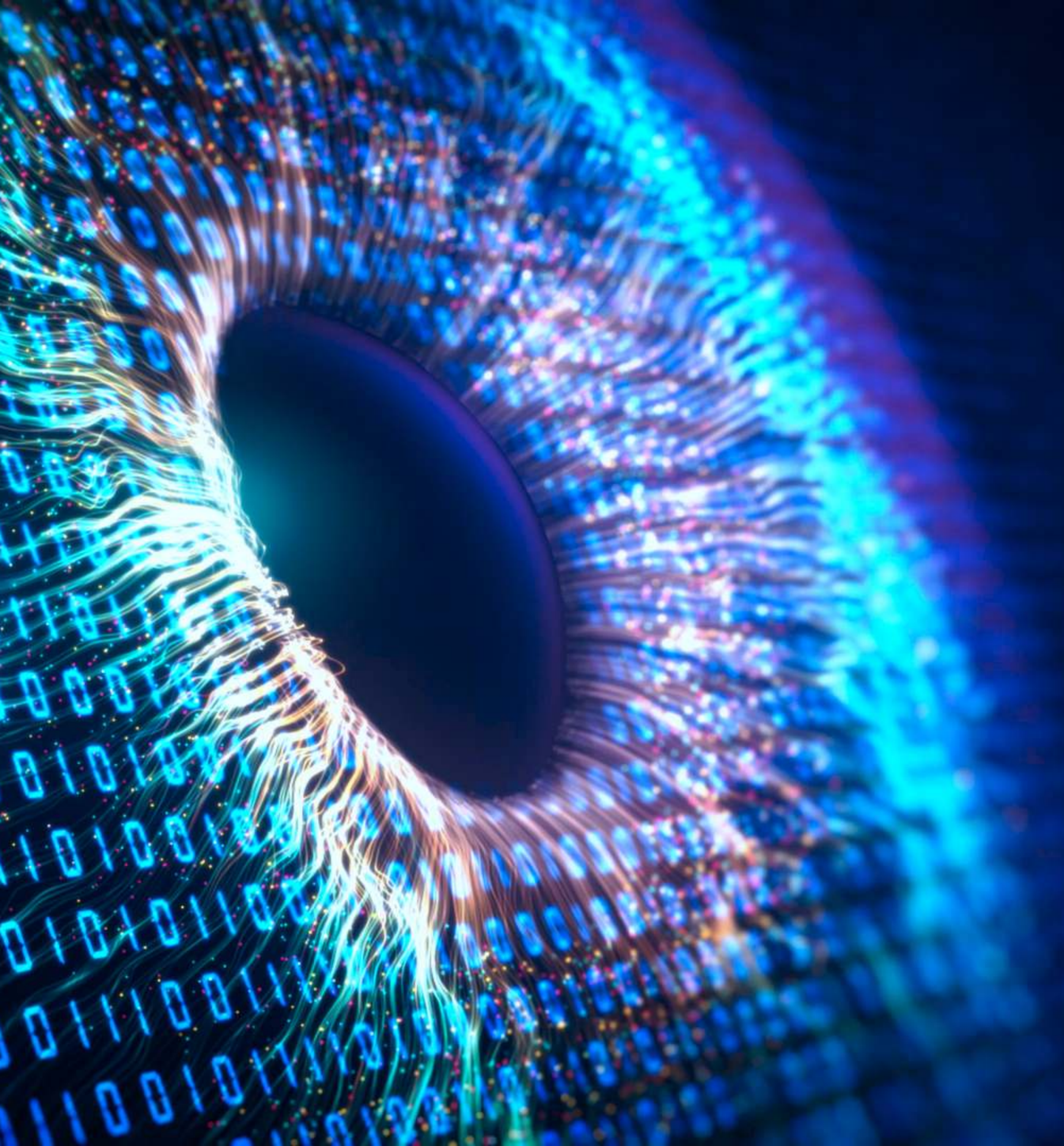


AI SPRINT

Unlocking AI's potential with AI-SPRINT

A framework for efficient AI application design and deployment

Introducing AI-SPRINT: Unlocking the Full Potential of AI Applications!

