CLOSED-WORLD EDGE DIMENSIONING: ENGSET-BASED RESOURCE OPTIMIZATION FOR SPATIAL MICRO-ZONES IN INDUSTRY 4.0

Damian Kmiecik [1,2], *Adrian Dziembowski* [2]

[1] "PerfectSoft" Damian Kmiecik, Poznań, Poland
[2] Institute of Multimedia Telecommunications, Poznan University of Technology, Poznań, Poland

ABSTRACT

Contemporary Edge AI systems in Industry 4.0 suffer from severe hardware over-provisioning due to heuristic peak capacity scaling or flawed infinite-source Poisson queuing assumptions. This paper introduces a closed-world dimensioning paradigm for Multi-Tier Edge architectures using finite-source teletraffic theory. By treating heavy video analytics as overflow traffic, we model the system via a multi-rate Engset distribution and a modified Fredericks-Hayward approach to capture exact variance distortion. We introduce a Time-Fractioning Algorithm to mathematically resolve multi-agent frame-sharing. Furthermore, using a binomial Temporal Ensembling model, we prove that a controlled frame loss target not greater than 1% preserves effective system recall. Validated against the synthetic NVIDIA PhysicalAI-SmartSpaces dataset, the proposed framework achieves a direct hardware resource reduction of 7.8% compared to computer vision peak provisioning, and effectively eliminates the 11.8% over-provisioning flaw of classical Erlang B models, all while strictly maintaining safety-critical Service Level Agreements.

Index Terms— Edge AI, Computer Vision, Teletraffic Theory, Engset Model, Resource Dimensioning, Industry 4.0.

1. INTRODUCTION

Edge Artificial Intelligence (Edge AI, [1]) enables real-time visual analytics for Industry 4.0 [2] applications, e.g., Automated Optical Inspection (AOI, [3]) or collision prediction [4]. Despite the maturity of Deep Neural Network (DNN) architectures, the methodologies for dimensioning the underlying hardware infrastructure remain predominantly heuristic.

Conventional deployments rely on *Peak Provisioning*, allocating a dedicated 1:1 hardware ratio for each video stream. However, advanced architectures like Vision Transformers [5] performing dense 4K analytics require $150-200$ ms per forward pass [6]. This imposes a strict physical limit of only $5-10$ frames per second (FPS) on a single industrial-grade accelerator. Consequently, scaling this heuristic approach to multi-camera environments requires massive resource allocation that often exceeds Capital Expenditure (CAPEX, [7]) budgets and rigid Size, Weight, and Power constraints.

Conversely, theoretical models in Multi-access Edge Computing (MEC, [8]) predominantly assume infinite-source Poisson arrival processes (e.g., $M/M/1$ or $M/M/c$ queues) [9, 10, 11]. In closed industrial facilities, the mobile agent population (e.g., personnel, robots) is strictly finite. This closed-world constraint enforces a negative feedback loop: an agent occupying resources in one zone cannot simultaneously trigger inference in another. Ignoring this property forces conventional models to assume a peakedness factor of $Z = 1$, failing to account for variance truncation and resulting in hardware over-provisioning.

To address these limitations, we merge computer vision engineering with finite-source multi-rate teletraffic theory. The main contributions of this paper are:

- Formulation of a Multi-Tier Edge AI pipeline as a closed overflow network.
- The Time-Fractioning Algorithm, converting physical multi-agent frame sharing into statistically independent variables for recursive Engset state equations.
- A multi-service dimensioning methodology (BPP-BPM overflow model [12, 13]) extending Fredericks-Hayward method [14] to extract exact overflow variance distortion ($Z_0 < 1$).
- A proof that intentional teletraffic frame loss ($P_{\text{loss}} \leq 0.01$) is safely mitigated by Temporal Ensembling, reducing hardware without degrading effective system recall.

2. SYSTEM ARCHITECTURE

To establish a mathematically rigorous dimensioning framework, we model the system as a cascaded, multi-tier overflow network embedded within a closed physical environment. To avoid precision loss, we treat the video frame as the atomic unit of time. Both arrival and service intensities are calculated strictly based on frame indices.

2.1. Multi-Tier Edge AI Pipeline

The proposed computational architecture dynamically allocates resources based on spatial triggers across distinct tiers. Resource consumption is measured in **Allocation Units (AU)**, defined as a standardized, discrete fraction of the hardware accelerator's compute and memory capacity.

