Guide for Applicants

dAIEDGE 3rd Open Call

Collaborative Projects

Submission starts on **14th August 2025 at 9:00** (Brussels Time),

with deadline on **14th October 2025 at 15:00** (Brussels Time).

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

dAIEDGE has received funding from the European Union's Horizon Europe research and innovation programme, under Grant Agreement No. 101120726.

Acronyms and definitions

Acronym	Meaning
AI	Artificial Intelligence
EC	European Commission
EU	European Union
FAQ	Frequently Asked Questions
GfA	Guide for Applicants
IPR	Intellectual Property Rights
NoE	Network of Excellence
OC	Open Call

CONTENTS

1. Basic Info about dAIEDGE

dAIEDGE is a Network of Excellence (NoE) for distributed, trustworthy, efficient and scalable AI at the Edge.

dAIEDGE seeks to strengthen and support the development of a dynamic European cutting-edge Artificial Intelligence (AI) ecosystem under the umbrella of the European Lighthouse for AI, and to sustain the development of advanced AI. dAIEDGE fosters the exchange of ideas, concepts, and trends on cutting-edge next generation AI, creating links between ecosystem actors to help both the European Commission (EC) and the European Union (EU) identify strategies for future developments in Europe.

dAIEDGE main objective is to advance Europe's innovation and technology base by developing a comprehensive policy and governance approach to AI.

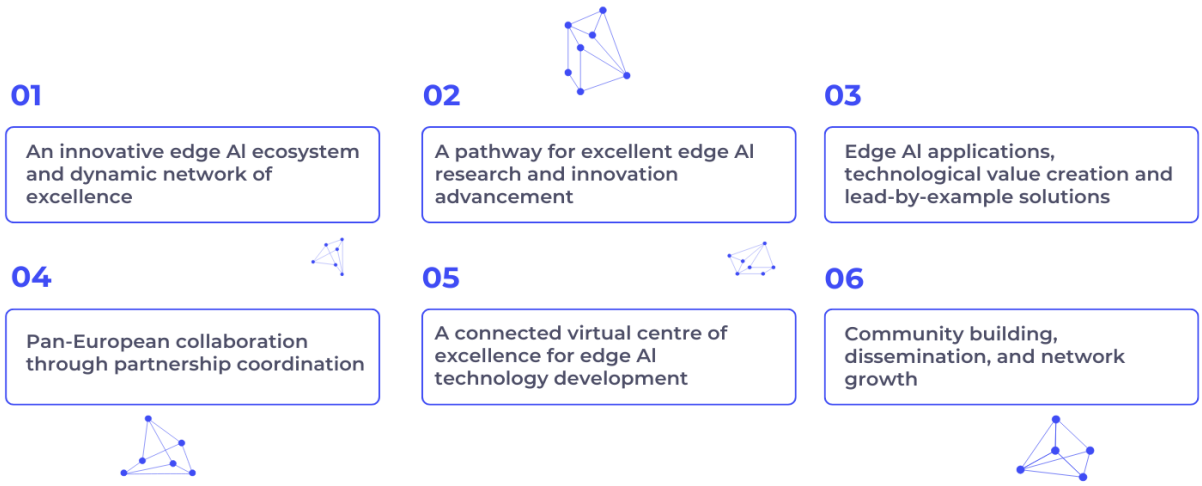


Figure 1 – dAIEDGE objectives

dAIEDGE project partners are a multidisciplinary mix of European-level experts with complementary expertise aligned with the project objectives. dAIEDGE website and Frequently Asked Questions document provide more information about the 35 partners from 15 European Countries coordinated by Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI).

This document summarises the main points of the **dAIEDGE 3rd Open Call for Collaborative Projects**, which will be open **from the 14th of August 2025 at 9:00 (Brussels Time) to the 14th of October 2025 at 15:00 (Brussels Time).**

You can submit your application to this open call and find more information at:
https://daiedge-3oc.fundingbox.com

If you have any technical problems or doubts, tell us at: dAIEDGE.help@fundingbox.com

2. What do we offer?

The open call aims to expand the role of the dAIEDGE Network of Excellence in the edge AI continuum, encompassing micro-, deep-, and meta-edge, with a focus on standardisation, open-source contributions, and infrastructure development. Participants can submit a proposal for research and development activities to solve or push forward the research in the identified Challenges.

The 3rd dAIEDGE Collaborative Projects Open Call aims to distribute up to 600.000 € among up to 10 Proposals to solve the Industrial Challenges defined by the dAIEDGE Consortium. Selected proposals will also have a possibility to use resources from the dAIEDGE Virtual Lab or to add new resources to it or to re-use developments of the dAIEDGE project or dAIEDGE use-cases (see detailed definition of the Challenges in the Annex 1 to this GfA and dAIEDGE VLab details in the Annex 2).

Each third-party project selected under dAIEDGE Collaborative Projects Open Call can receive a **maximum amount of up to €60 000.** For more information on the Support Programme and payment arrangements, check Section 5.

During the two stages of the dAIEDGE Collaborative Projects Support Programme, the beneficiaries will receive **technical mentoring from dAIEDGE technical experts**. They will be supported with the project's development and demonstration of the added value of their solutions to address the dAIEDGE Industrial Challenges.

dAIEDGE is planning some dissemination/information activities about this Open Call. They will be announced at the dAIEDGE Open Call website and dAIEDGE Social Media channels.

3. Eligibility Criteria

To participate in the dAIEDGE Support Programme, the applicant has to meet all the criteria described in Section 3 of this Guide, positively pass the evaluation process and finally sign the Sub-Grant Agreement with the dAIEDGE Consortium.