AI-SPRINT, the "**A**rtificial **I**ntelligence in **S**ecure **PRI**vacy-preserving computing co**NT**inuum" project, offers an innovative framework equipped with cutting-edge design and runtime management tools. Our solutions empower businesses to seamlessly design, partition, and operate AI applications across a wide array of cloud-based solutions and AI-based sensor devices. Emphasizing resource efficiency, performance, data privacy, and security guarantees, AI-SPRINT is the key to maximizing the potential of your AI endeavors.

Solving Real-World Challenges

For Application Developers:

- Manual AI application pipeline design and management become a thing of the past.
- AI-SPRINT brings high-level QoS annotations and automated partitioning of DNNs for optimized performance.

For Application Managers:

- Say goodbye to time-consuming manual configuration with AI-SPRINT's automated application containerization.
- Simplify performance appraisal through AI application performance profiling and design space exploration.

For Infrastructure Providers and System Integrators:

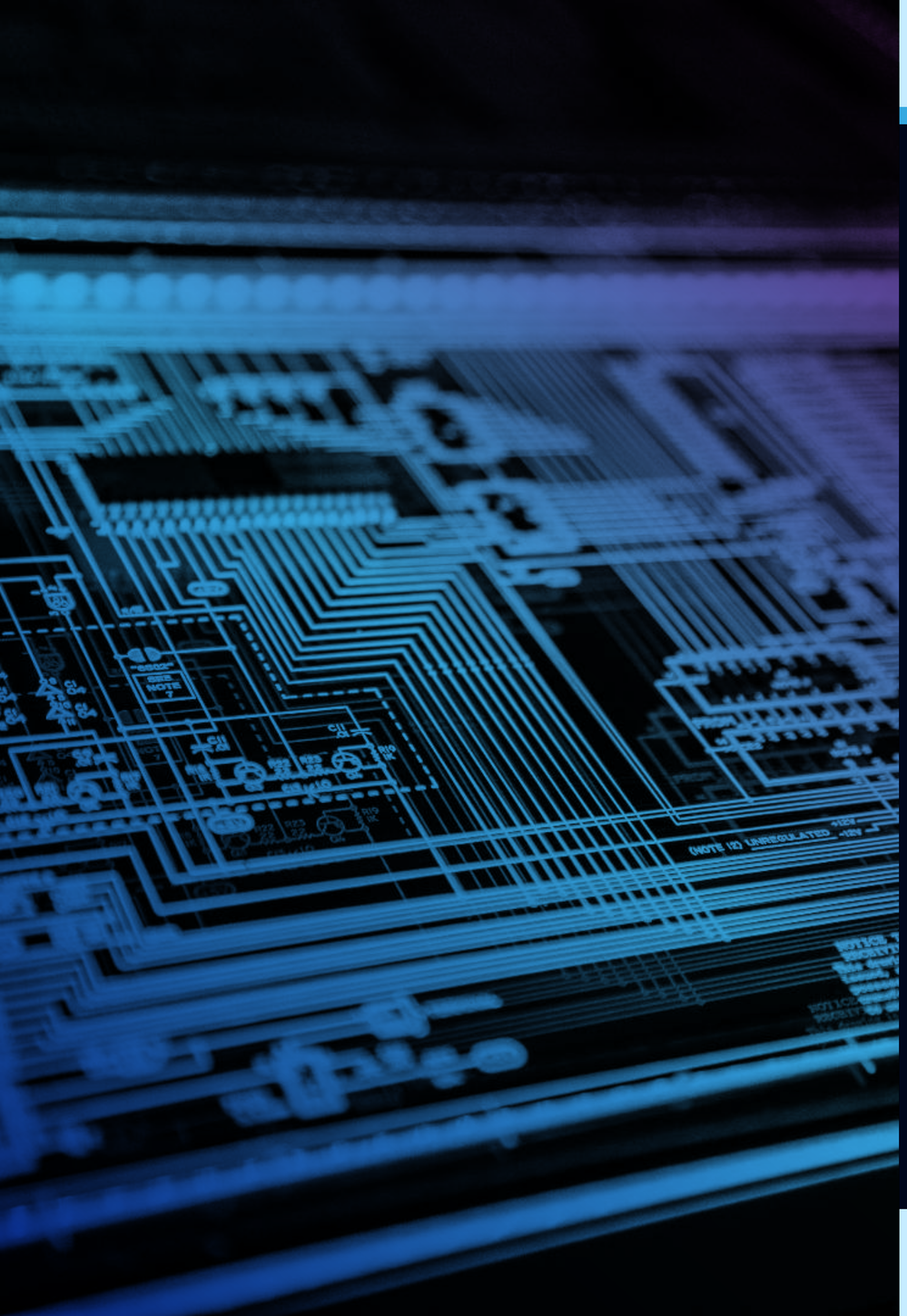
- Bid farewell to naive autoscaling solutions with AI-SPRINT's advanced resource allocation to handle load variations.
- Optimize energy and cloud operational costs with energy-aware runtime migration capabilities.