1. **Background Detection Tier (0 AU cost):** Continuous spatial monitoring that does not consume resources at the central server. Utilizing lightweight Video Motion Detection (VMD), this tier serves as a zero-cost spatial trigger.
2. **Tier-1 Edge Nodes (Primary Resources, V^s):** Localized edge accelerators (e.g., Jetson modules) with a finite capacity V^s. When localized spatial demand exceeds V^s, excess traffic is instantantly offloaded to the central tier.
3. **Tier-2 Central Edge Server (Secondary Resources, V^0):** A centralized GPU/TPU cluster executing heavy analytical requests (e.g., 4K AOI, 3D point clouds). Processing a frame at this tier requires a deterministic allocation cost of d_c AU.

2.2. Markovian Single-Source Parameters

In a closed-world industrial environment, the number of mobile agents (e.g., personnel, automated guided vehicles) is finite and bounded by a maximum population N. For each unique agent $i \in \{1, 2, \ldots, N\}$, the spatial behavior is decomposed to define a two-state Markov chain (Idle/Busy):

$$\mu_i = \frac{1}{\bar{t}_{\text{busy},i}}, \quad \gamma_i = \frac{1}{\bar{t}_{\text{idle},i}}, \quad \alpha_i = \frac{\gamma_i}{\mu_i}, \tag{1}$$

where $\bar{t}_{\text{busy},i}$ and $\bar{t}_{\text{idle},i}$ represent the mean duration (in frames) the object i spends inside and outside the heavy analytics zones, respectively. Crucially, the closed-world assumption enforces a negative feedback loop: an agent occupying resources in the *Busy* state is physically excluded from generating new requests, naturally truncating the traffic variance tail.

2.3. Spatial Modeling and Time-Fractioning

In industrial vision networks, multiple moving agents frequently share a single video frame within micro-zones. Because the number of Multiply-Accumulate (MAC) operations executed during the forward pass on a dedicated AI accelerator is deterministic for a fixed input tensor, the GPU hardware cost remains strictly constant regardless of the object count. Processing time variations occur solely during CPU-bound post-processing (e.g., Non-Maximum Suppression), which does not consume GPU Allocation Units.

However, physical frame-sharing violates the assumption of statistical independence required by the Engset model. To formulate valid inputs for recursive state equations, we introduce the *Time-Fractioning Algorithm*. Let $N_{f,k}$ denote the number of agents present in zone k during frame f. The fractional load generated by agent i in zone k is:

$$c_{i,f,k} = \begin{cases} \frac{1}{N_{f,k}} & \text{if agent } i \text{ is in zone } k, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Since an agent cannot occupy disjoint zones simultaneously, its total fractional cost for the central server during frame f is $c_{i,f} = \sum_k c_{i,f,k}$. Corrected, effective offered traffic $\alpha_{\text{eff},i}$ is:

$$t_{\text{eff\_busy},i} = \sum_f c_{i,f} \quad \Longrightarrow \quad \alpha_{\text{eff},i} = \frac{t_{\text{eff\_busy},i}}{\bar{t}_{\text{idle},i}}. \tag{3}$$

This transformation converts physical image sharing into statistically disjoint time allocations, ensuring that $\alpha_{\text{eff},i} < \alpha_i$. Although group movements introduce temporal correlations, the finite population (N) and closed-world negative feedback loop (agents clustered under one camera cannot trigger requests elsewhere) mathematically stabilize load fluctuations, preserving model validity.

To prevent redundant processing in overlapping fields of view, a CPU-bound *Zone-Centric Master Camera Handover* selects a single master RTSP stream per agent cluster based on zero-cost metadata generated by Tier-1 nodes (e.g., Track-IDs, coordinates). Selection prioritizes maximum cumulative bounding box area ($\sum \text{RoI}_{\text{Area}}$) for optimal resolution, using mean confidence scores as tie-breakers. This coordination secures central GPU resources for neural network inference.

