Integrating NVIDIA AI Microservices with the Eclipse Arrowhead framework for smart city applications

Eduard Cristian POPOVICI,
POLITEHNICA Bucharest, Splaiul Independentei Street, No. 313, 6th District, 060042 Bucharest, Romania
eduard.popovici@upb.ro

Octavian FRATU,
POLITEHNICA Bucharest, Splaiul Independentei Street, No. 313, 6th District, 060042 Bucharest, Romania
octavian.fratu@upb.ro

Alexandru VULPE,
POLITEHNICA Bucharest, Splaiul Independentei Street, No. 313, 6th District, 060042 Bucharest, Romania
alexandru.vulpe@upb.ro

Razvan CRACIUNESCU,
POLITEHNICA Bucharest, Splaiul Independentei Street, No. 313, 6th District, 060042 Bucharest, Romania
razvan.craciunescu@upb.ro

Andra Paula AVASILOAIE,
POLITEHNICA Bucharest, Splaiul Independentei Street, No. 313, 6th District, 060042 Bucharest, Romania
andra.avasiloaie@stud.etti.upb.ro

Abstract

In order to provide a solid, scalable solution for edge AI applications suited to smart city projects, this article suggests an architecture that combines NVIDIA AI Microservices with the Eclipse Arrowhead Framework. The integration addresses the demand for smooth, real-time AI-powered functions across heterogeneous devices and serves a variety of sectors, including social innovation, urban planning, and e-government. The framework seeks to improve citizen services and optimize urban resource management by utilizing Arrowhead's service-oriented skills and NVIDIA's cutting-edge AI models. As seen by applications like automated systems and industrial IoT, the study expands on developments in cloud-edge integration and service orchestration within the Arrowhead Framework. Few existing frameworks have specifically addressed the integration of high-performance AI microservices for smart city contexts, instead concentrating on general interoperability and dynamic service discovery. By using Docker for containerization, the suggested approach makes it possible to deploy AI services in a secure and scalable manner. While Arrowhead manages service registration, discovery, and secure communication, NVIDIA AI models take care of activities like data analysis and pattern identification. Workloads are balanced across cloud and edge settings because of the architecture's support for decentralized execution. The successful orchestration of AI microservices for applications such as environmental monitoring and traffic optimization is demonstrated by the preliminary implementation. Through simulated urban scenarios, the system's ability to process data with minimal latency and make dependable decisions across heterogeneous platforms is tested. By providing a model for improving urban infrastructure, this framework can greatly increase the effectiveness of smart city operations for both practitioners and scholars. Additionally, it establishes the framework for incorporating upcoming advancements in AI into public services. To guarantee compatibility, scalability, and security, the study presents a novel method of integrating Arrowhead's orchestration tools with NVIDIA's AI Microservices. The framework provides a creative answer to contemporary urban problems by considering the particular requirements of smart cities.

Keywords: smart infrastructure management, urban AI solutions, IoT interoperability, edge-to-cloud integration, NVIDIA GPU cloud model integration (NMI)

1. Introduction

While urban settings are becoming more complicated and there is a growing need for sustainable, effective solutions, smart cities are at the forefront of technological innovation. To better resource management, public services, and citizen well-being, these cities mostly rely on cutting-edge technologies.

With the processing capacity required to evaluate big volumes of information, automate decision-making procedures, and provide real-time solutions, artificial intelligence (AI) has emerged as an important enabler in this transition.

Simultaneously, the Internet of Things (IoT) has connected disparate devices to gather and share data, extending the networked infrastructure of cities. However, there are a number of obstacles to overcome when incorporating AI capabilities into these IoT- enabled ecosystems, such as interoperability, scalability, and safe orchestration across many platforms. In order to solve them, strong frameworks that facilitate smooth communication and effective resource use between edge and cloud computing environments are needed.

In order to meet the particular requirements of smart city applications, this article presents a novel architecture that combines the Eclipse Arrowhead Framework with NVIDIA AI Microservices. By bridging the gap between service-oriented frameworks and high-performance AI models, this integration makes improved urban services possible in fields like social innovation, urban planning, and e-government. The suggested solution makes use of cutting-edge orchestration techniques and containerization technologies to guarantee the safe, scalable, and real-time deployment of AI services.

In addition to expanding on developments in edge-cloud integration and service-oriented architectures, this study fills a knowledge vacuum regarding the application of AI microservices in smart city settings. By providing new avenues for scholars, practitioners, and politicians to develop smarter, more connected cities, the framework seeks to transform urban infrastructure.

In order to improve industrial cyber-physical systems (CPS) through automation and interoperability, the authors of [1] presented a framework for incorporating artificial intelligence (AI) capabilities into the Eclipse Arrowhead Framework. The creation of an AI Toolbox that enables service-based operations in safety-critical industrial use cases is highlighted in the paper. Our research focuses on smart city applications, with a particular emphasis on orchestrating NVIDIA AI Microservices for urban planning and e-government scenarios, whereas their work focuses on generic industrial applications.

The authors of [2] addressed dynamic data flow management and service composition in their design for integrating edge and cloud services into industrial IoT systems. They employ production and condition monitoring use cases to validate their methodology. Our work applies these concepts to smart city domains, utilizing NVIDIA's AI capabilities for real-time decision-making in urban settings, while their concentration is on industrial IoT with dynamic service orchestration.

In [3], the authors created a safe onboarding process for Internet of Things (IoT) devices using the Eclipse Arrowhead Framework in a System of Systems (SoS) setting. It places a strong emphasis on building trust using secure communication protocols and a chain of certificates. Our research expands on Arrowhead's onboarding concepts to guarantee the safe integration and orchestration of NVIDIA AI Microservices in distributed smart city systems, even though this work focuses on secure onboarding procedures.

In order to dynamically orchestrate and choreograph operations in CPS, the authors of [4] investigated automating engineering toolchains using the Eclipse Arrowhead Framework. Use cases from the industrial toolchain validate the methodology. While dynamic orchestration is used in both works, our study focuses on smart city applications rather than engineering procedures and applies these ideas to the deployment of NVIDIA AI services in distributed, real-time urban contexts.

Using the Arrowhead Framework to integrate edge and cloud computing, the study explores service orchestration for machine learning-based object identification in industrial vehicles in [5]. Reliable, low-latency operations are the main focus. We generalize and expand the orchestration framework to include larger smart city applications, utilizing NVIDIA's AI Microservices to improve a variety of urban services including traffic management and social innovation, even if they focus on object recognition in industrial vehicles.

The rest of the paper is organized as follows. The next section provides an overview of the Eclipse Arrowhead Framework, highlighting its recent advancements towards version 5.0, with a focus on cloud-edge integration and service orchestration. Section 3 explores NVIDIA AI Microservices, emphasizing their relevance to smart city applications and their integration potential. Section 4, starts with some lessons learned from prior research and outlines our proposed architecture, detailing the integration of Arrowhead and NVIDIA AI for enabling advanced urban applications. Section 6 concludes the paper by summarizing our contributions and outlining future research directions.

2. Eclipse arrowhead framework: advancements towards version 5.0

One important service-oriented architecture (SOA) that was created to solve the problems of security, scalability, and interoperability in distributed systems is the Eclipse Arrowhead Framework. The framework, which originated in industrial automation, has developed to enable a broad range of applications, such as cyber-physical systems (CPS), IoT ecosystems, and smart city infrastructures.