The proposals that do not comply with the criteria described in this section will be excluded. The eligibility criteria will be checked during the whole evaluation process.

3.1. Who are we looking for?

dAIEDGE Collaborative Projects Open Call looks for **legal entities**, applying **individually** or as a **consortium of up to 2** entities.

The eligible entities are:

i. Research and Technology Organisations (RTO),

ii. Academia or

iii. SMEs[1], including Startups, legally registered as a company at the moment of the application submission to this Open Call,

registered in one of the following eligible countries:

- EU Member States[2] and its Overseas Countries and Territories (OCT), or
- Horizon Europe Associated Countries[3].

Applicants subject to EU restrictive measures under Article 29 of the Treaty on the European Union (TEU) and Article 215 of the Treaty on the Functioning of the EU (TFEU)[4] are not eligible to participate in this Open Call.

[1] An **SME** will be considered as such if it complies with the European Commission's Recommendation 2003/361/EC. As a summary, the criteria defining an SME are:
• Headcount in Annual Work Unit (AWU) less than 250;
• Annual turnover less or equal to €50 million OR annual balance sheet total less or equal to €43 million.
Note that the figures of partners and linked enterprises should also be considered as stated in the SME user guide. For detailed information check EU recommendation:
https://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition_en

[2] Following the Council Implementing Decision (EU) 2022/2506, as of 16th December 2022, no legal commitments can be signed with Hungarian public interest trusts established under Hungarian Act IX of 2021 or any entity they maintain. Affected entities may continue to apply to calls for proposals. However, in case the Council measures are not lifted, such entities are not eligible to participate in the dAIEDGE Open Call. In case of consortium, co-applicants will be invited to remove or replace that entity. Tasks and budget may be redistributed accordingly

[3] AC as of 12.12.2024: Albania, Armenia, Bosnia and Herzegovina, Canada, Faroe Islands, Georgia, Iceland, Israel, Kosovo, Moldova, Montenegro, North Macedonia, Norway, Serbia, Türkiye, Tunisia, United Kingdom, Ukraine, for the most up-to-date list please first part of this document. For the avoidance of doubt, New Zealand and Switzerland are not eligible in this open call.

[4] Please note that the EU Official Journal contains the official list and, in case of conflict, its content prevails over that of the EU Sanctions Map

dAIEDGE Consortium partners, their affiliated entities, board members, employees or permanent collaborators CANNOT be beneficiaries of dAIEDGE Open Call.

3.2. What types of activities can be funded?

The activities that qualify for financial support under dAIEDGE Collaborative Projects Open Call are the **research and development activities** to solve or push forward the **research in the Industrial Challenges defined by the dAIEDGE Consortium**, with the possibility of using resources from dAIEDGE Virtual Lab or adding new resources to it, or to re-use developments of the dAIEDGE project or dAIEDGE use-cases.

Each single applicant or consortium must select and reply in its proposal to **one** of the following Industrial Challenges (details in Annex 1):

1. Balancing Energy Efficiency and Trustworthiness in Federated Learning Across the Edge–Cloud Continuum
2. Enabling On-Device Training Support in TFLite Micro for Resource-Constrained Edge Devices
3. Very low consumption of speech enhancement AI algorithm at the edge
4. High-Performance Bayesian-based AI Agents for Embedded Intelligence in Smart Environments
5. Towards a Unified Semantic and Hyperspatial World Model: Enabling Smart Edge Systems through Domain-Aware Graphs and Agent Governance
6. Automated Edge deployment, tuning and Performance Evaluation of Binary Neural Networks targeting FPGA enabled platform
7. Towards Automatic DMWay Configuration with Generative AI
8. Health Monitoring AI for Edge Devices with Privacy-First
9. Edge AI framework for web browsers empowering Federated Machine Learning applications addressing Classification problems over multiple format data (text and images)
10. Off-Chip Weights for Streaming Architectures
11. 2D Vehicle Detection Network on PYNQ-Z1
12. Real-Time Vision Transformer (ViT) on FPGA SoC for Image Classification
13. Generative AI for object detection in Edge scenarios

The activities within the dAIEDGE Project should start at a **Technological Readiness Level** (TRL) of **2-3 and to achieve a TRL 4-5**. You can also check the FAQ document for a detailed explanation of TRL.

Your project should have a clear **European Dimension**, meaning fostering projects that generate a substantial positive impact on the European society and economy.

3.3. Ground rules

When applying to the dAIEDGE Open Call, please also consider the following:

- **Be on time and use our system:** only proposals submitted through the online form before the deadline of 14th October 2025 at 15:00 (Brussels Time) will be evaluated and considered for funding. If you submit the form correctly, the system will send you a confirmation of your submission. Get in touch with us if it is not the case.

- **English Language:** proposals must be written in English in all mandatory (including the attachments/uploaded files) parts to be eligible. Only parts written in English will be evaluated. If the mandatory parts of the proposal are in any other language, the entire proposal will be rejected. If only non-mandatory parts of a proposal are submitted in a language different from English, those parts will not be evaluated but the proposal is still eligible.

- **Every question deserves your attention:** all mandatory sections - generally marked with an asterisk - of the proposal must be completed. The data provided should be actual, true and complete, and should allow assessment of the proposal. Additional material, not specifically requested in the online application form, will not be considered for the evaluation.

- **Be exhaustive:** applicants have to verify the completeness of the form, as it won't be possible to add any further information after the deadline. After the proposal is submitted, the form can be modified until the deadline.