3. TRAFFIC CHARACTERIZATION

Real-world industrial environments exhibit high kinetic heterogeneity. Mechanized traffic (e.g., automated guided vehicles) typically have deterministic trajectories, whereas pedestrian traffic is highly stochastic, featuring unpredictable pathing and high variance in appearance times. To rigorously evaluate the proposed model under conditions of high uncertainty, the input spatial data is explicitly filtered to exclude deterministic machine movement, focusing the analysis entirely on the behavior of human agents. Because smart factories are strictly governed by takt times and Standard Operating Procedures (SOPs), human workflows are wide-sense stationary. Therefore, steady-state profiling is sufficient for offline capacity planning, deliberately avoiding the massive GPU OPEX required for continuous adaptive profiling.

Applying a global average to these highly variable traffic parameters suppresses natural system variance, leading to inaccurate dimensioning. Thus, we implement a 3-step extraction and clustering pipeline to feed a multi-rate Engset model.

3.1. Data Purification and Outlier Isolation

The input spatial data is filtered to remove elements that violate stochastic principles:

1. **Stationary Objects:** Personnel at fixed workstations are excluded, as they deterministically saturate resources (occupying hardware for 100% of the video frames).
2. **Background Noise:** Agents appearing only incidentally are isolated into a minimal-impact Engset class (e.g., $\alpha \approx 0.001$), preserving the N without skewing the service intensity estimators of the primary workforce.

3.2. Autonomous Behavioral Profiling (K-Means)

For the filtered subset of N_{active} mobile agents, we apply K-Means clustering in the normalized two-dimensional feature space $[\bar{\gamma}_i, \bar{\mu}_i]$. The optimal number of clusters B is determined autonomously by maximizing the Silhouette Score, ensuring maximal intra-cluster cohesion and inter-cluster sepa-

ration (supported by the Elbow Method on the Within-Cluster Sum of Squares – WCSS). For each extracted cluster c, the algorithm computes centroid coordinates $\bar{\gamma}_c$ and $\bar{\mu}_c$, defining the baseline offered traffic α_c for distinct operational profiles. The exact empirical clustering results, including the optimal number of clusters and the achieved Silhouette Score, are detailed in Section 7.

3.3. Traffic Class Formulation and Spatial Multiplication

Intelligent Edge Orchestration links hardware demand to the specific neural network architecture assigned to a spatial zone, defining the computational cost d_c in AU (e.g., lightweight object detection requiring single-stage networks like YOLO-Nano [15] and $d = 2 - 5$ AU versus heavy 3D point cloud processing requiring massive vision transformers and $d = 30 - 90$ AU [16]).

In classical multi-rate teletraffic theory, a class is defined by a unique pair of offered traffic (α) and requested capacity (d). Since clustered agents migrate across D distinct physical zones with varying d_c demands, we introduce a *Spatial Multiplication Algorithm* yielding $M = B \times D$ input classes. Agents are treated as multi-service sources, with their populations (N_c) fractionally assigned based on their spatial migration probability density. The vector for each class $c \in \{1, 2, \ldots, M\}$ assumes a unified format:

$$\text{Class}_c = \{N_c, \gamma_c, \mu_c, \alpha_c, d_c\}. \tag{4}$$

This rigorous parameterization allows the underlying engine to execute the multi-rate Kaufman-Roberts recursion over all permissible system states ($x \geq d_c$). By isolating distinct service requirements, the model accurately reflects the dynamics of multiplexing asynchronous video streams with various hardware demands on a shared GPU.

4. MULTI-TIER OVERFLOW TRAFFIC ANALYTICS (BPP-BPM OVERFLOW MODEL)

In the defined multi-tier architecture, Tier-1 edge nodes traditionally process primary stochastic traffic, while the residual stream overflowing to the Tier-2 central cluster (V^0) exhibits a distorted variance-to-mean ratio. Conventional MEC models rely on Poisson assumptions ($Z = 1$), which ignore the population limit, while standard Engset models applied to an overflow stream critically underestimate hardware requirements. To extract the exact aggregate peakedness factor (Z_0), we employ BPP-BPM overflow model [12, 13], an advanced extension of the Fredericks-Hayward approach for finite-source systems.

4.1. Bypassing Fictitious Resources ($V^s \to 0$)

Typically, *BPP-BPM overflow model [12, 13]* utilizes a modified Equivalent Random Traffic (ERT) method with Rapp's approximations to calculate Equivalent Fictitious Primary Resources (EFPR). However, in our proposed spatial architecture, Tier-1 edge nodes strictly execute zero-cost Video Motion Detection (VMD) triggers. This restricts their capacity for heavy DNN analytics to an asymptotic limit of $V^s \to 0$. Consequently, the complex EFPR transformation is mathematically bypassed, as 100% of the spatial traffic overflows directly to the central cluster. Setting $V^s \to 0$ allows for isolating and validating the Engset dimensioning apparatus.