For End Users:

- Place your trust in AI-SPRINT's secure computing and storage environments, even with untrusted providers.
- Achieve a higher level of data anonymization for enhanced data protection.

Empowering Your AI Journey

AI-SPRINT overcomes existing technological challenges for designing and executing AI applications, particularly in the edge-to-cloud continuum. Our Python-based AI applications, with extensive AI technology integration, enable seamless operation across cloud servers, edge servers, and AI-enabled sensors. With AI-SPRINT Studio, we abstract applications from underlying computing resources, allowing developers to focus solely on algorithms and application logic.



Key Components of AI-SPRINT Design Tools

Design and Programming Abstractions:

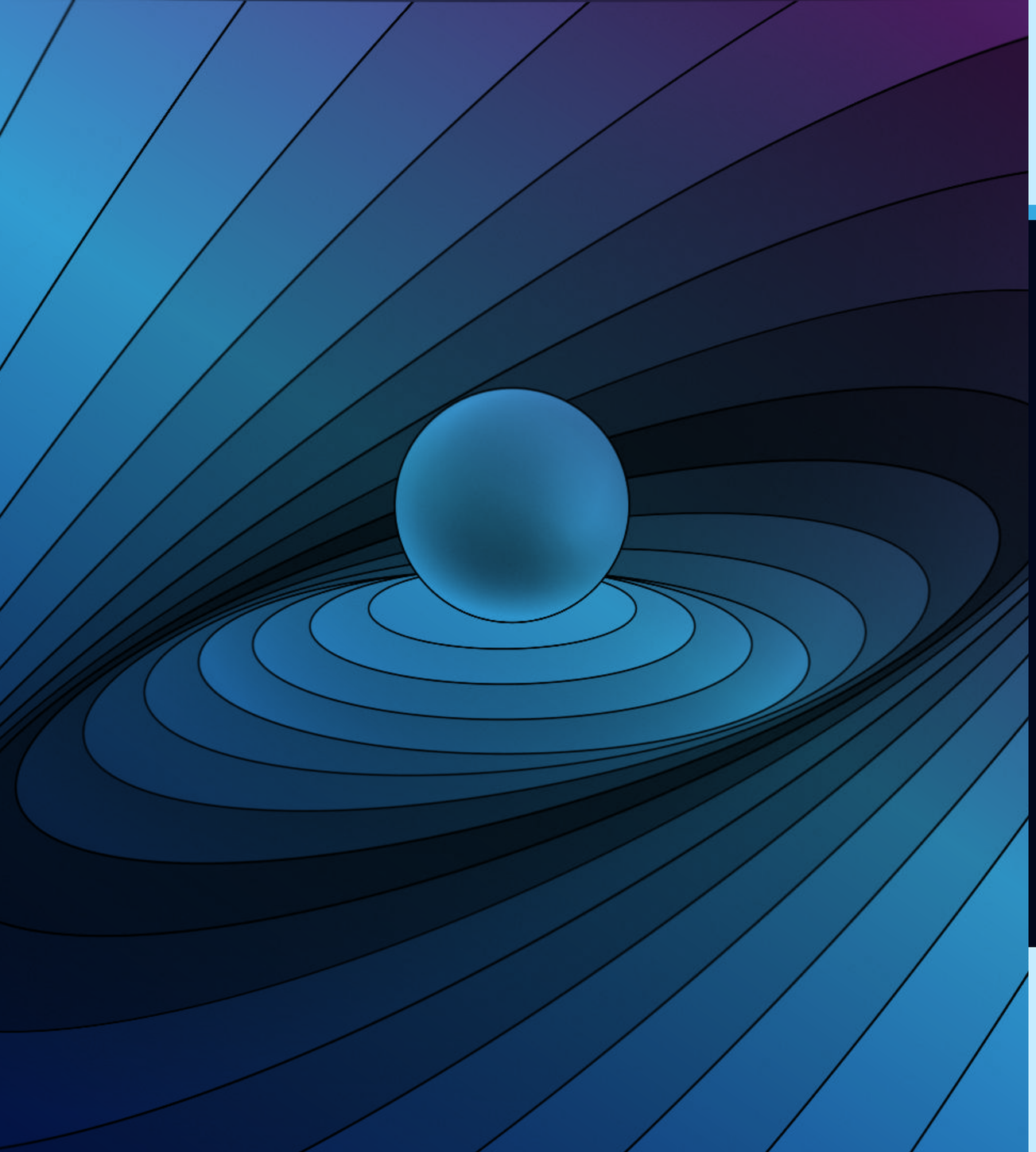
- Effortlessly parallelize compute-intensive tasks using specialized resources like GPUs and AI-enabled sensors.
- Enrich code with high-level annotations for QoS constraints, code dependencies, and performance parameters for efficient task allocation.

