SOC-E

WHITEPAPER

Time-Sensitive Networking to Meet Hard Real-Time Boundaries in Edge AI Applications for NGVA Land Vehicles

Astarloa, Armando. (SOC-E)

Abstract

This paper presents an AI-based video analytics application deployed on an edge-computing device for real-time object detection. The primary objective is to demonstrate the acceleration of AI inference and video compression using hardware accelerators embedded on RelyUm Time-Sensitive Networking (TSN) Endpoint Switch XMC Mezzanines. Detected object information, including location and size, is transmitted as hard real-time traffic over TSN, ensuring timely and reliable delivery. The implementation aligns with the NATO Generic Vehicle Architecture (NGVA) for land systems. The paper outlines the hardware and software architecture, describes the design methodology, and discusses the implementation results.

1. Introduction

Edge Intelligence (EI), the convergence of edge computing and Artificial Intelligence (AI), is gaining momentum as a key enabler for low-latency, high-performance applications. One of the most compelling cases for EI is real-time video analytics [1].

Mission-critical domains such as Aerospace & Defense (A&D) impose stringent latency and reliability requirements that traditional cloud-based solutions struggle to meet—particularly with the growing resolution of video sources and the increasing demand for bandwidth and deterministic communication [2]

To overcome these limitations, edge computing platforms equipped with dedicated hardware accelerators for video encoding/decoding and AI inference have emerged as a practical solution. In these systems, inference is executed directly at the edge to minimize latency, while model training typically remains in the cloud, leveraging its extensive computational resources

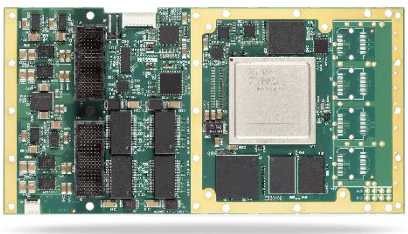

Figure 1: RelyUm AI-enabled XMC-TSN board [3].

This paper introduces a real-time AI video analytics application focused on detecting road signaling cones. The system is deployed on a RelyUm edge-computing device, shown in that leverages hardware acceleration to execute AI inference and video compression efficiently. Detected object data—including location and size—is transmitted as hard real-time traffic over Deterministic Ethernet (Time-Sensitive Networking - TSN). The use case is implemented within the framework of the NATO Generic Vehicle Architecture (NGVA) for land systems.

1.1 Edge Intelligence

Deep Neural Networks (DNNs) have emerged as one of the most powerful tools in artificial intelligence, offering solutions to a wide range of modern technological challenges. Their applications span numerous domains that impact everyday life—from autonomous driving systems to advanced healthcare diagnostics.

DNNs excel in learning hierarchical representations from raw data, automatically extracting high-level features layer by layer without the need for manual intervention. In many cases, their performance surpasses that of human experts in tasks such as image recognition, natural language processing, and pattern detection. This capability has led to DNNs becoming a standard in state-of-the-art AI technologies, demonstrating superior efficiency and accuracy across a wide array of systems and implementations [4].

Traditionally, the training and inference of DNNs have relied on cloud computing infrastructures, taking advantage of their massive computational resources. While this approach has enabled significant progress, it faces limitations in real-time applications due to the increasing volume of data generated at the network edge. The sheer scale of Big Data introduces latency, bandwidth constraints, and potential privacy concerns—creating bottlenecks that hinder the effectiveness of cloud-based DNN processing.

To address these challenges, Edge Intelligence (EI) has emerged as a transformative paradigm. EI combines Edge Computing (EC) and AI, enabling data processing to occur closer to the data source—at the edge of the network. By offloading computing tasks from the cloud to edge devices, EI reduces latency, increases responsiveness, and supports real-time decision-making.

The integration of DNNs into edge devices enables distributed and collaborative inference strategies, where models can be partially or fully executed on local hardware. This architecture is particularly valuable in time-critical environments such as autonomous systems, industrial automation, and smart surveillance, where immediate insights from data are essential.