4.2. Overflow Traffic Parameters

Crucially, because $V^s = 0$, the overflow traffic parameters (mean R_c and variance σ_c^2) are strictly equal to the offered traffic parameters generated by idle sources. For a finite-source Engset class, the exact parameters yield a peakedness factor strictly less than unity:

$$\begin{aligned} R_c = N_c \frac{\alpha_c}{1+\alpha_c}, \quad \sigma_c^2 = N_c \frac{\alpha_c}{(1+\alpha_c)^2} \\ \implies Z_c = \frac{1}{1+\alpha_c} < 1 \end{aligned} \tag{5}$$

When the secondary resource is modeled via the generalized Fredericks-Hayward framework, this variance distortion forces an algebraic simplification where the input arrival component $\frac{R_c}{Z_c}$ resolves directly to a state-independent term:

$$\frac{R_c}{Z_c} = \frac{N_c \frac{\alpha_c}{1+\alpha_c}}{\frac{1}{1+\alpha_c}} = N_c \alpha_c \tag{6}$$

This elegant reduction proves that the macrostate space perfectly adapts to the unique nature of the closed-world sources, fully preserving the true underlying variance structure through individual peakedness filtering.

4.3. Loss Probability

To dimension the central server, the macrostate probabilities $[P_n]$ of the secondary resource are derived using the multi-rate Kaufman-Roberts recursive equation, generalized via the Fredericks-Hayward approximation to incorporate the aggregate peakedness factor Z_0:

$$n[P_n]_{\frac{V^0}{Z_0}} = \sum_{c=1}^{m} \frac{R_c}{Z_c} d_c [P_{n-d_c}]_{\frac{V^0}{Z_0}} \tag{7}$$

where V^0 is the baseline hardware capacity of the Tier-2 central server, d_c represents the computational cost of class c expressed in Allocation Units, while R_c and Z_c correspond to the mean overflow traffic and individual peakedness factor derived for each finite-source Engset profile, respectively.

The ultimate loss probability (Call Congestion) B_c^0 for class c on the central accelerator is strictly evaluated over the active blocking boundary states. In a closed-world system, active agents currently utilizing the server are physically blocked from triggering concurrent tasks. Therefore, incoming inference requests see a state distribution shifted by the remaining idle source population. The Call Congestion is evaluated within this normalized state space as follows:

$$B_c^0 = \sum_{n=\lfloor \frac{V^0}{Z_0} \rfloor - d_c + 1}^{\lfloor \frac{V^0}{Z_0} \rfloor} [P_n]_{\frac{V^0}{Z_0}} \tag{8}$$

By forcing the objective optimization function strictly through this loss probability metric rather than standard static time congestion, the model intrinsically evaluates the operational feedback loop, eliminating the over-provisioning vulnerabilities of traditional models.

5. INVERSE DIMENSIONING AND RESILIENCE

The ultimate objective of the proposed framework is the mathematical minimization of hardware CAPEX without compromising the reliability of the computer vision pipeline.

5.1. Inverse Dimensioning and SLA Formulation

Modern multi-object tracking architectures robustly interpolate missing spatial coordinates [17], inherently tolerating marginal frame drops. We exploit this algorithmic resilience by defining a strict Service Level Agreement (SLA) threshold, $P_{\text{loss}}^{\text{target}}$ (e.g., ≤ 0.01). This transforms hardware scaling into a constrained probabilistic optimization.

The *Inverse Dimensioning Algorithm* reverses standard capacity evaluation. By iteratively solving the BPP-BPM overflow model [12, 13] state equations (incorporating the aggregate peakedness factor Z_0), it determines the minimal optimal capacity V_{opt}^0 that satisfies the SLA:

$$V_{\text{opt}}^0 = \min\{V \in \mathbb{N} \mid P_{\text{loss}}(V, Z_0) \leq P_{\text{loss}}^{\text{target}}\}. \quad (9)$$

Because the blocking probability function monotonically decreases with capacity V, the minimal boundary can be isolated deterministically by the solver.