The TOSCARIZER:

- Generate TOSCA documents for optimal and base component placement, streamlining the virtual infrastructure provisioning process.
- Trigger the creation and destruction of virtual infrastructures using the Infrastructure Manager.

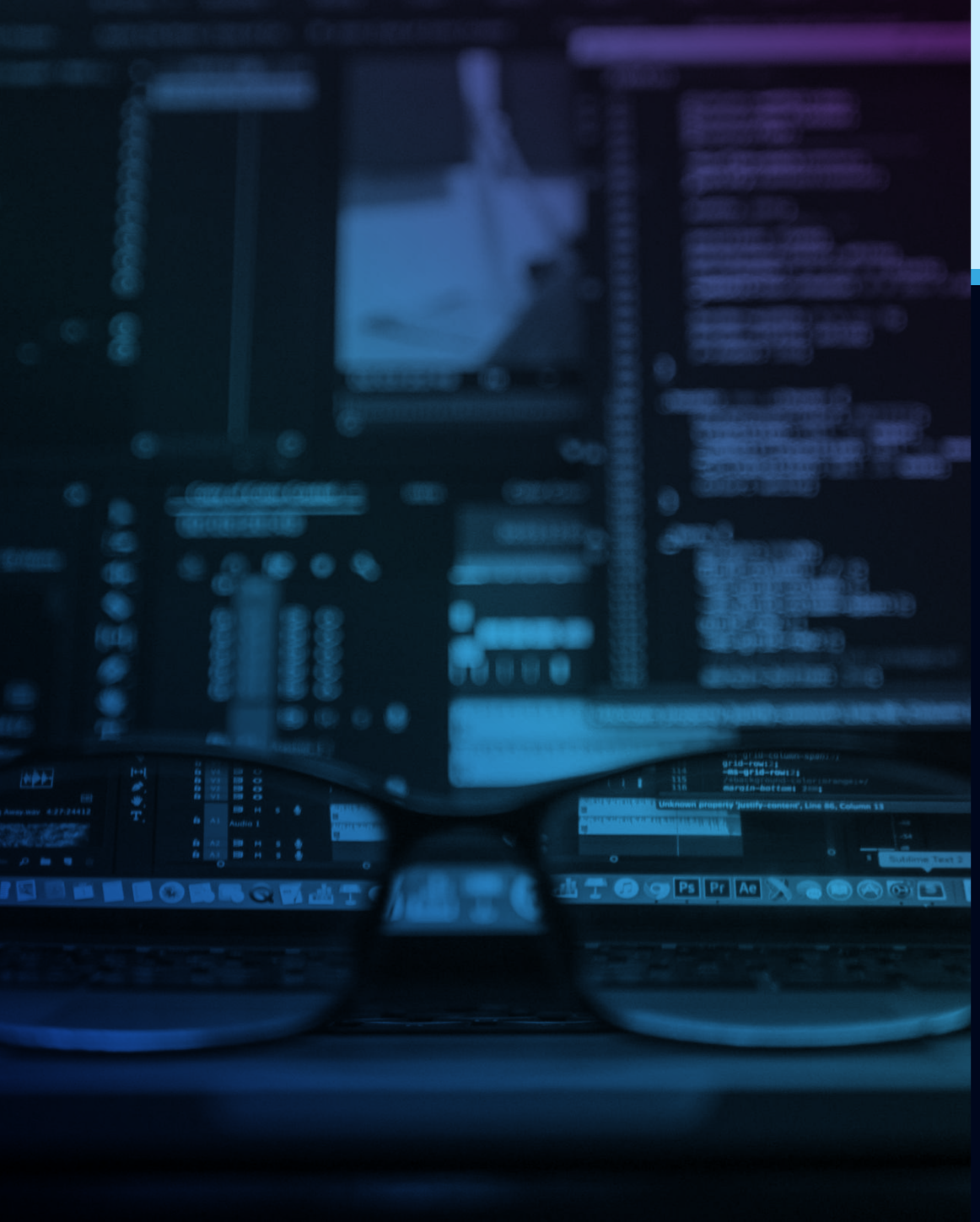
Performance Models and Design Space Exploration:

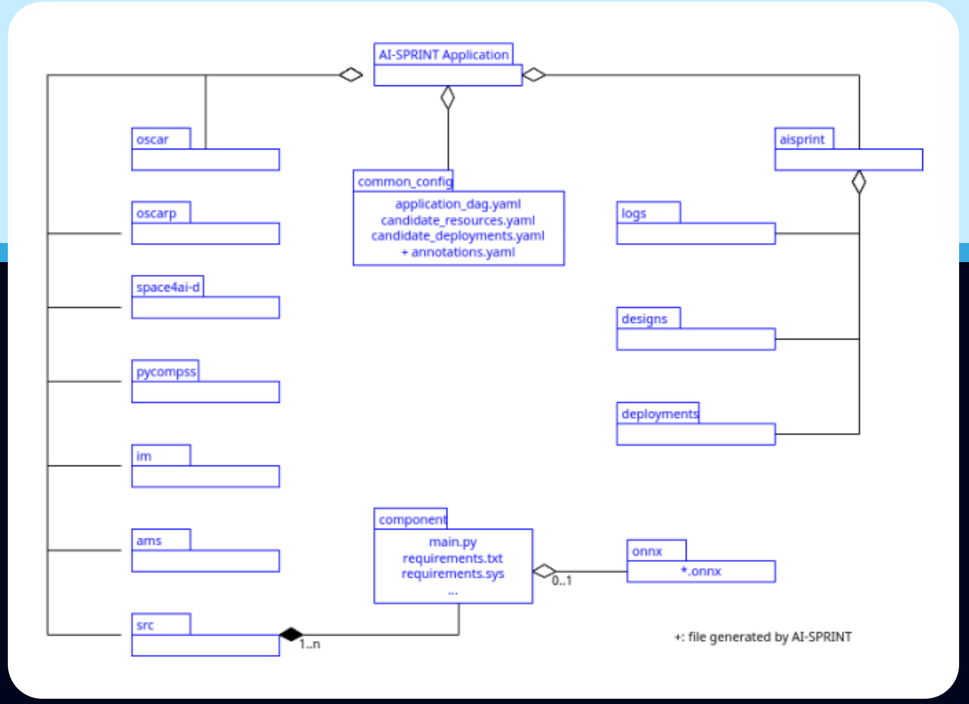
- AI-SPRINT's performance modeling approach, powered by ML and aMLLibrary, ensures accurate performance prediction.
- Identify optimal resource selection and component placement for enhanced resource efficiency and cost reduction with SPACE4AI-D.



