 21 days ago	◈ ICOS-AI/uc1_robot_anomaly_detection private Updated 27 days ago
◈ ICOS-AI/uc1_corn_weed_detection_yolo private Updated 29 days ago	◈ ICOS-AI/ICOS-AI_prediction_adaptive_cpu_utilisation_ Updated Feb 26 · ♥ 2
◈ ICOS-AI/quantized_models private Updated Dec 4, 2024	◈ ICOS-AI/lstm_baseline private Updated Dec 4, 2024
◈ ICOS-AI/distilled_model private Updated Dec 4, 2024	

Datasets 6 ⊕

▦ ICOS-AI/Electricity_dataset_sample ▦ Viewer · Updated 13 days ago · ▦ 17.6k · ± 81	▦ ICOS-AI/scaphandre_power_consumption ▦ Viewer · Updated 29 days ago · ▦ 300 · ± 90
▦ ICOS-AI/scaphandre_ram_usage ▦ Viewer · Updated 29 days ago · ▦ 300 · ± 89	▦ ICOS-AI/scaphandre_cpu_usage ▦ Viewer · Updated 29 days ago · ▦ 300 · ± 93
▦ ICOS-AI/synthetic_cpu_utilization ▦ Viewer · Updated Feb 26 · ▦ 50.3k · ± 15	▦ ICOS-AI/live_cpu_utilization ◈ Preview · Updated Dec 4, 2024 · ± 15

ICOS

Who we are

The ICOS project brings together a consortium of leading academic and industrial partners from across Europe, with expertise in cloud computing, edge computing, artificial intelligence, and cybersecurity. The project is funded by the European Union's HORIZON research and innovation programme under grant agreement No 101070177.

The partners involved in the development of the Intelligence Layer and integrated components include:

icos-project.eu

icos_project

icos_project

@icos_project