The framework offers significant enhancements in service orchestration, cloud-edge integration, and dynamic flexibility as it moves closer to version 5.0. With the help of these improvements, developers may create reliable, effective systems of systems (SoS) that can easily coordinate services across several platforms. Starting with a summary of the framework and its guiding principles, the ensuing sections explore the salient features of this development.

Overview of the eclipse arrowhead framework

A service-oriented architecture (SOA) called the Eclipse Arrowhead Framework [6] was created to make it easier to develop and implement automation and digitalization solutions in a variety of fields, including as smart cities, industrial automation, and the Internet of Things (IoT). It facilitates the development of complex systems of systems (SoS) by encouraging interoperability and integrability through late binding methods and loosely connected services.

The utilization of self-contained local clouds is a key idea in the Arrowhead Framework. In terms of real-time data processing, data and system security, automation system engineering, and automation system scalability, these local clouds offer enhancements and assurances.

Service Registry, Orchestrator, and Authorization are required fundamental technologies that oversee service registration, discovery, orchestration, and security in every local cloud. Inter-cloud communication and sophisticated service management are made possible by supporting key systems like the Event Handler, Gatekeeper, and Gateway.

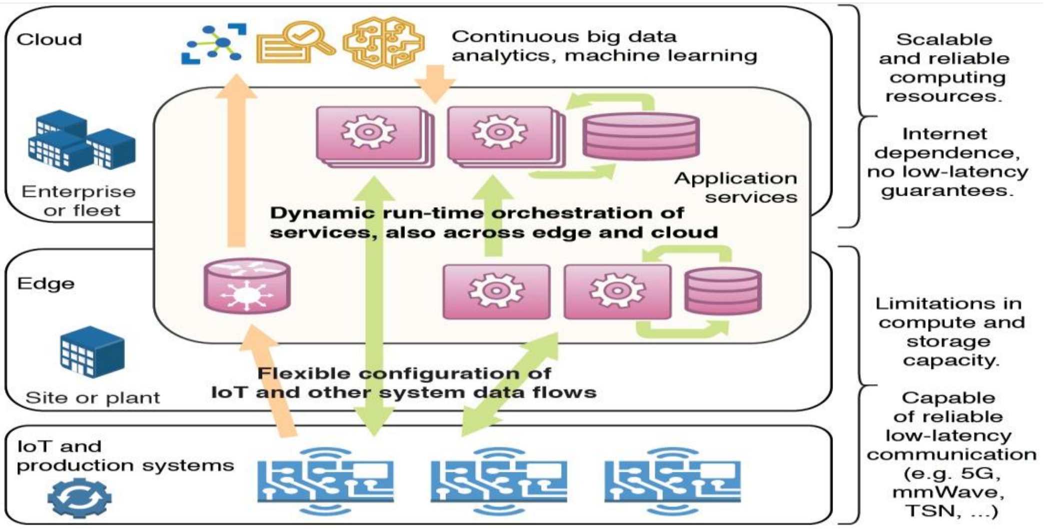
In order to achieve the goals of facilitating industrial IoT and collaborative automation, the framework's architecture is based on a microservices approach and makes use of SOA principles [7]. The abstraction of communication, services enable late (runtime) binding and loose coupling between service providers and users. Higher security, hard real-time speed, scalability, and easier engineering are made possible by the introduction of the local cloud idea.

Numerous industrial control systems, including distributed control systems (DCSs) and supervisory control and data acquisition (SCADA), manufacturing execution systems (MESs), programmable logic controllers (PLCs), and Internet of Things (IoT) solutions like building energy management and industrial gateways for smart city applications, have used the Arrowhead Framework [8].

The Eclipse Arrowhead Framework tackles the difficulties of merging disparate systems and technologies by offering a strong and adaptable architecture, opening the door for cutting-edge automation and digitalization solutions in a world growing more interconnected by the day.

2.1. Progress in cloud-edge integration

Significant progress has been achieved by the Eclipse Arrowhead Framework in facilitating smooth cloud-edge integration, which is essential for contemporary distributed systems. With the growing need for real-time data processing and decision- making in IoT and industrial applications, the framework offers tools to optimize workload distribution between cloud and edge settings. Even in environments with limited resources, this integration guarantees low latency, scalability, and reliable performance [2] [3].



Cloud
Continuous big data analytics, machine learning
Enterprise or fleet
Application services
Dynamic run-time orchestration of services, also across edge and cloud
Edge
Site or plant
Flexible configuration of IoT and other system data flows
IoT and production systems
Scalable and reliable computing resources.
Internet dependence, no low-latency guarantees.
Limitations in compute and storage capacity.
Capable of reliable low-latency communication (e.g. 5G, mmWave, TSN, ...)

Fig. 1. Dynamic service orchestration across heterogeneous cloud and edge in the Arrowhead Framework. Source: Dynamic Execution of Engineering Processes in Cyber-Physical Systems of Systems Toolchains [4]

The Orchestrator and Gatekeeper, two of the framework's key services, have been improved to handle safe data flows and service dependencies across cloud and edge levels. These improvements guarantee that mission-critical applications can function flawlessly in a variety of contexts and allow real-time synchronization. Such capabilities are essential for situations like industrial automation and smart city infrastructure that call for autonomous system coordination [5], [9].

These services can be implemented outside of the car since cloud connectivity can become so commonplace and dependable. Furthermore, as suggested by [2], connection between other cars and the roadside infrastructure may play a significant role in enabling autonomous driving (Fig. 2).

Additionally, Arrowhead offers a uniform deployment methodology for cloud and edge applications through its use of containerization technologies like Docker. By making system growth and replication easier, this method frees engineers to concentrate on usefulness rather than intricate infrastructure. The Arrowhead Framework maintains its position as a foundation for service-oriented architectures in dynamic, distributed environments by utilizing these developments.

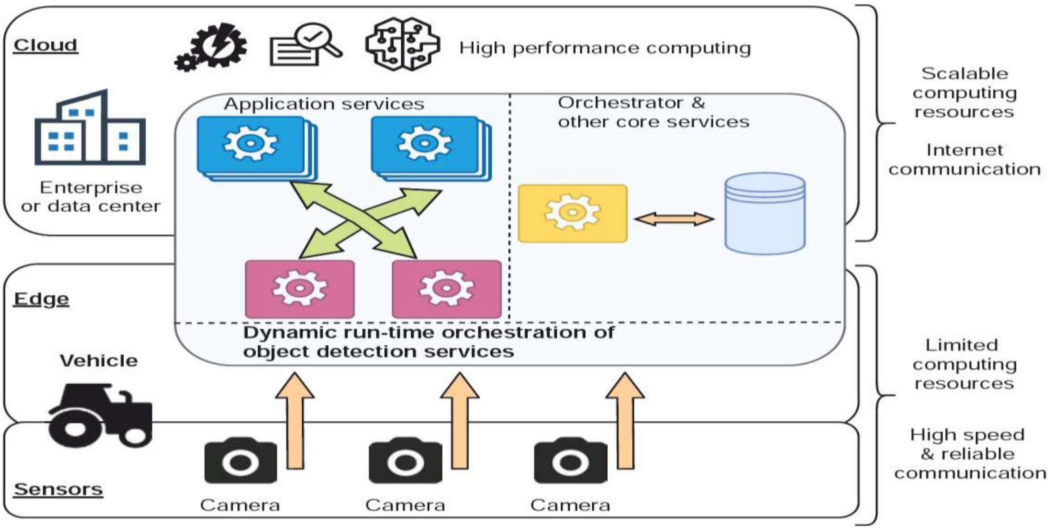


Fig. 2. High level overview of a hybrid cloud’s SOA for object detection task.
Source: Dynamic Edge and Cloud Service Integration for Industrial IoT and Production Monitoring Applications of Industrial Cyber-Physical Systems [2]

2.2. Service orchestration enhancements

The service orchestration features of the Eclipse Arrowhead Framework have been greatly improved to satisfy the needs of contemporary distributed systems. The framework's service orchestration now allows for dynamic service binding, allowing for real-time adjustment to shifting operational needs. This is accomplished by means of sophisticated orchestration methods that enable the automatic construction of services across heterogeneous platforms, guaranteeing the smooth interoperability of varied applications [4] [9].