Redefine Your AI Potential with AI-SPRINT

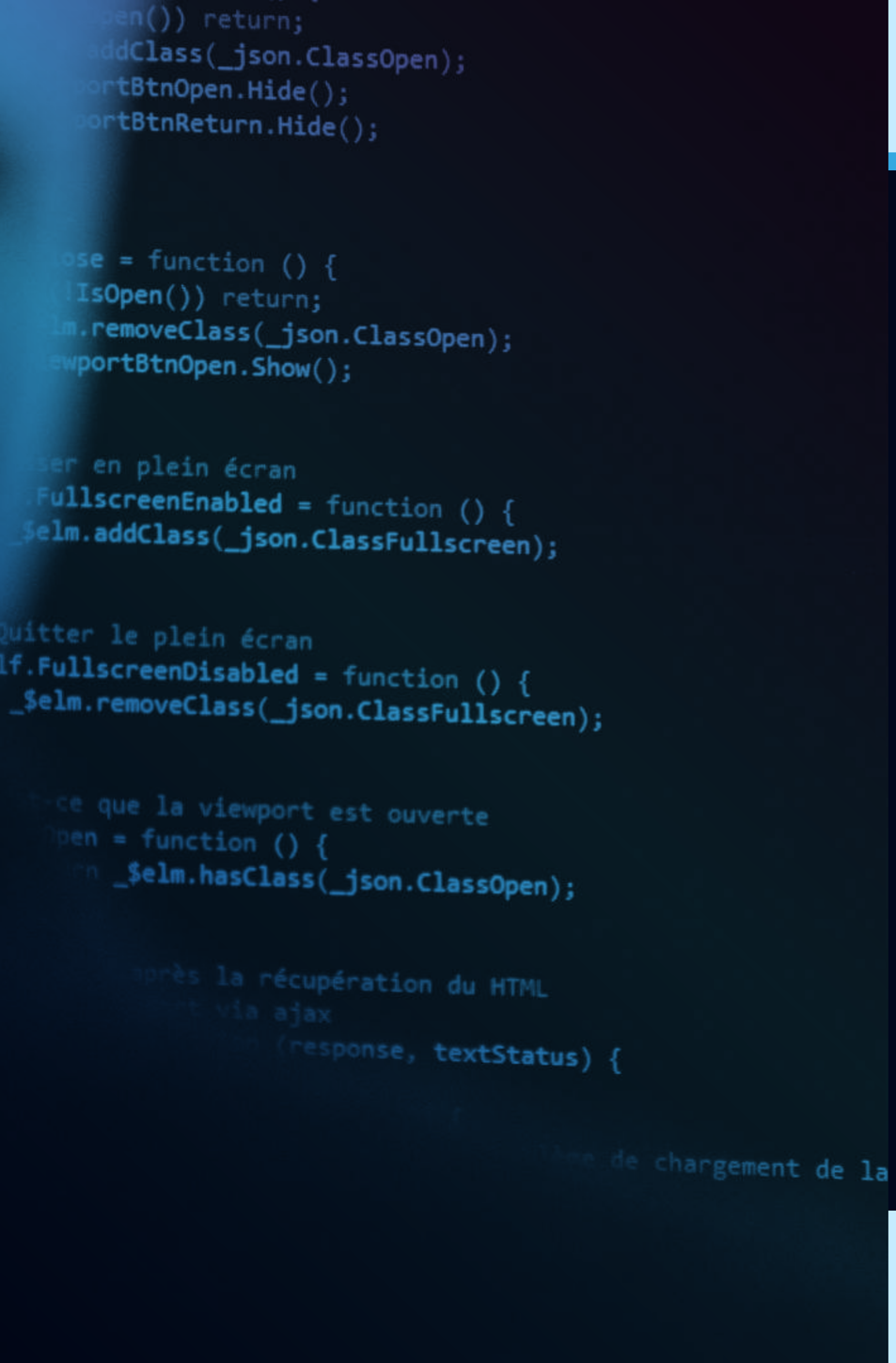
Our solutions produce tangible artifacts, including TOSCA descriptions, application components/partitions images, and application performance models. Trust in AI-SPRINT to revolutionize your AI application development and deployment, optimizing performance, ensuring data privacy, and positioning your business for unmatched success in the AI-driven landscape.