- **Conflicts of interest:** the existence of a potential conflict of interest among applicants and one or more Consortium partners will be taken into consideration. Consortium partners, their affiliated entities, employees, or persons treated as personnel (e.g. working under B2B contract) cannot take part in the dAIEDGE programme. All cases of potential conflict of interest will be assessed case by case. See EC definition of Conflict of Interest[5] and check our FAQ for more info.

- **Healthy finances and a clean sheet are mandatory:** we don't accept entities that are under liquidation or are an enterprise under difficulty[6] according to the Commission Regulation No

[5] EC definition of Conflict of Interest: https://commission.europa.eu/strategy-and-policy/eu-budget/protection-eu-budget/conflict-interest_en

[6] An enterprise will be considered an undertaking in difficulty if more than half of the capital has disappeared. This refers to the loss of "subscribed share capital". If profit and loss reserves deficit more than 50% of share capital, there is a potential problem with the company. (Article 2, item 18 point a) and b)

(a) In the case of a limited liability company (other than an SME that has been in existence for less than three years […]), where more than half of its subscribed share capital has disappeared as a result of accumulated losses. This is the case when deduction of accumulated losses from reserves (and all other elements generally considered as part of the own

651/2014, art. 2.18, or that are excluded from the possibility of obtaining EU funding under the provisions of both national and EU law, or by a decision of both national or EU authority. We also don't accept entities that are meeting national regulations regarding bankruptcy.

- **It is your proposal:** Proposals should be based on your original work. If the proposal incorporates intellectual property (IP) not originally yours, the right to use the IP must be clear. Any work related to the implementation of the project described in the application may not violate the IPR of third parties, and the IPR of the application project may not be the subject of a dispute or proceedings for infringement of third-party IPR.
  dAIEDGE Consortium encourages the use of Open Source Software and Hardware Solutions. IPR and confidentiality issues will be assessed case by case and relevant provisions will be included in each Sub-Grant Agreement if necessary.

- Applicants must ensure that their proposals have an exclusive focus on **civil applications**. Military use is not allowed and such projects will not be funded by the dAIEDGE.

- **Focus on quality**: to streamline the process, we encourage applicants to focus on a single, high-quality submission, even though a maximum of three applications (one per Industrial Challenge) is permitted per entity or consortium. Please note that dAIEDGE will only fund any given team member or legal entity once. If multiple applications from the same applicant are shortlisted to be selected, only the highest-ranked proposal across all Industrial Challenges will be funded.
  Beneficiaries from dAIEDGE's 1st or 2nd Open Call are ineligible to receive funding in the 3rd Open Call (neither team members nor any legal entities can be funded twice by dAIEDGE).

- **Gender Equality Plan:** public bodies, higher education institutions, and research organisations from EU countries and associated countries must have a Gender Equality Plan (GEP)[7].

funds of the company) leads to a negative cumulative amount that exceeds half of the subscribed share capital. For the purposes of this provision, 'limited liability company' refers in particular to the types of company mentioned in Annex I of Directive 2013/34/EU (1) and 'share capital' includes, where relevant, any share premium

.

(b) In the case of a company where at least some members have unlimited liability for the debt of the company (other than an SME that has been in existence for less than three years […]), where more than half of its capital as shown in the company accounts has disappeared as a result of accumulated losses. For the purposes of this provision, 'a company where at least some members have unlimited liability for the debt of the company' refers in particular to the types of company mentioned in Annex II of Directive 2013/34/EU.

Please note that, if your SME exist from less than three years, you won`t be considered as undertaking any difficulties.

[7] For more details please check https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/democracy-and-rights/gender-equality-research-and-innovation_en

- **Acceptance of the Open Call rules:** to apply for this Open Call, applicants have to accept its rules and regulations detailed in this Guide for Applicants.

If extra hints on how to prepare the application are needed, applicants can check out the section: extra Hints to submit your proposal.

4. How will we evaluate your proposal?

dAIEDGE evaluation process is transparent, fair and equal to all participants, with a clearly defined complaint procedure (see section Complaints). dAIEDGE will evaluate the proposals in 5 phases, as shown in the following graphic:

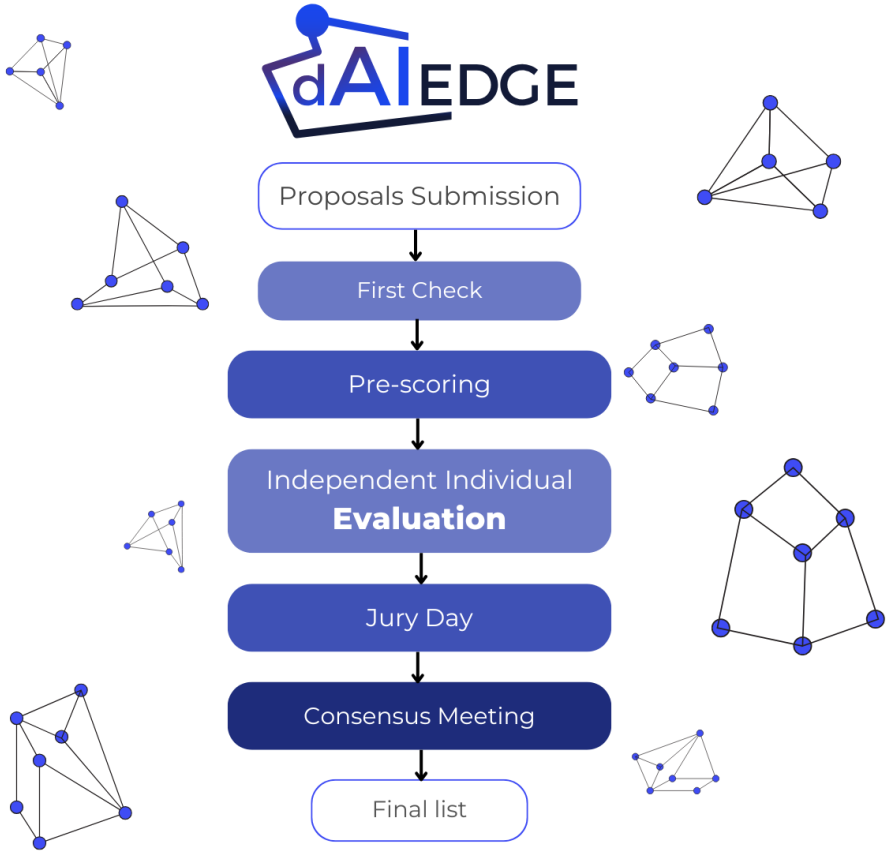


Figure 2 – dAIEDGE evaluation process

For this call, dAIEDGE is looking for the best fit for the dAIEDGE project, and a high volume of applications is anticipated. Since the primary concern is quality over quantity, applicants are encouraged to put effort into presenting their applications in the best possible way, offering as much detail as possible. This will assist us in evaluating the applications and identifying how the proposals align with the overall scope of dAIEDGE (see Section 3.2).

4.1. First Check

After the closure of the Open Call, the proposals will be reviewed to ensure they meet the conditions outlined in Section 3. This assessment will be based on the statements provided in the proposal.

At this stage, the eligibility criteria are checked against the Declaration of Honour or self-declarations included in the application form, and they will be continuously verified throughout the evaluation process, including the final formal check.

Projects that do not comply with the above-mentioned criteria will be rejected, and this could happen at any stage of the evaluation process.

4.2. Pre-scoring

As we expect a lot of applications for this call, a pre-scoring system may be included. If we receive more than 100 eligible proposals, the eligible proposals will be automatically scored by the application submission system according to the following criteria:

- Experience background (15 % / up to 15 points)
- Proven experience in AI (15% weight/up to 15 points),
- Team skills (15% weight/up to 15 points),
- Dedication to the project (15% weight/up to 15 points)
- Innovation level (maturity of the technological solutions) - 15% weight / up to 15 points
- Competitiveness (15% weight/up to 15 points)
- Social & environmental impact (5%/up to 5 points)
- Previous projects funded by EU (5 % weight/up to 5 points)

In total, your project can receive up to 100 points. As an applicant, you should select the option that best describes the real characteristics of your project in relation to each criterion and confirm that no false declarations are made.

If your score places you in the pool of up to 100 top-ranked applications, we will forward your application to independent experts for individual evaluation (See Section 4.3 below).

We will inform you about the results of the first check and the pre-scoring phase soon after these steps have been completed.

4.3. Independent Individual Evaluation

In this phase, each application that passed the pre-scoring evaluation will be evaluated by 2 external independent experts.

The applications will be evaluated within the following awarding criteria:

(1). EXCELLENCE will evaluate:

Ambition: Applicants have to demonstrate how their proposal contributes to the project scope, has a European dimension and is beyond the State of the Art. They need to describe the innovative

approach behind the project (like ground-breaking objectives, novel concepts and approaches, new products, services or business).

Innovation: Applicants should provide information about the level of innovation and about the degree of differentiation that their project will bring.

Soundness of the approach and credibility of the proposed methodology.

Gender dimension: Applicants have to demonstrate to what extent the gender dimension has been integrated into the research and innovation content.

(2). IMPACT will analyse:

Market opportunity: Applicants have to demonstrate a clear idea of what they want to do and whether the new/improved product has market potential.

Competition: Applicants have to provide information about the degree of competition of their product/service and why their idea is disruptive and breaks the market. They should explain why their products/services can be differentiated from the competition.

Commercial Strategy and Scalability: Applicants must demonstrate the level of scalability of the new/improved product, meaning by not addressing a specific problem but being able to solve a structural problem in a specific sector/process/etc.

Environmental and social impact: Applicants should demonstrate their project contribution towards environmental, social and economic impacts to contribute to sustainable development, Green Deal and other European policies.

(3). IMPLEMENTATION will consider:

Team: Applicants have to demonstrate the team's management and leadership qualities, their ability to take a concept from ideas to market, their capacity to carry through their ideas and understand the dynamics of the market they are trying to tap into. The team should be cross-functional, with a strong background and skills base and considering its gender balance.

Resources. Applicants should demonstrate the quality and effectiveness of the resources assigned to get the objectives proposed.

The evaluators will score each criterion on a scale from 0 to 5:

0 - **The proposal fails to address the criterion** or it cannot be assessed due to missing or incomplete information.

1 - **Poor** – The criterion is inadequately addressed or there are serious inherent weaknesses.

2 - **Fair** – The proposal broadly addresses the criterion, but there are significant weaknesses.

3 - **Good** – The proposal addresses the criterion well but a number of shortcomings are present

4 - **Very good** – proposal addresses the criterion very well but a small number of shortcomings are present.

5 - **Excellent** – The proposal successfully addresses all relevant aspects of the criterion. Any shortcomings are minor.

Each evaluator will produce an **Individual Evaluation Report**. Once the Individual Evaluation Reports are submitted, **the final score for each individual criterion** will be calculated as the average of the scores provided by each evaluator. The **final score per each application** will be calculated as the sum of the scores for each individual criterion.