1.2 Convolutional Neural Network for Real-Time Video Analytics

DNNs have become foundational tools in modern artificial intelligence, offering transformative capabilities across sectors where precision, autonomy, and situational awareness are paramount. Within this broad category, Convolutional Neural Networks (CNNs) have emerged as a key enabler of real-time visual analytics, particularly in mission-critical applications such as autonomous navigation, target recognition, and perimeter surveillance—core to A&D operations [5].

In such high-assurance environments, systems must operate under stringent constraints: low latency, high reliability, and deterministic performance, often in the absence of cloud connectivity. To meet these requirements, CNNs are increasingly deployed on edge-computing platforms, where sensor data is processed locally on ruggedized, resource-constrained devices. These edge systems are capable of executing AI inference in real time, enabling fast decision-making without relying on external communication links—an essential feature for tactical platforms and autonomous systems operating in contested or disconnected environments.

Object detection at the edge relies on CNNs to parse images into regions, extract features, and identify object locations and classes in real time. Several network architectures have been developed for this task: R-CNN, which generates region proposals before classification; SSD (Single Shot Detector) , which integrates object localization and classification in a single pass; RetinaNet, which uses focal loss to improve detection in imbalanced datasets; YOLO (You Only Look Once), which provides ultra-fast inference by framing detection as a single regression problem across the entire image.

This use-case adopts YOLO due to its high inference speed and suitability for real-time applications. These characteristics make it well-matched for A&D scenarios such as real-time battlefield awareness, vehicle hazard detection, and UAV-based surveillance, where quick and accurate responses are critical [6].

YOLO typically leverages Darknet as its convolutional backbone. The network divides an image into a grid, with each cell predicting multiple bounding boxes and associated class probabilities. Each detection includes: Bounding box coordinates (x, y, width, height); a confidence score indicating object presence and, the predicted object class.

Low-confidence predictions are filtered, resulting in a compact, actionable set of detections suitable for downstream processing.

To ensure the robust deployment of such systems in the A&D domain, this solution is implemented on a deterministic, edge-ready platform leveraging RelyUm technology. As an example, RELY-TSN12 (12-ports TSN bridge) and RELY-TSN-PCIe (TSN PCIe NIC), provide Time-Sensitive Networking (TSN) and Deterministic Ethernet support, ensuring reliable and bounded-latency communication. These features are essential for integrating real-time AI analytics into NGVA-compliant land systems, avionics payloads, or ISR platforms, where timing precision and interoperability with legacy systems are non-negotiable.

1.3 Time-Sensitive Networking in the context of NATO Generic Architecture for Land Systems

Since its standardization in 1983, Ethernet has evolved far beyond its original scope as a local networking solution for computer systems. Today, Ethernet has become the de facto standard for fieldbus communication across a wide range of industries, including industrial automation, energy, automotive, and increasingly, the A&D sector.

However, the traditional Ethernet lacked native support for real-time communication, a critical requirement in military and safety-critical systems. While various proprietary solutions were developed to address these limitations, their lack of interoperability created vendor lock-in and increased integration complexity.

To overcome these constraints, the IEEE TSN Task Group introduced a suite of open standards that extend Ethernet with deterministic communication capabilities [7]. Originally stemming from the Audio Video Bridging (AVB) initiative, TSN ensures predictable latency, low jitter, and guaranteed delivery times over standard Ethernet infrastructure.

At the core of TSN is the Time-Aware Shaper which structures network traffic into repetitive cycles. Each cycle is divided into time slots assigned to different traffic classes, enabling coexistence of high-priority scheduled traffic with best-effort and reserved streams. This mechanism ensures that critical data—such as control messages or sensor fusion results—are transmitted with guaranteed timing, even under high network load.