The incorporation of adaptable orchestration procedures that maximize resource allocation across edge and cloud settings is a significant improvement over version 5.0. Upgrades to the framework's Orchestrator enable dynamic management of intricate service dependencies (Fig. 3). Applications needing fault tolerance and high availability, such industrial automation systems and smart city infrastructures, especially depend on this capability [5].

Additionally, the Arrowhead Framework now supports orchestration and service choreography, allowing services to be coordinated remotely without the need for a centralized control point. Large-scale IoT and CPS installations benefit greatly from this decentralized approach's increased scalability and robustness. The framework's capacity to coordinate services across organizational and geographic borders is further expanded by improved event handling and inter-cloud communication modules [3] [4].

3. NVIDIA AI Microservices: innovations for smart cities

Advanced AI capabilities are essential as smart cities develop in order to handle massive volumes of data and make choices in real time. To meet these needs, NVIDIA AI

Microservices offers scalable, modular solutions that make use of state-of-the-art AI models that are optimized for GPU deployment.

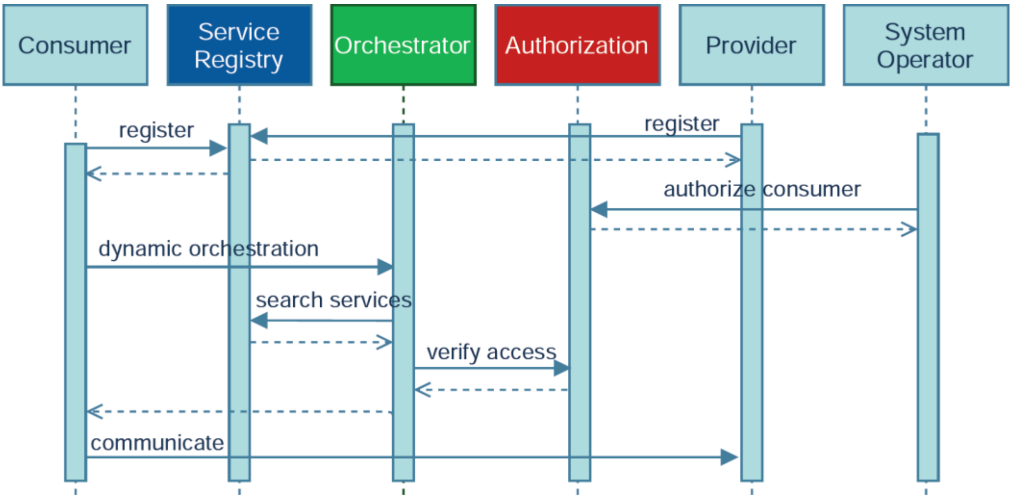


Fig. 3. Sequence of dynamic orchestration process using the Arrowhead Framework.
Source: Service Orchestration for Object Detection on Edge and Cloud Industrial Vehicles [5]

Through the NVIDIA GPU Cloud (NGC) platform, developers gain access to a rich ecosystem of pre-trained models, containers, and software development kits, streamlining the integration of AI into urban applications. Innovations which enable seamless deployment of AI services across edge and cloud environments, fostering data-driven decision-making in domains like e-government, urban planning, and social innovation.

3.1. NVIDIA NGC and model integration

With its collection of pre-trained models, deep learning containers, and tools tailored for GPU acceleration, NVIDIA NGC acts as a single location for AI developers. These tools make it easier to incorporate AI into a variety of systems, cutting down on development time and complexity. NGC gives users access to models for tasks including environmental analysis, traffic monitoring, and citizen behavior prediction for smart city applications. These models may be refined and implemented on NVIDIA's hardware platforms.

The incorporation of AI models into microservices architecture is a fundamental component of NVIDIA's methodology. Developers can incorporate particular features, such object identification or natural language processing, into stand-alone services thanks to this modular design. Then, by coordinating these services, intricate workflows that tackle the particular difficulties of metropolitan settings can be developed. Because of the microservices architecture's scalability and flexibility, cities can modify AI-driven solutions to meet changing needs.

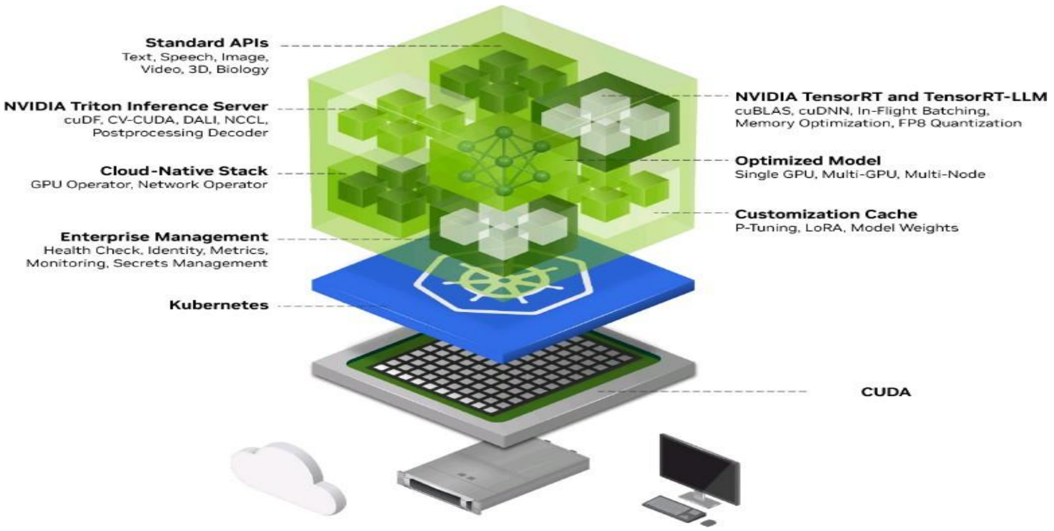

Standard APIs
Text, Speech, Image, Video, 3D, Biology
NVIDIA Triton Inference Server
cuDF, CV-CUDA, DALI, NCCL, Postprocessing Decoder
Cloud-Native Stack
GPU Operator, Network Operator
Enterprise Management
Health Check, Identity, Metrics, Monitoring, Secrets Management
Kubernetes
NVIDIA TensorRT and TensorRT-LLM
cuBLAS, cuDNN, In-Flight Batching, Memory Optimization, FP8 Quantization
Optimized Model
Single GPU, Multi-GPU, Multi-Node
Customization Cache
P-Tuning, LoRA, Model Weights
CUDA

Fig. 4. NVIDIA NIM Microservices Architecture.
Source: Orchestrating Accelerated Virtual Machines with Kubernetes Using NVIDIA GPU Operator [10]

NVIDIA's NIM (NVIDIA Inference Microservices) separates particular AI functions, like object detection and natural language processing, into stand-alone services, as shown in Fig. 4. These microservices can be coordinated to create intricate workflows that tackle the particular difficulties presented by metropolitan settings. Because of the architecture's scalability and adaptability, cities can modify AI-driven solutions to meet changing needs [10].