Cases of significant divergence between the evaluators' scoring will be discussed by evaluators during a Consensus Group. In case no consensus is reached between the evaluators, an additional evaluator will be included to provide an extra evaluation.

The final score threshold for each individual criterion is 3 out of 5 points. T. The overall threshold, applying to the **final score per application**, is 10 out of 15 points

Once the initial ranking is obtained, **ties** (if any) will be solved using the following criteria in order of priority:

- The highest score in the Excellence criterion.
- Gender balance (prioritizing teams with more female team members) among the personnel responsible for carrying out the activities.
- The highest score in the Impact criterion.

As a result of the Independent Individual Evaluation, a **'Ranking List per Challenge'** and a **"General Ranking List"** will be produced. All proposals that reach the threshold or are scored above the threshold, will be considered for the next phase.

Please note that we need time to process all the proposals in this phase, so you probably won't hear back from us for a while.

4.4. Jury Day + Consensus Meeting

After the Independent Evaluation, around **30 top highest-scoring applications** will be invited to an online Jury Day (pitching session), where applicants will present their project, team, and vision in front of the **Selection Committee** and external independent experts and reply to their questions. Finalists will be ranked based on the following Awarding Criteria:

- Team vision and project vision
- Potential impact on the project's goals

- Positive effect on the applicant's processes, products, or services

The Jury Day will be followed by a Consensus Meeting during which finalists will be selected.

4.5. Consensus Meeting

The Selection Committee (including 2 Experts who participate in the Independent Evaluation) will review and discuss the results of the Independent Evaluation and the online pitching session. They will decide on a *Provisional List of FSTP Recipients* and *Reserve List* (if needed) based on the Independent Evaluation scores, results of online pitching sessions (see the criteria in sections 4.3 and 4.4), and considering how the proposed projects are aligned with dAIEDGE scope. The Selection Committee will make the final decision by a consensus or majority of ⅔ votes.

dAIEDGE aims to select 10 Proposals for 10 different Industrial Challenges, but the exact number of proposals approved will be decided based on the overall **quality** of the proposals. Although the highest-scoring proposals are usually chosen for funding, the Selection Committee can reject a candidate for valid reasons, such as not fitting dAIEDGE goals and scope, limited potential impact, commercial competition issues, serious ethical concerns, or possible conflicts of interest.

Applicants will be notified after the Jury Day and Consensus Meeting about their status.

4.6. Formal Check, Sub-Grant Agreement Preparation and Signature

Before the selected applicants get started with the dAIEDGE Collaborative Projects Programme, they need to sign the Sub-Grant Agreement with the dAIEDGE Consortium.

Before signing the Agreement, applicants should provide documents regarding their formal status. The dAIEDGE Consortium will verify the documents to prove the applicants' eligibility.

Applicants should be extremely vigilant to:

1. **The nature of the documents requested.** If the documents provided do not prove the eligibility, applicant participation will end there.

2. **The deadlines for submission of these documents.** If an applicant fails to deliver the requested documents on time, without a clear and reasonable justification, we will have to exclude that proposal from further formal assessment. Another applicant from the *Reserve List* will then take the place.

5. The Collaborative Projects Support Programme and Payment Arrangements

Once the applicant's eligibility has been confirmed following the formal check, and the Sub-Grant Agreement is signed, the applicant will become an official beneficiary of the dAIEDGE Programme.

Beneficiaries will enrol in an up to 7 months Technical Support Programme, receiving technical mentoring from dAIEDGE experts for the development of their projects. Each third-party project can receive a maximum lump sum of up to €60 000.

The lump sum is a simplified method of settling expenses in projects financed from Horizon Europe Programme funds. It means that the beneficiary is not required to present strictly defined accounting documents to prove the cost incurred (e.g. invoices) but is obliged to demonstrate that the implementation of the project is in line with the milestones set for it.

Simply speaking, it means that the Selection Committee will carefully assess beneficiaries' progress and the quality of their work during Milestone Reviews after each stage of the Programme. The milestones (deliverables, KPIs, and ethical recommendations, as well as the budget) will be fixed in the **'Individual Mentoring Plan'** elaborated at the beginning of the programme.

The lump sum method does not exempt beneficiaries from collecting documentation to confirm the costs under their national fiscal regulations.

The grant will be paid after the Milestone Review of each stage, following this scheme:

Stage	Name	Duration	Deliverable	Funding
Stage 1	Plan Phase	1 month	Individual Mentoring Plan	15% of the grant, up to €9 000
Stage 2	Implementation Phase	6 months	Proof of Demonstration	85% of the grant, up to €51 000
			TOTAL	up to €60 000

A delayed payment mechanism could be applied to the payments if decided by the dAIEDGE Consortium. This mechanism implies that up to 15% of each tranche will be paid to the beneficiaries who completed the given stage once the whole dAIEDGE Project is completed. This should happen approximately 9 months after the end of the Project. The expected end of the dAIEDGE Project is September 2026. Relevant provisions will be included in the Sub-Grant Agreements. Please bear in mind that dAIEDGE project might be extended.

Milestones Review

Figure 3 – dAIEDGE milestone review and payment process

Beneficiaries performance during the Collaborative Projects Support Programme will be reviewed at the Milestone Review (established every time a payment is due), according to the following criteria:

- Deliverable quality (30%).
- Technical performance indicators (60%).
- Deadline Compliance (10%).

Each criterion will be scored from 0 to 10 and, based on the weight of each criterion, the final score will be calculated.