Streamlining ML-Based Application Development

AI-SPRINT empowers businesses to efficiently design and deploy machine learning (ML) applications. Our user-centric approach involves application programmers and SysOps as the primary stakeholders. They collaborate by providing essential application code and configuration files. This collaborative effort adheres to a well-structured framework



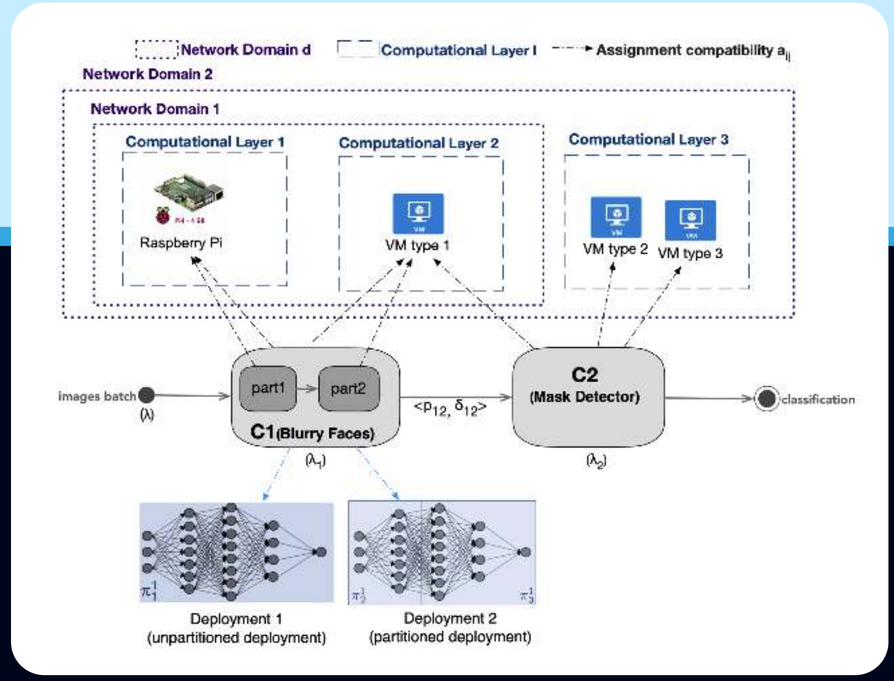
Application DAG and Components

Within our system, the dependencies of application components are managed through an Application Directed Acyclic Graph (DAG). This DAG comprises:

- Vertices: Representing application components, each vertex signifies a fundamental element of the application.
- Edges: These connections denote dependencies between components, defining the execution sequence. Each edge is represented as a tuple, including source and destination components, along with transition probabilities. For instance, consider an application with two consecutive components, as defined in the dag.yaml file:

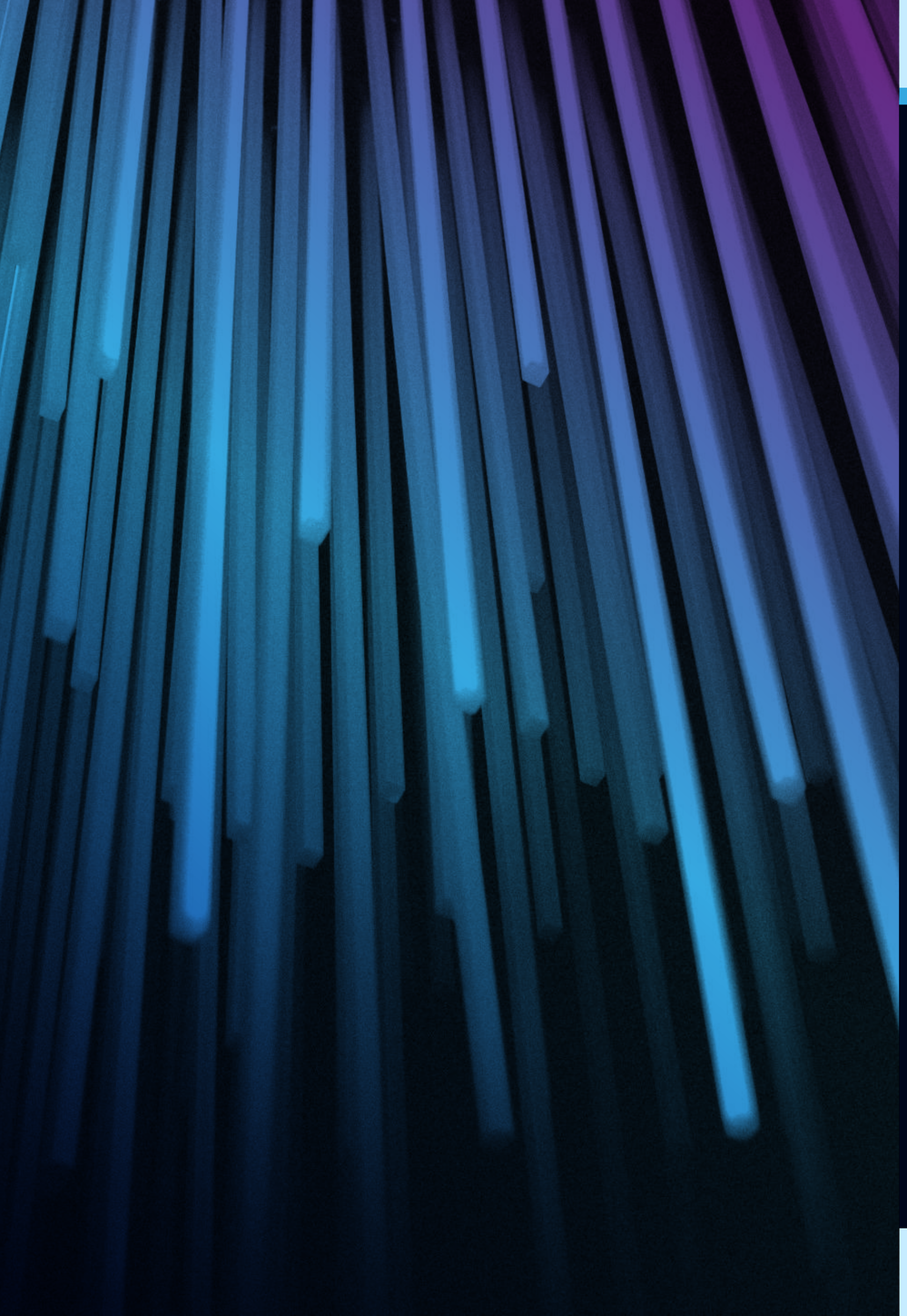
yaml
DAG YAML file
System:
name: Example Application
components: [component1, component2]
dependencies: [[component1, component2, 1]]