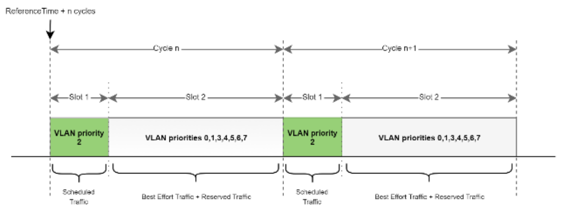

Figure 2: Example of a basic, two slot, TSN configuration.

Figure 2 illustrates a simplified example of TSN scheduling, where scheduled traffic is isolated from lower-priority communication, preserving quality of service (QoS) across the network. In typical deployments, up to eight distinct traffic classes can be scheduled within each communication cycle, offering flexible control over latency and bandwidth allocation.

The NGVA defines standardized and modular architecture for land-based military platforms, as specified in the AEP-4754 standard [8]. NGVA promotes interoperability, modularity, and reuse across NATO member systems, reducing integration complexity and lifecycle costs.

One of the most challenging aspects of NGVA-compliant system design is ensuring support for real-time and deterministic communication between heterogeneous subsystems—ranging from fire control systems to sensor fusion units and mission management components. This is precisely where TSN becomes essential.

NGVA mandates the use of Ethernet at the Data Link Layer for all inter-subsystem communication. As shown in Figure 3, Volume III of AEP-4754 defines the NGVA Data Infrastructure, which encompasses multiple layers: User Application, Data Model, Transport, Network, and Data Link/Physical. While the application layer remains system-specific, all other layers are standardized to support consistent communication semantics, bandwidth guarantees, and timing precision across all compliant nodes.

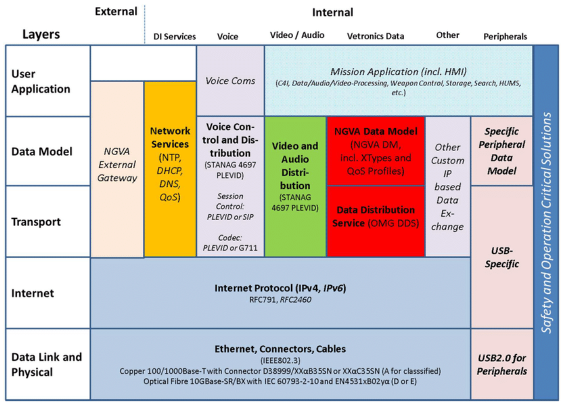

Figure 3: Definition of NGVA data infrastructure [8].

By integrating TSN into NGVA-based systems, military platforms benefit from: Deterministic data delivery for control and mission-critical functions; reduced latency for real-time video; navigation and targeting systems; interoperability among multi-vendor equipment; scalability and standardization for future upgrades.

2. SoC Architecture for AI Video Processing Acceleration

2.1 High Level Architecture

The integration of AI with FPGA-based systems represents a powerful solution for meeting the stringent performance and reliability requirements of real-time applications in A&D. Particularly in scenarios involving unmanned platforms, autonomous vehicles, or surveillance systems, this combination offers low-latency inference, deterministic communication, and optimal power efficiency at the tactical edge.

In this context, this work presents a hardware-accelerated AI system for drone-sized object detection, implemented on the RelyUm TSN Endpoint Switch XMC Mezzanine (XMC-10TSN series). This module integrates the AMD-Xilinx Zynq® UltraScale+™ MPSoC, which provides a robust platform for high-performance Edge Intelligence. The system is specifically designed to address real-time processing and low-latency communication needs in embedded, mission-critical environments.

To ensure rapid object identification and prompt communication to other subsystems, the architecture relies on two principal components. First, the Zynq UltraScale+ device incorporates a Deep Learning Processor Unit (DPU) within its programmable logic, which performs the inference of convolutional neural networks (CNNs) trained for this task. Second, a Time-Sensitive Networking (TSN) switch IP block (SocTek) embedded in the same FPGA fabric handles the transmission of crucial information—such as object size and location—over a deterministic Ethernet link. This configuration aligns with the IEEE 802.1 TSN standard and is key for enabling synchronized and predictable communication across the vehicle network.