According to this final score, beneficiaries over the threshold (7 points) will successfully receive the corresponding part of the grant and continue the programme. The beneficiaries who have not reached the threshold (7 points) will be invited to leave the programme without receiving the corresponding payments.

The Selection Committee will review and validate the evaluations, paying special attention to the 'under threshold' cases, if any, by taking into consideration all possible objective reasons for underperformance (i.e. external factors that might have influenced the beneficiaries' performance). The dAIEDGE Selection Committee will make the final decision and approve the payments or invite beneficiary projects which have not reached the threshold to leave the programme.

6. Contact us

How can we help you?

If you have any questions regarding our application process, feel free to email us at dAIEDGE.help@fundingbox.com

Responses to any questions are provided on an individual basis, do not constitute a change to this Guide for Applicants, and are provided for informational purposes.

In case of any technical issues or problems, please include the following information in your message:

- your username, project acronym, phone number and email address;
- details of the specific problem (e.g. error messages you encountered, bug description, i.e. if a dropdown list isn't working, etc.); and
- screenshots of the problem.

Complaints

If you believe that a mistake has been made after receiving the results of one of the evaluation phases (when foreseen), you may submit a complaint. To do so please email us your complaint in English at dAIEDGE.help@fundingbox.com and include the following information:

- your contact details (including email address);
- the subject of the complaint;
- information and evidence regarding the alleged breach.

You have three (3) calendar days to submit your complaint, starting from the day after the communication with the results was sent. We will review your complaint within seven (7) calendar days from its reception. If we need more time to assess your complaint, we will inform you about the extension by email. We will not review anonymous complaints as well as complaints with incomplete information.

Please take into account that the evaluation is run by experts in the relevant field, and we do not interfere with their assessment. Therefore, we will not evaluate complaints related to the results of the evaluation other than those related to procedural or technical mistakes.

7. Final provisions

Any matters not covered by this Guide will be governed by Polish law and rules related to the Horizon Europe Programme and general rules of EU grants.

Please consider that we make our best effort to keep all provided data confidential; however, for the avoidance of doubt, you are solely responsible for indicating your confidential or sensitive information as such.

Your Intellectual Property Rights (IPR) will remain your property.

For the selected beneficiaries, the Sub-grant agreement will include a set of obligations towards the European Commission (for example: promoting the project and giving visibility to the EU funding, maintaining confidentiality, understanding potential controls by the EC/ECA, EPPO and OLAF).

The dAIEDGE Consortium might cancel the call at any time, change its provisions or extend it. In such a case we will inform all applicants about such change. The signature of the Sub-grant agreement is an initial condition to establish any obligations among applicants and any Consortium partners (with respect to the obligation of confidentiality of the application).

8. Extra hints before you submit your proposal

A proposal takes time and effort, and we know it. Here are a few crucial points you should read before submitting your proposal.

- Is your proposal in line with what dAIEDGE is looking for? You are not sure? You can consult Section 3.1 and Section 3.2.
- Did you present your project in a way that will convince evaluators? Not sure if you did? Go back to Section 4.
- Does your project fulfil all eligibility requirements described in the Guide? Check again Section 3.
- Are you sure you can cope with our process of the Sub-Grant Agreement signature and payment arrangements for selected proposals? You may want to go over Section 4.5.
- Did you check our Sub-Grant Agreement template? You didn't? Check it here.
- Do you need extra help? Contact us.

Good luck!

ANNEX 1: dAIEDGE Challenges

ANNEX 2: DAIEDGE Lab

daiedge.eu

dAIEDGE 3rd Open Call. COLLABORATIVE PROJECTS.

Challenges

1. Balancing Energy Efficiency and Trustworthiness in Federated Learning Across the Edge–Cloud Continuum
2. Enabling On-Device Training Support in TFLite Micro for Resource-Constrained Edge Devices
3. Very low consumption of speech enhancement AI algorithm at the edge
4. High-Performance Bayesian-based AI Agents for Embedded Intelligence in Smart Environments
5. Towards a Unified Semantic and Hyperspatial World Model: Enabling Smart Edge Systems through Domain-Aware Graphs and Agent Governance
6. Automated Edge deployment, tuning and Performance Evaluation of Binary Neural Networks targeting FPGA enabled platform
7. Towards Automatic DMWay Configuration with Generative AI
8. Health Monitoring AI for Edge Devices with Privacy-First
9. Edge AI framework for web browsers empowering Federated Machine Learning applications addressing Classification problems over multiple format data (text and images)
10. Off-Chip Weights for Streaming Architectures
11. 2D Vehicle Detection Network on PYNQ-Z1
12. Real-Time Vision Transformer (ViT) on FPGA SoC for Image Classification
13. Generative AI for object detection in Edge scenarios

CHALLENGE 1. Balancing Energy Efficiency and Trustworthiness in Federated Learning Across the Edge–Cloud Continuum

The challenge focuses on developing and evaluating strategies for balancing energy efficiency and trustworthiness in federated learning (FL) across the edge–cloud continuum. Participants are expected to design mechanisms that dynamically adapt training decisions (e.g., which devices participate or defer training) based on factors such as battery level and device availability, while carefully considering the trade-offs between energy savings and trust dimensions, including data privacy, model accuracy, and system robustness.

Key research questions include:

How can energy-aware scheduling of FL clients be made without compromising privacy or significantly degrading model accuracy?

What mechanisms can quantify and manage the impact of skipping or offloading training (due to energy constraints) on both trustworthiness and model performance?