5.2. System-Level Recall and Temporal Ensembling

A common objection to teletraffic optimization in the computer vision community is the intentional acceptance of frame loss ($P_{\text{loss}} > 0$). In safety-critical applications, a standard alternative is to operate at a globally reduced frame rate (e.g., $2-5$ FPS) to ensure zero buffer overflows. However, industrial environments exhibit severe perceptual distortions, such as artificial glare and motion blur. At low frame rates, a single distorted frame results in a complete classification failure for that time window.

To ensure robustness, Edge AI systems process high-density video streams (e.g., 30 FPS) and compute final classifications using *Temporal Ensembling*. In a sliding window of W frames, the algorithm requires a minimum subset of M successfully processed frames to output a prediction matching the nominal neural network accuracy (R_{nom}).

We model the probability of a successful system-level classification ($R_{\text{eff\_system}}$) using the binomial distribution:

$$R_{\text{eff\_system}} = R_{\text{nom}} \cdot \sum_{k=M}^{W} \binom{W}{k} (1 - P_{\text{loss}})^k (P_{\text{loss}})^{W-k}. \quad (10)$$

Given an optimized queue target of $P_{\text{loss}} = 0.01$, an ensembling window of $W = 30$, and a threshold of $M = 20$, the cumulative probability is > 0.99999.

While blocking systems can exhibit bursty or correlated loss patterns, blockages under our strict target ($P_{\text{loss}} \leq 0.01$) are highly transient. Because the source population is finite (N=50), the closed-world negative feedback loop physically prevents prolonged hardware saturation and limits the burst size of consecutive dropped frames. Since the execution time for a single inference request is a fraction of a second, this theoretical burst size remains strictly within the safety margin ($W-M$=10). Thus, Temporal Ensembling mitigates these transient drops, ensuring that the BPP-BPM overflow model [12, 13] yields significant CAPEX reduction while maintaining effective system recall ($R_{\text{eff\_system}} \approx R_{\text{nom}}$).

6. EXPERIMENTAL METHODOLOGY

6.1. Dataset

To empirically validate the proposed teletraffic framework, the mathematical model is evaluated against the synthetic *NVIDIA PhysicalAI-SmartSpaces* dataset (AI City Challenge, [18]), specifically the *Warehouse_001* scenario.

Utilizing exact Ground Truth bounding boxes establishes a nominal detection rate of $R_{\text{nom}} = 100\%$. This ensures strict methodological variable isolation, decoupling intrinsic AI perception errors (e.g., missed detections due to occlusions or glare) from hardware buffer saturation (P_{loss}). In production, hardware dimensioning must remain independent of classification accuracy to prevent perception failures (e.g., false negatives artificially lowering the load) from masking hardware capacity limits. Therefore, 100% accurate triggers allow for worst-case stress-testing under full nominal load. Furthermore, a synthetic dataset bypasses the unavailability of open-source industrial video repositories due to corporate NDAs. The dataset also guarantees immutable TrackIDs for agents migrating across overlapping camera fields of view, ensuring the total population N remains constant, mathematically satisfying the closed-world constraint.

6.2. Comparative Baselines

To rigorously evaluate the CAPEX optimization achieved by the Inverse Dimensioning Algorithm, BPP-BPM overflow model [12, 13] is benchmarked against three prevalent engineering baselines:

1. Peak provisioning (heuristic 1:1 mapping),
2. Infinite-source Erlang B (Poisson arrivals, $Z = 1$),
3. Standard Engset model (ignoring overflow variance).

7. EXPERIMENTAL RESULTS

The evaluation sequentially validates the data purification process, computes the minimum required hardware footprint, and compares the efficiency of the proposed analytical framework against established baselines.

7.1. Behavioral Profiling and Parameter Extraction

Applying the K-Means clustering algorithm to the filtered pool of $N = 50$ active mobile agents extracted distinct operational profiles. The autonomous pipeline determined the optimal number of behavioral clusters as $B = 3$, achieving a Silhouette Score of 0.4619. By combining these behavioral profiles with the physical inspection zones, the Spatial Multiplication Algorithm generated a final input vector of $M = 20$ independent traffic classes fed directly into the multi-rate solver.

7.2. Empirical Hardware Reduction

With the exact traffic parameters established, the Inverse Dimensioning Algorithm calculated the minimum central buffer capacity (V_{opt}^0) required to sustain an SLA of $P_{\text{loss}} \leq 0.01$.