3.2. Microservices for AI-driven edge applications

Modern applications that need to process and make decisions in real time now rely heavily on the integration of AI microservices at the edge. The deployment of AI models as modular components is made easier by NVIDIA's microservices architecture, which enables certain features—like speech processing, object identification, and picture recognition—to operate effectively on edge devices. In addition to guaranteeing scalability, this design facilitates the integration of several services in intricate processes that are suited for certain use cases such as applications for smart cities, industrial IoT, and driverless cars.

NVIDIA Jetson Nano and Triton Inference Server are used as edge and cloud components in real-world applications, respectively. Triton provides high-performance inference on the cloud, while Jetson Nano provides low-latency on-board computing. Frameworks like as Eclipse Arrowhead facilitate the smooth orchestration of these elements, allowing for secure communication between edge and cloud nodes and dynamic service discovery [5]. The hardware configuration, which links edge and cloud services for real-time object identification, is seen in Fig. 5 and includes a Jetson Nano, Raspberry Pi, and Triton Inference Server.

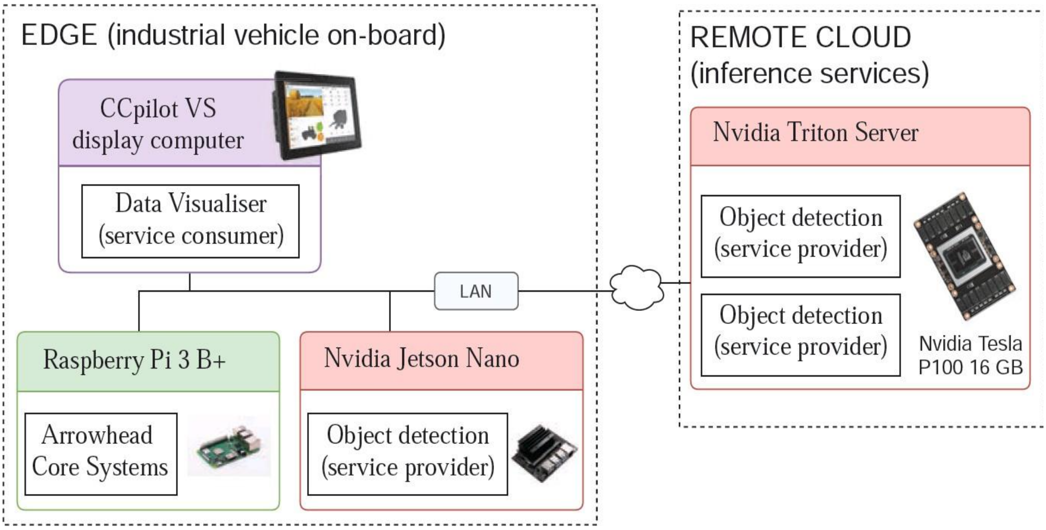

EDGE (industrial vehicle on-board)
CCpilot VS display computer
Data Visualiser (service consumer)
LAN
Raspberry Pi 3 B+
Arrowhead Core Systems
Nvidia Jetson Nano
Object detection (service provider)
REMOTE CLOUD (inference services)
Nvidia Triton Server
Object detection (service provider)
Object detection (service provider)
Nvidia Tesla P100 16 GB

Fig. 5. Hardware and software components for AI-driven edge applications, including NVIDIA technologies. Source: Service Orchestration for Object Detection on Edge and Cloud Industrial Vehicles [5]

.

Fig. 6 shows the Triton Inference Server high-level architecture.

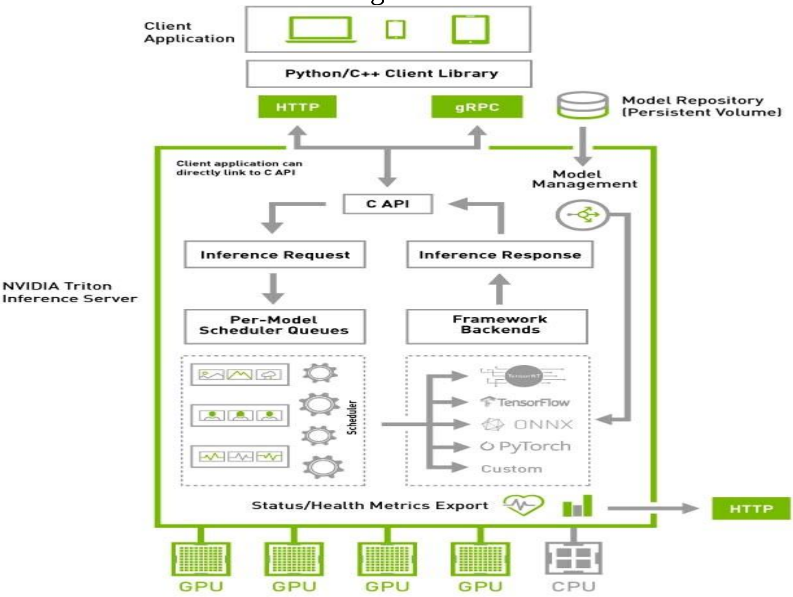

Client Application
Python/C++ Client Library
HTTP
gRPC
Model Repository (Persistent Volume)
Client application can directly link to C API
C API
Model Management
NVIDIA Triton Inference Server
Inference Request
Inference Response
Per-Model Scheduler Queues
Framework Backends
Scheduler
TensorRT
TensorFlow
ONNX
PyTorch
Custom
Status/Health Metrics Export
HTTP
GPU
GPU
GPU
GPU
CPU

Fig. 6. NVIDIA Triton Inference Server architecture for edge-cloud AI workflows. Source: NVIDIA Triton Inference Server. Architecture [11]

Fig. 7 illustrates response times for different configurations, highlighting the performance difference between Jetson Nano (edge) and Triton Server (cloud).

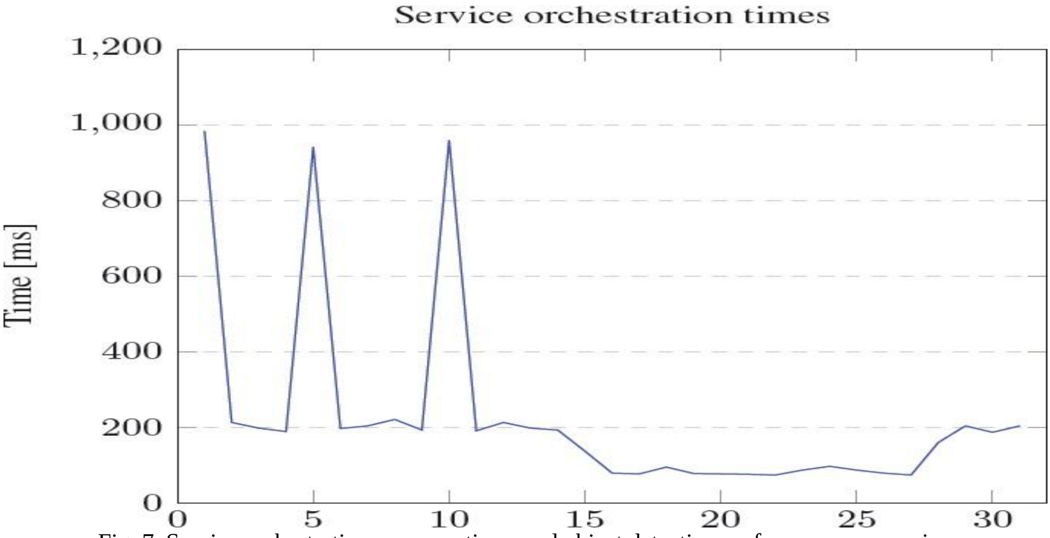


Fig. 7. Service orchestration response times and object detection performance comparison.
Source: Service Orchestration for Object Detection on Edge and Cloud Industrial Vehicle [5].