EXPECTED DELIVERABLE:

- A prototype FL system or simulation framework (Software)
- Source code with documentation
- Experimental results demonstrating energy–trust trade-offs under different scenarios
- A final report explaining design decisions, energy/trust metrics used, and key findings
- Optional: policy recommendation for trust-aware energy optimization in FL

REFERENCE INFO.

- Datasets: e.g., MNIST
- FL Frameworks: e.g., Flower
- Trust Metrics: Data locality, privacy leakage risk estimation, model accuracy, participation fairness
- Energy Models: Simulated or real device profiles (e.g., Jetson Nano, Raspberry Pi)
- Measurement of CO2 emissions: e.g., ECO2AI library
- Benchmarks:
- Energy consumption per round (joules)
- Accuracy drop (%) vs. baseline
- Privacy exposure (qualitative or through metrics like mutual information leakage)

DIAGRAM & TECHNICAL SPEC:

- Input Spec (incl. pre-processing):
- Sensor or user-generated data (pre-processed according to use case: normalization, segmentation, etc.)
- Metadata from devices (battery level, connectivity, energy profile)
- Output Spec (incl. post-processing):
- Trained global model
- Logs on energy usage, device participation, accuracy trends
- Optional: Trustworthiness risk indicators (e.g., proxy for privacy exposure, data remoteness)

REQUISITES:

Mandatory:

- Integration of energy-awareness in FL client selection
- Monitoring and reporting of at least one trust-related impact (e.g., data movement, privacy trade-offs, accuracy degradation)
- Use of a standard FL framework (e.g., Flower, TFF)

Preferential:

- Real-world device simulation or hardware-in-the-loop
- Implementation of adaptive policies balancing trust and energy
- Use of differential privacy or secure aggregation mechanisms
- Exploration of battery-aware fairness strategies in FL

EXPECTED RESULT:

- Energy Reduction: ≥20% lower cumulative energy consumption compared to baseline FL (all clients participate)
- Accuracy Drop Tolerance: ≤5% accuracy loss compared to full participation baseline
- Trustworthiness Indicators:
- Reduction in data movement across trust boundaries (qualitative or proxy metrics)
- Participation fairness maintained within ±15% variance across devices
- Adaptivity: Demonstrated system reaction to dynamic availability (e.g., devices skipping rounds due to low battery) with graceful degradation

CHALLENGE 2: Enabling On-Device Training Support in TFLite Micro for Resource-Constrained Edge Devices

TensorFlow Lite for Microcontrollers (TFLite Micro) provides inference-only support for running machine learning models on ultra-low-power devices. This challenge aims to extend TFLite Micro to support on-device training (ODT), allowing models to adapt in real-time to local data. Participants will investigate architectural extensions, memory and compute optimizations, and integration pathways to contribute this functionality upstream to the TFLite Micro open-source repository. A preliminary work carried out by the dAIEdge Consortium will be used as baseline for extending the ODT capabilities of TFLite Micro. The challenge is a strategic step by dAIEdge toward enabling decentralized, adaptive AI on the tiniest edge platforms.

EXPECTED DELIVERABLE:

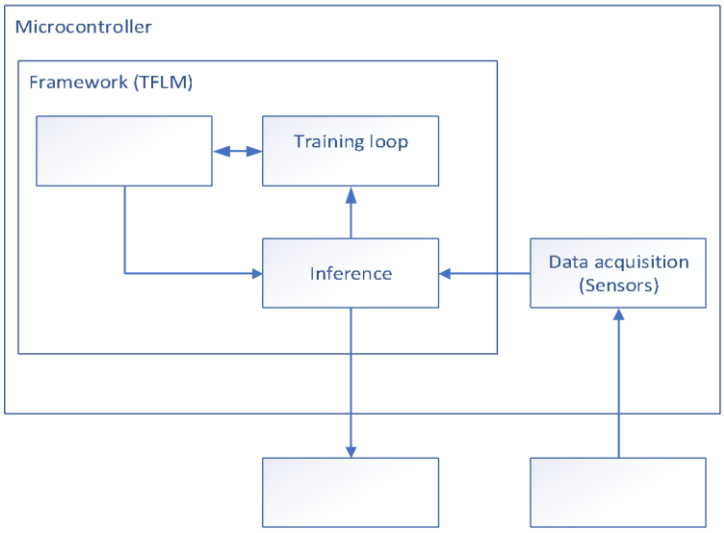
- Functional software prototype of on-device training support in TFLite Micro
- Technical documentation of the architecture, APIs, memory model, and supported use cases
- Benchmark report comparing performance (inference + training) before and after modifications
- Pull request or contribution proposal to the TFLite Micro OSS (Open Source Software) repository
- Optional: Demonstration video of the prototype running on real hardware

REFERENCE INFO:

- TensorFlow Lite Micro (https://www.tensorflow.org/lite/microcontrollers)
- STM32, or other Cortex-M platforms
- TinyML benchmark models (keyword spotting, person detection, anomaly detection)
- Google's Person Detection and Speech Commands v0.02 datasets
- Baseline: Inference-only performance of TFLite Micro

DIAGRAM & TECHNICAL SPEC:

Global Picture of the Challenge:



Input Specification:

- Raw sensor data (e.g., audio, IMU, image)
- Preprocessing: quantization, normalization

Output Specification:

- Classification, detection, or anomaly result

- Training statistics (loss, accuracy, training time, etc.)