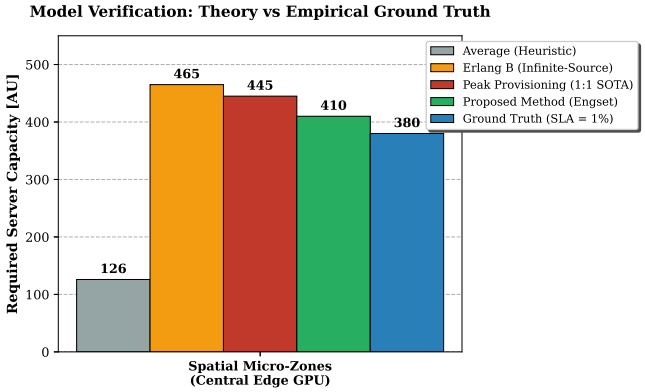

Fig. 1. Empirical verification of Edge AI dimensioning.

As illustrated in Fig. 1, classical infinite-source MEC models (Erlang B) heavily overestimate the required capacity by demanding 465 AU. The standard Computer Vision heuristic of 1:1 Peak Provisioning requires 445 AU. By leveraging the Call Congestion Engset metric to safely truncate the extreme variance tails of the spatial traffic, the proposed methodology dimensions the system at exactly 410 AU. This achieves a direct hardware resource reduction of **7.8%** against Peak Provisioning and mathematically corrects the **11.8%** over-provisioning error of the Erlang B baseline.

7.3. SLA Validation and Ensembling Stability

Despite the aggressive truncation of extreme variance tails (resulting in significant CAPEX savings), the optimized architecture strictly respects the defined SLA. Substituting the measured overflow frame loss into the binomial ensembling model confirms that the effective system recall ($R_{\text{eff\_system}}$) remains > 0.99999.

This empirical result, shown in Fig. 2, proves that highly multiplexed Edge AI environments can safely discard standard 1:1 heuristic dimensioning without affecting safety-critical perception capabilities.

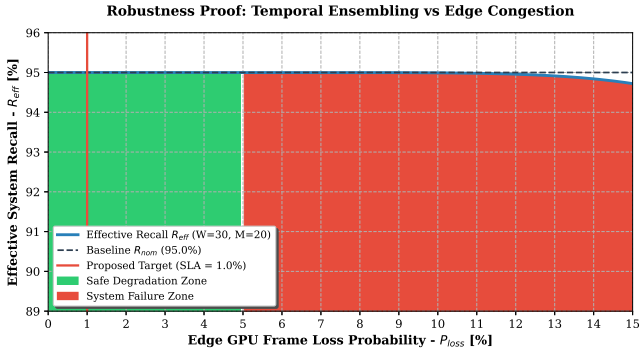

Fig. 2. Effective System Recall ($R_{\text{eff\_system}}$) modeled via Temporal Ensembling. A controlled loss target of $P_{\text{loss}} \leq 0.01$ maintains system accuracy equivalent to the baseline.

8. CONCLUSION AND DEPLOYMENT IMPLICATIONS

This paper presented a closed-world teletraffic framework for precise multi-tier Edge AI dimensioning. By replacing heuristic peak provisioning and flawed infinite-source Poisson assumptions ($Z = 1$) with a multi-service BPP-BPM overflow model [12, 13], the proposed methodology yields an empirically exact hardware capacity requirement. Beyond mathematically proven CAPEX minimization, this approach resolves critical deployment bottlenecks in Industry 4.0. First, reducing the physical accelerator count proportionally lowers recurring per-GPU enterprise software licensing costs (OPEX). Second, it provides IT orchestrators with minimized GPU limits for shared Kubernetes clusters, preventing AI workloads from monopolizing hardware resources needed by parallel industrial workloads like Digital Twins. Ultimately, bridging telecommunications traffic engineering with computer vision establishes a new standard for economically viable and reliable Edge AI scaling. Future work will extend this framework by enabling Tier-1 edge nodes to additionally process specific heavy analytics tasks locally.

9. ACKNOWLEDGMENT

Publication of this article was financially supported by the Ministry of Science and Higher Education of the Republic of Poland. The research and development of the software architecture and algorithmic core described in this study were self-funded by Damian Kmiecik through his private commercial activity under the brand **PerfectSoft**. No public funding was utilized for the technical realization of the project to date.