The models that Triton will make available for inferencing are stored in a file-system- based repository called the model repository. Inference requests are sent to the relevant per-model scheduler after arriving at the server via HTTP/REST, GRPC, or the C API. Several batching and scheduling techniques are implemented by Triton, and they can be set up model-by-model. It is optional for each model's set scheduler to batch inference requests before forwarding them to the appropriate model-type framework backend. To generate the desired outputs, the framework backend uses inferencing with the inputs supplied in the request. After formatting the outputs, a reply is sent.

3.3. Relevance to the urban environments

In order to handle the intricate problems of contemporary cities, such as traffic control, energy optimization, public safety, and environmental monitoring, urban environments are depending more and more on data-driven solutions. These situations are especially well-suited for NVIDIA AI Microservices, which provide scalable, effective, and flexible solutions that work flawlessly in real-time edge and cloud deployments. Through better infrastructure management and more intelligent service delivery, municipalities may improve the quality of life for their residents by incorporating AI capabilities into urban systems.

Certain features, such object detection, natural language processing, and predictive analytics, can be customized for city-specific requirements thanks to NVIDIA's microservices architecture's modular design. For example, the cloud-based Triton Inference Server analyzes aggregated data for long-term urban planning, while AI models operating on NVIDIA Jetson Nano may track traffic flow in real-time at the edge. Cities can manage a variety of workloads with this dual-layered architecture, from quick local decisions to in-depth data analysis.

These microservices can also be integrated with current smart city systems, like IoT networks and urban data hubs, thanks to NVIDIA's support for open standards and interoperability. Cities can embrace cutting-edge AI technologies and fully utilize their current infrastructure thanks to this compatibility. Urban environments can adjust to changing needs with the help of NVIDIA's scalable solutions, encouraging innovation in fields like resource management, social innovation, and e-government.

NVIDIA microservices make a substantial contribution to the development of smarter, more efficient cities by tackling important urban challenges with cutting-edge AI and machine learning capabilities. These solutions create the groundwork for sustainable urban growth and development in addition to improving operational efficiency.

4. Proposed architecture: integrating Arrowhead and NVIDIA AI

The suggested architecture creates a smooth, scalable, and secure solution for smart city applications by fusing the advantages of NVIDIA AI Microservices and the Eclipse Arrowhead Framework. Advanced analytics and real-time decision-making are made possible by NVIDIA's high-performance AI models and microservices, while Arrowhead's strong service orchestration capabilities provide as the foundation for handling service registration, discovery, and secure communication.

This integration optimizes resource allocation and improves heterogeneous device interoperability by utilizing the modularity and flexibility of both frameworks to enable dynamic edge-to-cloud deployments. The architecture offers a potent toolkit for contemporary cities by addressing urban issues like environmental monitoring, transportation optimization, and smart infrastructure management.

4.1. Lessons learned from prior research

The suggested architecture incorporates recognized best practices and fills in noted gaps by largely drawing on insights from earlier work. We discovered the value of service orchestration for dynamic environments through the Eclipse Arrowhead Framework's advances. According to research, elements like secure service discovery and authorization —which are best illustrated in industrial IoT applications—can be successfully modified for urban settings.

A scalable basis for connecting disparate services was made possible by the Arrowhead Framework's support for Docker-based containerized deployments [2] [3]. These ideas are expanded upon in our architecture to address the intricacies of smart cities, where services and devices need to function dependably in large, dispersed settings.

The proven effectiveness of microservices in industrial applications is expanded upon by the incorporation of NVIDIA AI Microservices into smart city workflows. The scalability and performance benefits of dividing computation-intensive jobs across edge and cloud were demonstrated by earlier implementations, such as object detection using NVIDIA Triton Inference Server [5]. Through the integration of these insights, our architecture guarantees that urban AI systems can handle large amounts of data with low latency and remain flexible enough to adjust to changing needs.

Additionally, the importance of striking a balance between local and centralized processing is highlighted by the synergy between edge and cloud computing seen in related studies. For example, the cloud-based Triton Inference Server's computing capability combined with the NVIDIA Jetson Nano's real-time edge inference capabilities provides a versatile framework for applications such as predictive maintenance and traffic flow monitoring [4] [5]. Even in situations when network connectivity is sporadic, operational continuity is guaranteed by this dual-layered strategy.

Finally, the results of the orchestration process evaluation are included into our design. The importance of dynamic orchestration in controlling resource dependencies and boosting resilience in distributed systems has been highlighted in earlier research [4] [5]. Through the integration of NVIDIA's modular AI capabilities with Arrowhead's sophisticated orchestration methods, we guarantee secure, scalable, and resilient smart city solutions that can tackle a range of issues, from environmental sustainability to e- government.

The suggested architecture is based on these lessons, which highlight interoperability, scalability, and adaptability while taking into account the particular requirements of urban settings.

4.2. System components and interoperability

The suggested architecture combines NVIDIA AI Microservices with the Eclipse Arrowhead Framework to produce a unified system designed for smart city applications. Effective data processing and service orchestration across urban areas are made possible by this integration, which guarantees smooth compatibility between numerous components.

The core components are:

- Eclipse Arrowhead Framework: serves as the backbone for service-oriented architecture, managing service registration, discovery, and orchestration. Its core systems —Service Registry, Authorization, and Orchestrator—ensure secure and efficient communication between services.
- NVIDIA AI Microservices: provide modular AI functionalities, such as object detection and natural language processing, optimized for deployment on NVIDIA GPUs. These microservices can be deployed both at the edge and in the cloud, offering flexibility in processing and scalability.
- Edge devices: utilize NVIDIA Jetson Nano modules to perform real-time data processing and inference tasks locally, reducing latency and bandwidth usage.
- Cloud infrastructure: employs NVIDIA Triton Inference Server hosted on powerful GPUs like the Tesla P100, enabling high-performance AI inference for more computationally intensive tasks.

The interoperability mechanisms are:

- Service orchestration: the Arrowhead Orchestrator dynamically manages the interaction between AI microservices and other system components, ensuring optimal resource allocation and service composition.
- Secure communication: the Authorization system within Arrowhead enforces security policies, ensuring that only authorized services can communicate, thereby maintaining data integrity and confidentiality.

- Dynamic service discovery: the Service Registry allows for the real-time discovery of available services, enabling the system to adapt to changes and scale as needed.

Figure 8 depicts the high-level architecture for the integration of the Arrowhead Framework and NVIDIA AI microservices for smart city applications. Edge Devices, such as the Jetson Nano, manage local AI functions, whereas Arrowhead Core Services facilitate service discovery, authorization, and resource coordination via its sub- components: Service Registry, Authorization System, and Orchestrator. The Cloud (Triton Inference Server) delivers high-performance AI processing, while Smart City Applications encompass functions such as traffic control and environmental monitoring.

This design emphasizes seamless edge-cloud interaction for scalable, real-time urban solutions. By leveraging these components and interoperability mechanisms, the architecture facilitates efficient data flow and processing across the smart city ecosystem, enhancing services such as traffic management, environmental monitoring, and public safety.