In this example, component1 must complete before component2 begins, with a transition probability of 1.



Candidate Resources and Deployments

AI-SPRINT provides businesses with the flexibility to define the resources available for running application components. This resource allocation involves explicitly associating each component with suitable candidate resources. This resource allocation is governed by two key files: candidate_resources.yaml and candidate_deployments.yaml.



Candidate Resources

- **Network Domains:** These encompass various computational layers, each with unique network communication properties. They establish connections between devices, ensuring seamless data flow.
- **Computational Layers:** Each network domain comprises computational layers, including physical devices, virtual resources (e.g., cloud VMs), or native cloud functions (e.g., AWS Lambda). Layer types determine resource allocation during deployment.
- **Resource Attributes:** Each resource is defined by attributes like name, description, processors, and more, specifying resource capabilities and specifications.

Candidate Deployments

- **Components:** The candidate_deployments.yaml file plays a central role, linking application components with compatible resources specified in candidate_resources.yaml. The Components section in this file defines containers for each component and their respective resource assignments.
- **Candidate Execution Layers:** AI-SPRINT allows components to be deployed across diverse computational layers, facilitating optimal resource selection during deployment.
- **Resource Specifications:** Each component's container is configured with key attributes like memory size and computing units, detailing its resource requirements.
- **Security Parameters:** AI-SPRINT bolsters security with trustedExecution for memory encryption, networkProtection for encrypted communications, and fileSystemProtection for automatic file encryption on untrusted storage

Redefine Your AI Potential with AI-SPRINT

Our solutions produce tangible artifacts, including TOSCA descriptions, application components/partitions images, and application performance models. Trust in AI-SPRINT to revolutionize your AI application development and deployment, optimizing performance, ensuring data privacy, and positioning your business for unmatched success in the AI-driven landscape.

Consortium

POLITECNICO MILANO 1863

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

TECHNISCHE UNIVERSITÄT DRESDEN

GREGOIRE

skaylink

CLOUD & HEAT

TTAnalysis

Trust-IT Services
communicating to markets

IDC
ANALYZE THE FUTURE

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

7bulls.com

Scan the QR Code
and visit our website

AI SPRINT

Follow us
on our social

The AI-Sprint project has received funding from the European Union Horizon 2020 research and innovation programme under Grant Agreement No. 101016577.