The system dimensioning and overflow simulations were conducted using the PerfectSoft Network Dimensioning Engine (Academic Edition) [19] provided by PerfectSoft IT (https://perfectsoft.it/rnd/network-dimensioning-engine/).

10. REFERENCES

[1] Raghubir Singh and Sukhpal Singh Gill, "Edge ai: A survey," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 71–92, 2023.

[2] Chunguang Bai, Patrick Dallasega, Guido Orzes, and Joseph Sarkis, "Industry 4.0 technologies assessment: A sustainability perspective," *International Journal of Production Economics*, vol. 229, pp. 107776, 2020.

[3] Abd Al Rahman M. Abu Ebayyeh and Alireza Mousavi, "A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry," *IEEE Access*, vol. 8, pp. 183192–183271, 2020.

[4] Steffen Hagedorn, Marcel Hallgarten, Martin Stoll, and Alexandru Paul Condurache, "The integration of prediction and planning in deep learning automated driving systems: A review," *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 5, pp. 3626–3643, 2025.

[5] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, Sept. 2022.

[6] Linyi Jiang, Yifei Zhu, Hao Yin, and Bo Li, "Hyperion: Low-latency ultra-hd video analytics via collaborative vision transformer inference," 2025.

[7] Hauke Hansen, Wolfgang Huhn, Olivier Legrand, Daniel Steiners, and Thomas Vahlenkamp, *CAPEX Excellence: optimizing fixed asset investments*, John Wiley & Sons, 2011.

[8] Abderrahime Filali, Amine Abouaomar, Soumaya Cherkaoui, Abdellatif Kobbane, and Mohsen Guizani, "Multi-access edge computing: A survey," *IEEE Access*, vol. 8, pp. 197017–197046, 2020.

[9] Djamila Talbi and Zoltan Gal, "Integrating reinforcement learning into m/m/1/k retry queueing models for 6g applications," *Sensors*, vol. 25, no. 12, 2025.

[10] Huan Zheng and Shunfu Jin, "A multi–source fluid queue based stochastic model of the probabilistic offloading strategy in a mec system with multiple mobile devices and a single mec server," *International Journal of Applied Mathematics and Computer Science*, vol. 32, no. 1, pp. 125–138, 2022.

[11] Bin Han, Vincenzo Sciancalepore, Yihua Xu, Di Feng, and Hans D. Schotten, "Impatient queuing for intelligent task offloading in multiaccess edge computing," *IEEE Transactions on Wireless Communications*, vol. 22, no. 1, pp. 59–72, 2023.

[12] Mariusz Głąbowski, Damian Kmiecik, and Maciej Stasiak, "On increasing the accuracy of modeling multi-service overflow systems with erlang-engset-pascal streams," *Journal of Electronics MDPI*, vol. 10, no. 4, pp. 1–26, 2021.

[13] Mariusz Głąbowski and Damian Kmiecik, "Modelling Pascal traffic in overflow systems," *International Journal of Electronics and Telecommunications*, vol. 66, no. 1, pp. 155–160, 2020.

[14] A. Fredericks, "Congestion in blocking systems – a simple approximation technique," *Bell System Technical Journal*, vol. 59, no. 6, pp. 805–827, July–August 1980.

[15] Daghash K. Alqahtani, Muhammad Aamir Cheema, and Adel N. Toosi, "Benchmarking deep learning models for object detection on edge computing devices," in *Service-Oriented Computing*, Walid Gaaloul, Michael Sheng, Qi Yu, and Sami Yangui, Eds., Singapore, 2025, pp. 142–150, Springer Nature Singapore.

[16] Hao Liu and Suhaib A. Fahmy, "Ask the expert: Collaborative inference for vision transformers with near-edge accelerators," 2026.

[17] S. Hassan, G. Mujtaba, A. Rajput, et al., "Multi-object tracking: a systematic literature review," *Multimedia Tools and Applications*, vol. 83, pp. 43439–43492, 2024.

[18] NVIDIA, "PhysicalAI-SmartSpaces: AI City Challenge (Track 1: Multi-camera 3d perception)," 2025, [Online]. Available: https://www.aicitychallenge.org/2025-track1/.

[19] PerfectSoft, "PerfectSoft Network Dimensioning Engine (NDE)," 2026, [Online]. Available: https://perfectsoft.it/rnd/network-dimensioning-engine/.