4.3. The cloud-edge deployment model

By carefully allocating workloads between cloud servers and edge devices, the Cloud-Edge Deployment Model in the suggested architecture aims to maximize data processing and service delivery, as stated in [12]. In order to improve performance, lower latency, and guarantee scalability in smart city applications, this strategy makes use of the computing advantages of both environments.

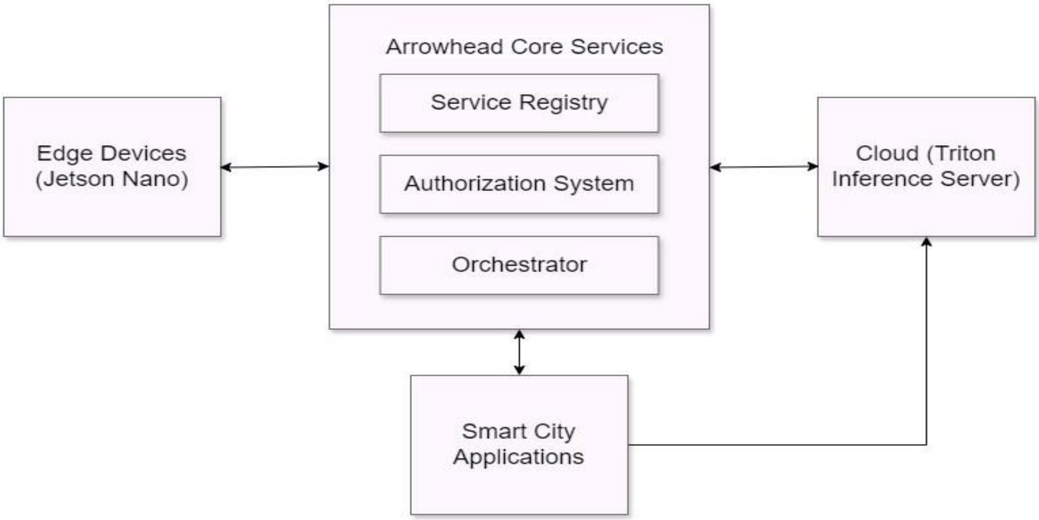


Fig. 8. High Level System Architecture Integrating Arrowhead Framework and NVIDIA AI Microservices
Source: Author's representation of the architecture

The key components of the Cloud-Edge Deployment Model are:

- Edge devices (e.g., NVIDIA Jetson Nano): these devices are deployed close to data sources, such as sensors and cameras, enabling real-time data processing and immediate response actions. By handling tasks like initial data filtering and local inference, edge devices minimize the need to transmit large volumes of raw data to the cloud, thereby conserving bandwidth and reducing latency.

- Cloud infrastructure (e.g., NVIDIA Triton Inference Server): the cloud serves as a centralized hub for intensive computational tasks, large-scale data analysis, and long-term storage. It supports complex AI model training and provides aggregated insights that inform strategic decision-making across the smart city ecosystem.

The operational workflow of the Cloud-Edge Deployment Model is the following:

- Data collection: sensors and IoT devices collect data from various urban environments, including traffic patterns, environmental conditions, and public safety metrics.
- Edge processing: edge devices perform preliminary data processing, such as noise reduction, data normalization, and executing AI inference tasks for immediate insights.
- Data transmission: processed data and relevant metadata are transmitted to the cloud for further analysis. This selective data transfer reduces network congestion and ensures that only valuable information is communicated.
- Cloud analysis: the cloud infrastructure conducts comprehensive data analysis, model training, and cross-referencing with historical data to generate actionable insights and predictive analytics.
- Feedback loop: insights and updated models from the cloud are deployed back to edge devices, enabling them to operate with enhanced intelligence and autonomy.

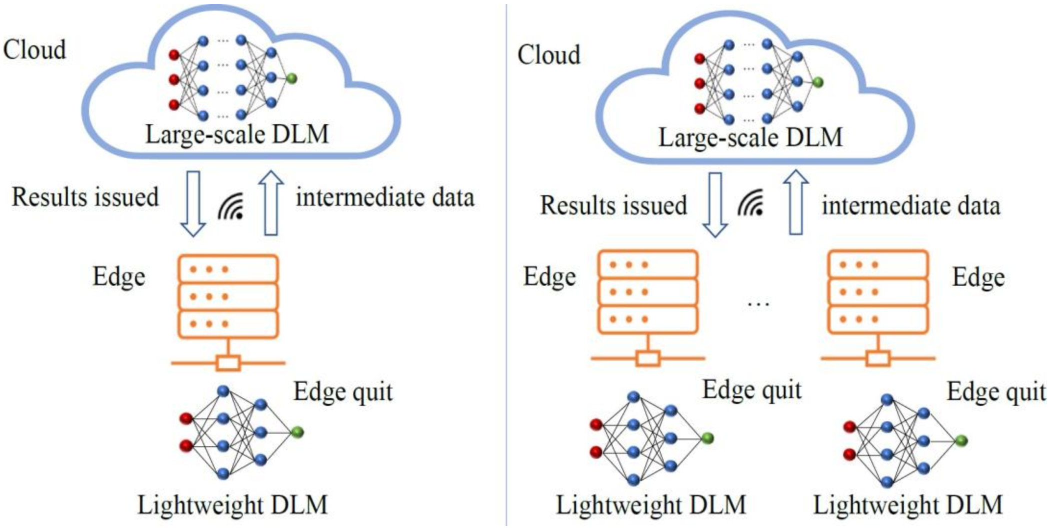


Fig. 9. Cloud-Edge deployment framework for DLMs with (a) an edge node and a local decision-making point before the cloud, (b) more edge nodes, the results being aggregated for decision-making
Source: Cloud-edge-device collaboration mechanisms of deep learning models for smart robots in mass personalization [12]

In the deployment model [12], Fig. 9 shows how cloud servers and edge devices interact, with a focus on data flow and processing activities:

- Edge nodes: performing local inference and pre-processing tasks to reduce bandwidth consumption.
- Cloud servers: managing computationally intensive tasks, including training and large-scale data aggregation.

- Feedback mechanism: highlighting how models and insights flow back from the cloud to the edge for enhanced decision-making.

4.4. Use cases: smart city scenarios

The combination of NVIDIA AI Microservices with the Eclipse Arrowhead Framework provides revolutionary solutions for a range of urban situations in the context of smart cities. Through cutting-edge AI-driven solutions, this collaboration improves resource management, public safety, and operational efficiency.

Traffic management and optimization

On edge devices such as the Jetson Nano, NVIDIA's AI models can be used to analyze real-time traffic data in order to improve signal timings and minimize congestion. In order to provide coordinated responses to changing traffic conditions, the Arrowhead Framework makes it easier for traffic sensors, AI services, and control systems to communicate with one another. Better traffic flow and lower emissions are the results of this integration. License plate identification is one of the biggest problems facing intelligent traffic control systems. Diverse and reliable training data is necessary to build a model that will function across numerous nations and localities with various laws, ordinances, and environmental conditions. SmartCow created an Omniverse extension to create synthetic data in order to supply sufficient and varied training data for the model (Fig. 10) [13] [14].