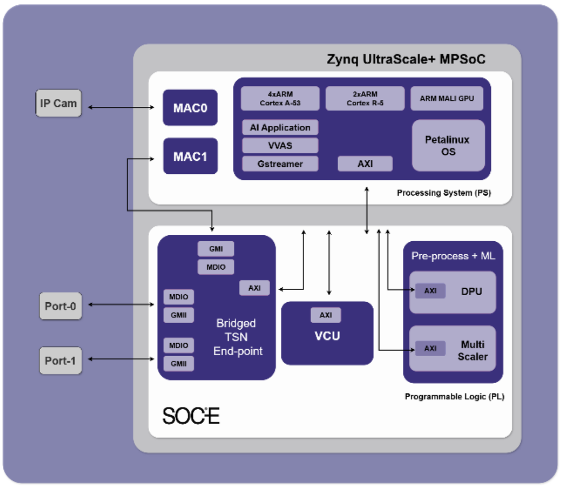

Figure 4: High-Level Architecture of the AI-TSN acceleration system implemented on the MPSoC device of RelyUm XMC board.

As shown in Figure 4, the hardware architecture is composed of a close interaction between the Programmable Logic (PL) and the Processing System (PS) sections of the Zynq device. On the software side, a Linux-based operating system runs on the PS, supported by libraries and tools such as GStreamer for multimedia handling, the Xilinx Vitis Video Analytics SDK (VVAS), DPU runtime components, and custom application software. These elements manage the capture, processing, and interpretation of video streams from onboard sensors.

The PL portion handles the critical acceleration tasks. The DPU is tailored for convolutional network inference and is configurable according to application needs, allowing for different levels of parallelism and performance, depending on available logic resources and architectural parameters. The TSN switch implemented in programmable logic ensures that data flows through the network with guaranteed bandwidth and latency, a requirement for applications where timing precision is crucial.

Additional processing is provided by a Multi-Scaler IP, which prepares video input by adjusting resolution, scaling, and color parameters to match the CNN's expected format. Video encoding and decoding are handled by the Video Codec Unit (VCU), which implements H.264/H.265 compression standards in hardware. While codec acceleration resides in the PL, video interfaces such as HDMI, DisplayPort, and MIPI-CSI are handled by the PS domain, ensuring complete system integration and performance optimization.

This integrated platform, leveraging RelyUm's TSN switching capabilities and the compute density of the Zynq MPSoC, illustrates a practical and efficient approach for deploying AI-based situational awareness in Aerospace and Defence platforms. It provides a foundation for high-speed, deterministic communications and robust AI inference at the tactical edge—ideal for real-time detection, tracking, and decision-making in critical missions.

2.2 Desing and Data Flow

Proper training of the neural network is critical for achieving optimal performance in the deployment of AI-based real-time detection systems. In this implementation, the neural network is tailored for the detection of road cones—an application relevant to autonomous navigation and unmanned ground vehicle perception within the A&D domain.

The development process involves several key phases: selecting and preparing the dataset, adjusting the YOLO network parameters, training and validating the model, quantizing the network for deployment, and generating an implementable model compatible with the system's DPU IP. These steps leverage the Darknet AI framework, which provides both the YOLO architecture and the GPU-accelerated compiler for efficient model training. Quantization and deployment are performed using the Xilinx Vitis AI toolkit, which transforms the trained model into a version suitable for execution at the edge.

The training dataset includes a variety of image types to ensure model robustness across different operational scenarios. These images vary in resolution, color fidelity, lighting conditions, object positioning, and cone density. Prior to training, annotation files are generated containing the class labels and the positional vectors of objects in each image. The YOLO network is trained using the Darknet compiler on a GPU to accelerate convergence. The resulting model achieves a classification accuracy of 96%.