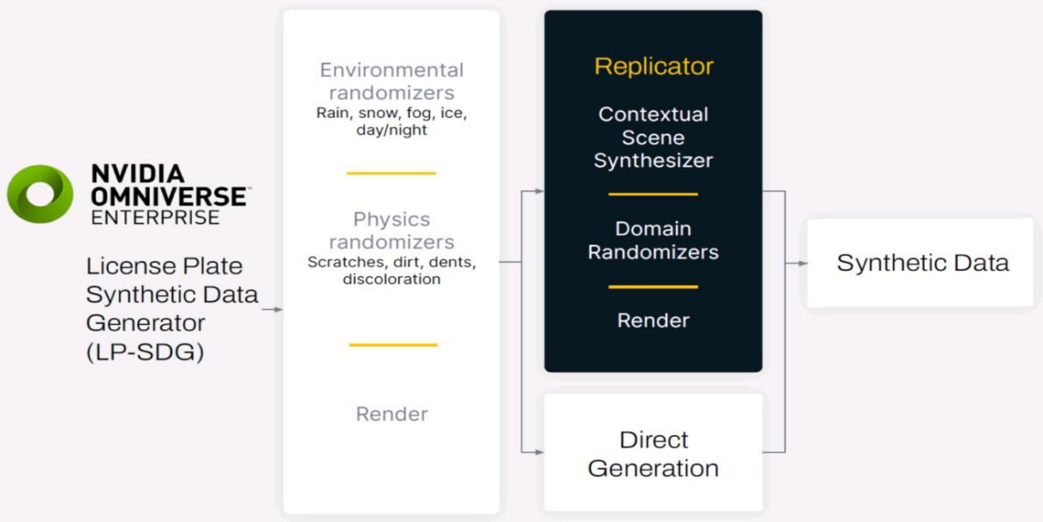

NVIDIA OMNIVERSE ENTERPRISE
License Plate Synthetic Data Generator (LP-SDG)
Environmental randomizers
Rain, snow, fog, ice, day/night
Physics randomizers
Scratches, dirt, dents, discoloration
Render
Replicator
Contextual Scene Synthesizer
Domain Randomizers
Render
Direct Generation
Synthetic Data

Fig. 10. The SmartCow License Plate Synthetic Data Generation workflow in NVIDIA Omniverse. Source: Developing Smart City Traffic Management Systems with OpenUSD and Synthetic Data [13]

Omniverse extensions are reusable parts or instruments that offer strong features to enhance workflows and pipelines. Developers can quickly share an extension they have created in Omniverse Kit to users of Omniverse USD Composer, Omniverse USD Presenter, and other apps. An environmental randomizer and a physics randomizer are used by SmartCow's extension, License Plate Synthetic Generator (LP-SDG), to increase the diversity and realism of synthetic datasets (Fig. 11).

Rain, snow, fog, dust, and other meteorological conditions are all simulated by the environmental randomizer in the digital twin environment. Scratches, dirt, dents, and discolouration that can impair the model's ability to identify the license plate number are simulated using the physics randomizer.

The light sources, textures, camera angles, and materials were all changed using domain randomization after the scene was created. The integrated Omniverse Replicator APIs were used to automate complete this full operation. Bounding box annotations and other output variables required for training were included in the exported data. Three thousand actual photos were used to train the original model. Understanding the baseline model's performance and verifying elements like accurate bounding box measurements and light variation were the objectives.

Public safety and surveillance

Enhancing public safety through real-time object detection and anomaly recognition is made possible by integrating NVIDIA's AI capabilities with citywide surveillance systems. The Arrowhead Framework guarantees safe and effective communication between emergency response teams, AI processing units, and surveillance cameras. Rapid incident detection and resource deployment are made possible by this architecture.

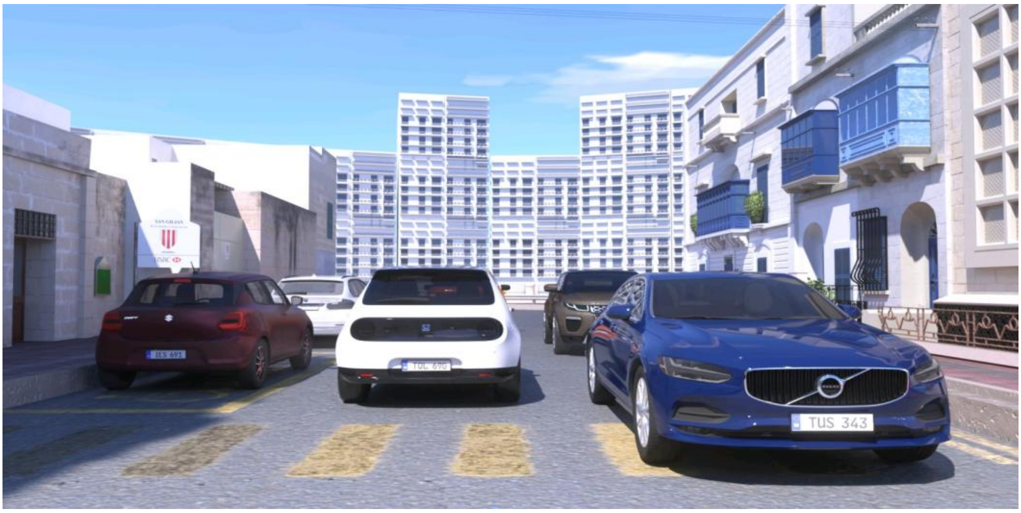


Fig. 11. Synthetically generated vehicles and license plates in NVIDIA Omniverse.
Source: Developing Smart City Traffic Management Systems with OpenUSD and Synthetic Data [13]

NVIDIA Metropolis and NVIDIA AI Blueprints are reference workflows for generative AI use cases that are utilized by both Lenovo Edge Guarding and Lenovo VINA [14]. The NVIDIA AI Blueprint for video search and summarization was unveiled today in advance of the Smart City Expo World Congress in Barcelona. It allows the deployment of AI agents that have been taught to evaluate visual data, compile vast amounts of data, and generate responses to user queries. By analyzing vast amounts of live or recorded footage, these visual agents may quickly search for incidents and detect and prevent action in any area equipped with sensor infrastructure. It is immediately applicable in smart city applications for public safety, parking and curbside management, and traffic management, and it is well-suited for sectors such as manufacturing, retail, and healthcare.

Real-time monitoring and predictive analysis of urban systems, as emphasized by [15], allow cities to foresee disturbances, enhance resource allocation, and execute preemptive measures. Integrating IoT sensor data with advanced analytics enhances operational efficiency and safety, proving crucial for effective infrastructure management and swift emergency response in contemporary urban settings.

Environmental monitoring

Environmental factors like temperature, noise levels, and air quality may be continuously tracked with the use of AI-powered sensors placed across the city. Real-time data aggregation and analysis are made possible by the orchestration of various services made possible by the Arrowhead Framework. This arrangement facilitates prompt actions, such issuing public health alerts or modifying public transit schedules during times of heavy pollution.

Today, a variety of initiatives to monitor and safeguard our planet are powered by aerial imagery from satellites and drones, which is accelerated by AI and NVIDIA GPUs [14] [16]. See how businesses in the NVIDIA Inception program are utilizing AI and aerial images to monitor thawing permafrost in the Arctic, stop natural gas leaks, and track global deforestation on the 50th anniversary of Earth Day. Inception is a virtual accelerator that provides essential tools to help product creation, prototyping, and deployment for businesses in data science and artificial intelligence.

Orbital Insight, a Palo Alto, California-based company, analyzes radar and satellite photos using convolutional neural networks for supply chain monitoring, real estate, mapping, and infrastructure. Through Amazon Web Services, NVIDIA GPUs speed up its geospatial AI algorithms. A 100x inference speedup on large satellite photos is made possible by the team's ability to grow and downsize their use of GPUs in the cloud, according to Manuel Gonzalez-Rivero, senior computer vision scientist at the business.