To enable efficient inference on edge devices, model quantization is essential. This process reduces the computational load by converting the 32-bit floating-point weights and activations into 8-bit integer values. Despite this reduction, quantization must preserve the model's accuracy to maintain detection reliability in mission-critical contexts. Vitis AI, in conjunction with the TensorFlow framework, is used to perform this quantization and to compile the model for the target DPU architecture. The resulting file includes all necessary instructions and runtime information for real-time inference execution.

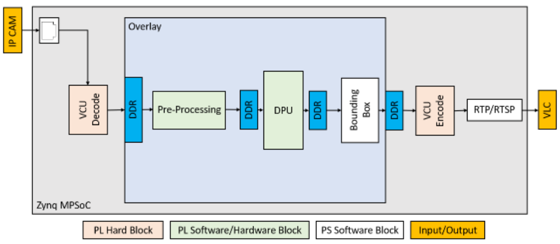

Figure 5: High level overview of the dataflow in the AI-TSN object detection application.

Figure 5 summarizes the data flow- The complete system utilizes the GStreamer framework for managing video flows over RTSP, coordinating image capture, preprocessing, AI inference, bounding box generation, and data transmission. Video frames enter and exit the system encoded in H.264, with the VCU decoding them into raw data. This raw video is then preprocessed by the Multi-Scaler IP, which converts the NV12 image format to BGR and resizes the frame to the 416×416 resolution expected by the YOLO network.

Once the video data is prepared, the DPU IP block performs accelerated inference, identifying objects and extracting vector-based information such as location, dimensions, and class confidence. This information is used to generate bounding boxes, which visually represent detected objects. The bounding box data, along with processed video, is transmitted via Time-Sensitive Networking (TSN) to other subsystems within the vehicle or platform.

To ensure interoperability and real-time performance, the system classifies outgoing traffic according to the NGVA standard and maps it onto TSN traffic classes:

- Hard real-time traffic: This includes the location and size of bounding boxes, which are critical for control and response functions. It is transmitted as Scheduled Traffic, with dedicated time slots in the TSN cycle, using a customized NGVA Brake Model message format.
- Soft real-time traffic: The annotated video stream, containing bounding boxes overlaid on the original feed, is sent to display systems as Reserved Traffic. While latency is minimized, this traffic does not require the same strict timing guarantees as scheduled data.
- Best-effort traffic: All remaining non-critical communications are transmitted using the Best-Effort TSN class and share a slot with the video stream, accommodating background and non-time-sensitive data.

By combining hardware-accelerated AI processing, deterministic TSN-based communication, and adherence to NGVA standards, the proposed system delivers a robust, real-time perception capability for Aerospace and Defence applications. It ensures accurate and low-latency object detection with guaranteed Quality of Service (QoS), suitable for integration into autonomous or crew-assisted mission systems.

2.3 Desing and Data Flow

For evaluation purposes, the system was tested using HD video streams at a resolution of 1280×720 pixels and a frame rate of 25 FPS. Several DPU architectures were explored during implementation, specifically the B4096, B3136, B2304, and B1600 configurations. Due to resource constraints on the hardware platform, the two most computationally demanding architectures (B4096 and B3136) were implemented with a single DPU core, whereas the B2304 and B1600 configurations allowed dual-core implementations, enabling greater parallelism.

Among all the evaluated configurations, the B4096 DPU architecture delivered the best latency performance and was, therefore, selected for the final system implementation. This version of the system executes the complete neural network inference pipeline with minimal processing delay, ensuring fast object detection suitable for real-time Aerospace and Defense applications, such as autonomous navigation and threat avoidance. The total latency for the capture, inference, and communication processes is approximately 80 ms. The DPU data controller operates at 250 MHz, and the DSP segments within the compute unit module are clocked at 500 MHz. The VCU also runs within the 250 MHz domain of the DPU controller. The bootneck, in this case, has been identified in the video source used in the test set-up.

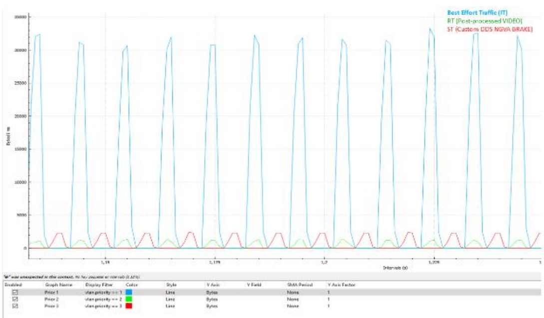

Figure 6: Generated TSN traffic combining Real-Time and Best-effort traffic.

Figure 6 shows the traffic distribution associated with the Time-Sensitive Networking (TSN) configuration used in the test setup. One dedicated slot is reserved for the transmission of control data using a customized NGVA Brake model frame format, while the post-processed video stream is transmitted within a shared slot alongside best-effort traffic, maintaining deterministic communication behavior.

3. Conclusions

This work demonstrates the feasibility and efficiency of deploying Edge AI vision systems for real-time object detection in NGVA-compliant land vehicle platforms. By integrating hardware-assisted neural network inference using AMD-Xilinx Zynq UltraScale+ MPSoC devices with time-sensitive networking (TSN) communication, the proposed solution achieves low-latency performance that meets the stringent real-time requirements of defense-grade vehicular systems.

The system leverages the powerful combination of a DPU and deterministic Ethernet enabled through a RelyUm TSN Endpoint Switch XMC Mezzanine (XMC-10TSN series), ensuring precise detection, fast inference, and predictable transmission of critical situational awareness data. The use of the SOC-E TSN technology enables seamless interoperability and QoS-compliant data exchange among NGVA sub-systems such as perception, actuation, and control units.

By aligning with the NGVA framework and leveraging scalable FPGA-based edge computing, the implementation not only enhances the autonomy and responsiveness of ground vehicles but also ensures compliance with modular, future-proof system integration standards. The modularity of the solution, its efficient use of TSN for real-time data delivery, and its capability to run complex AI models at the tactical edge confirm its suitability for next-generation defense mobility applications.

Future developments may include the extension of this framework to multi-sensor fusion scenarios, secure edge inference, and real-time system monitoring—all supported by RelyUm's portfolio of TSN-capable hardware and software solutions tailored for Aerospace & Defense systems.

Referencias

[1] O. Kechagias-Stamatis and N. Aouf, "Automatic Target Recognition on Synthetic Aperture Radar Imagery: A Survey," in IEEE Aerospace and Electronic Systems Magazine, vol. 36, no. 3, pp. 56-81, 1 March 2021.

[2] Z. Wan, "Cloud Computing infrastructure for latency sensitive applications," 2010 IEEE 12th International Conference on Communication Technology, Nanjing, China, 2010, pp. 1399-1402.

[3] SoC-e S.L, “RelyUm AI-enabled XMC-TSN board”, https://soc-e.com/products/xmc-tsn-xmc-10tsn/, 2025.

[4] J. Li, W. Liang, Y. Li, Z. Xu, X. Jia and S. Guo, "Throughput Maximization of Delay-Aware DNN Inference in Edge Computing by Exploring DNN Model Partitioning and Inference Parallelism," in IEEE Transactions on Mobile Computing, vol. 22, no. 5, pp. 3017-3030, 1 May 2023.

[5] Jiuxiang Gu etal., Recent advances in convolutional neural networks, Pattern Recognition, Volume 77, 2018, Pages 354-37.

[6] C. -J. Lin and J. -Y. Jhang, "Intelligent Traffic-Monitoring System Based on YOLO and Convolutional Fuzzy Neural Networks," in *IEEE Access*, vol. 10, pp. 14120-14133, 2022

[7] IEEE Time Sensitive Networking Task Group, 2018 IEEE 802.1 Standards. http://www.ieee802.org/1/pages/tsn.html.

[8] NATO, “AEP-4754 NATO GENERIC VEHICLE ARCHITECTURE (NGVA) for Land Systems Volume Iii: Data Infrastructure”, https://www.natogva.org, 2018.