5. Conclusions and future directions

In order to address important issues in smart city applications, this article proposed an architecture that blends NVIDIA AI Microservices with Eclipse Arrowhead Framework. The suggested system makes use of NVIDIA's high-performance AI models and Arrowhead's powerful service orchestration capabilities to facilitate secure connectivity, real-time data processing, and smooth integration between cloud and edge settings.

The integration of these technologies illustrates the capacity for developing intelligent, adaptive urban infrastructures that can dynamically address the changing requirements of contemporary cities. Use cases like traffic optimization and real-time environmental monitoring demonstrate how this architecture may augment operational efficiency, better resource management, and improve people' quality of life [17] [18].

As highlighted by [15], Romania's public administration is undergoing a digital transformation aimed at aligning with European Union standards. This process involves overcoming significant challenges, such as interoperability between systems and integration of innovative technologies, while capitalizing on opportunities to enhance service delivery.

The digitalization efforts in Romania's public administration highlight the importance of adopting interoperable systems and integrating advanced architectures such as Cloud- Edge to meet both national and European digital standards [19].

This work highlights the modularity of NVIDIA microservices and Arrowhead's orchestration capabilities, which allow cities to adapt to changing needs and seamlessly integrate new technologies. It also demonstrates the effective deployment of NVIDIA AI models on edge devices and cloud infrastructure to optimize resource allocation and enhance system responsiveness.

Finally, it applies the architecture to urban challenges like traffic management, environmental monitoring, and energy optimization, showcasing its practical value in improving urban living standards. The suggested paradigm has wide-ranging effects on researchers, practitioners, and policymakers by facilitating real-time decision-making, operational efficiency, and enhanced resource management.

Even while the suggested architecture takes care of important components of smart city operation, there are still areas that might be improved. Optimizing service orchestration's latency and throughput, especially in high-demand situations with massive data flows, could be the subject of future research. Future work can further enhance the integration of AI and service-oriented frameworks by expanding on the foundation laid by this study, for smart city ecosystems that are more responsive, intelligent, and sustainable.

Acknowledgements
This work was supported by a grant of the European Commission, CHIPS Joint Undertaking (G.A. no. 101111977) and of the Ministry of Research, Innovation, and Digitization, CNCS/CCCDI - UEFISCDI, project number PN-IV-P8-8.1-PME-2024-0011

„*Arrowhead flexible Production Value Network* (Arrowhead fPVN)", within PNCDI IV, and partially supported by grant of the Ministry of Research, Innovation, and Digitization, CNCS/CCCDI-UEFISCDI, project number PN-IV-P8-8.1-PRE-HE-ORG-2023-0063 „Arrowhead fPVN PI", within PNCDI IV.

References

[1] G. Hollósi, D. Ficzere, A. Frankó and et al., "AIMS5.0 AI Toolbox: Enabling Efficient Knowledge Sharing for Industrial AI," *NOMS 2024-2024 IEEE Network Operations and Management Symposium,* pp. 1-6, 2024.

[2] D. Hästbacka and et al., "Dynamic Edge and Cloud Service Integration for Industrial IoT and Production Monitoring Applications of Industrial Cyber-Physical Systems," *IEEE Transactions on Industrial Informatics,* vol. 18, no. 1, pp. 498-508, 2022.

[3] S. Maksuti and et al., "Automated and Secure Onboarding for System of Systems," *IEEE Access,* vol. 9, pp. 111095-111113, 2021.

[4] F. Montori, M. S. Tatara and P. Varga, "Dynamic Execution of Engineering Processes in Cyber-Physical Systems of Systems Toolchains," *IEEE Transactions on Automation Science and Engineering,* pp. 1-12, 2024.

[5] H. Pettinen and D. Hästbacka, "Service Orchestration for Object Detection on Edge and Cloud in Dependable Industrial Vehicles," *Journal of Mobile Multimedia,* vol. 18(1), pp. 1-26, 2021.

[6] Eclipse Foundation, "Eclipse Arrowhead," [Online]. Available: https://projects.eclipse.org/projects/iot.arrowhead. [Accessed November 2024].

[7] Arrowhead, "Vision, objective and strategy," [Online]. Available: https://arrowhead.eu/vision-objective-and-strategy/. [Accessed November 2024].

[8] Eclipse Foundation, "Eclipse Arrowhead: A Framework for IoT and System of Systems Solutions," 2020. [Online]. Available: https://www.eclipse.org/community/eclipse_newsletter/2020/july/1.php. [Accessed November 2024].

[9] D. Ficzere, G. Hollósi, A. Frankó and P. Varga, "AI Toolbox Concept for the Arrowhead Framework," *19th International Conference on Network and Service Management (CNSM),* pp. 1-7, 2023.

[10] C. Chaubal, "Orchestrating Accelerated Virtual Machines with Kubernetes Using NVIDIA GPU Operator. NVIDIA Developer," [Online]. Available: https://developer.nvidia.com/blog/orchestrating-accelerated-virtual-machines-with-kubernetes-using-nvidia-gpu-operator/. [Accessed 31 October 2022].

[11] NVIDIA, "NVIDIA Triton Inference Server. Architecture," [Online]. Available: https://docs.nvidia.com/deeplearning/triton-inference-server/archives/triton_inference_server_1150/user-guide/docs/architecture.html. [Accessed November 2024].

[12] C. Yang, Y. Wang, S. Lan and et al., "Cloud-edge-device collaboration mechanisms of deep learning models for smart robots in mass personalization," *Robotics and Computer-Integrated Manufacturing,* vol. 77, 2022.

[13] A. Docca and M. Viviani, "Developing Smart City Traffic Management Systems with OpenUSD and Synthetic Data. NVIDIA Developer," 1 August 2023. [Online]. Available: https://developer.nvidia.com/blog/developing-smart-city-traffic-management-systems-with-openusd-and-synthetic-data/. [Accessed November 2024].

[14] J. Esposito, "Enabling Next-Level Intelligence for the Cities of Tomorrow. Lenovo SSG ISG International Services," 5 November 2024. [Online]. Available: https://news.lenovo.com/lenovo-expands-generative-ai-solutions-for-smart-cities/.

[15] G. Suciu and C. Stalidi, "Digital Twins, the Software Solution for Safer Cities," *Scientific Bulletin of Communication and Networking Systems,* vol. 1, no. 1, 2023.

[16] I. Salian, "Earth to AI: Three Startups Using Deep Learning for Environmental Monitoring," 22 April 2020. [Online]. Available: https://blogs.nvidia.com/blog/geospatial-ai-earth-day/. [Accessed November 2024].

[17] C. Vrabie, "Artificial Intelligence Promises to Public Organizations and Smart Cities," *Digital Transformation. Lecture Notes in Business Information Processing,* vol. 465, 2022.

[18] C. Buttice, "Top 14 AI Use Cases: Artificial Intelligence in Smart Cities," 11 August 2022. [Online]. Available: https://www.techopedia.com/top-14-ai-use-cases-artificial-intelligence-in-smart-cities/2/34049. [Accessed November 2024].

[19] R. Damaschin and M. G. Mihaila, "Digitalizarea administratiei publice din Romania in raport cu tendintele europene," *Smart Cities International Conference (SCIC) Proceedings* , vol. 8, p. 47–64, 2023.