REQUISITES:

Mandatory:

- C/C++ programming for embedded systems
- Familiarity with TFLite Micro and memory-constrained environments
- Knowledge of backpropagation, gradient descent

Preferential:

- Experience contributing to open-source ML frameworks
- Familiarity with microcontroller platforms (e.g., ARM Cortex-M, RISC-V)
- Use of profiling/debugging tools for embedded AI

EXPECTED RESULT:

- Functional on-device training within <1MB RAM footprint
- Training throughput: ≥1 sample/sec for small models (e.g., 1D CNN for audio)
- <20% increase in inference latency after adding training support
- OSS contribution accepted or under review by TFLite Micro maintainers
- Demonstrated use case on physical device

CHALLENGE 3. Very low consumption of speech enhancement AI algorithm at the edge

To promote multimodal data exploitation in dAIEDGE, we propose to enrich the algorithm library developed in the project (Whereas the algorithms of this library would be mainly focused on images, we intend here to contribute to it also with algorithms on audios which could complement also fit overall purpose (also some of the Use Cases envisioned) with other modalities, in particular here with audio data).

This is highly relevant to release full potentiality of multimodality in a variety of Use Cases targeting AI at the Edge such as the Smart Warehouse monitoring use case depicting hereafter where multimodality could be used to significantly enrich intrusion detection capabilities based on multimodal data including audio and image data.

For this open call, we are looking for:

1) A HW platform on top of which a speech AI enhanced algorithm could be embedded and run at very low consumption (< 50 mW).
2) The development of one speech AI enhanced algorithm. LSTMs, Transformers or Mamba could be good candidates. The list is not exhaustive.

The whole system should be able to get on-boarded on a drone.

EXPECTED DELIVERABLE: On the software level, one must be able to input an ONNX speech enhancement model and a few audio files to the virtual lab and get metrics desired on the outputs of the algorithm: audio metrics (i.e. PESQ, SNR) and hardware metrics (i.e. inference time, consumption).

REFERENCE INFO

State of the art.

The current state of the art, in the context of implementation in very low consumption constraints, is the use of small LSTMs, which has its limitations. Meantime more recent results on Transformers and Mamba-type architectures have proved their worth “off-edge” and appear ready to move to the edge.

Datasets

Public datasets could be used for this open call: Librispeech, Voxceleb for example. Moreover, for some datasets, improvements have been done with the inclusion of noise to test the robustness of the algorithms.

Benchmarks

Experiments on some of these datasets are also available with the testing of denoising algorithms such as WaveUnet, DCCRN and TinyDenoiser). Results could be used as baselines for the benchmarks of the new algorithms and hardware implementation that should be done in this open call.

dAIEDGE VLab infrastructure can be used.

REQUISITES:

- Speech enhancement algorithms that will improve baselines defined in section 4: Mandatory
- Very low consumption speech enhancement pipeline: Mandatory
- Integration of the pipeline with ONNX: Preferential
- Integration in the dAIEdge-Vlab: Preferential
- Benchmark process automation

EXPECTED RESULT:

- Total time computation with specific constraints (to be defined)
- Total energy consumed with specific constraints (to be defined)

CHALLENGE 4. High-Performance Bayesian-based AI Agents for Embedded Intelligence in Smart Environments

SHORT DESCRIPTION:

Edge computing enables real-time processing for industrial and autonomous systems, where rapid, localized decision-making is essential. However, conventional AI systems often fail in dynamic settings. This limitation underscores the need for intelligent agents capable of autonomous operation at the edge, with the ability to continuously adapt to evolving conditions.

Active Inference (AIF) models provide a principled framework to address this challenge, empowering agents to perceive, reason, and act based on continuous feedback from their environment. Many Bayesian algorithms and routines are used within AIF models, which are core tools for perception and decision-making in uncertain environments. However, the efficient deployment of such routines in resource-constrained systems is complex.

This challenge invites participants to select an innovative environment use case (e.g., intelligent robotic systems in warehouses, real-time anomaly detection from satellite imagery, autonomous drones navigating urban airspaces, etc) and develop a Bayesian-based model for its operation. The focus is on algorithmic domain-specific optimization for embedded hardware, leveraging known constraints of probabilistic computations that can be exploited to accelerate implementation on hardware.

EXPECTED DELIVERABLE:

Implementation:

1. Selection of an application in a simulated smart environment, e.g., autonomous drones in smart cities, mobile robots in smart warehouses, embedded analytics in satellite payloads, etc.
2. Development of a Bayesian-based AI agent using algorithms such as Kalman Filters, Particle Filters, Bayesian (multinomial) logistic regression, etc.
3. Hardware-aware algorithm optimization on at least one hardware platform (see deployment section), showcasing substantial performance gains (latency, memory, energy) versus a standard implementation.

Evaluation:

A comprehensive evaluation of the optimized implementation versus a reference version regarding runtime, computational efficiency, memory usage, and inference accuracy on the selected embedded device (GPU, FPGA, or MCU).

Deployment & Demonstration:

Deployment of the optimized algorithm in a simulation environment on one of the following HW platform types: GPU, FPGA, Micro-controller Unit (MCUs).

Specific platform-based optimization to be considered:

- Custom CUDA kernels for GPUs (e.g., Nvidia Jetson)
- FPGA-based accelerator for real-time processing (e.g., Zync)
- Lightweight C implementations targeting MCUs, e.g., STM32

Bonus: integration and demonstration of the algorithm into a perception or control stack to create an Active Inference agentic system (e.g., SLAM, path planning, anomaly detection).

Note: No HW will be provided by dAIEDGE. Already having the selected HW platform on-premises is beneficial.

Documentation:

The report shall include architectural decisions, optimization techniques (e.g., matrix structure exploitation, loop unrolling, SIMD), benchmark results, and application potential.

REFERENCE INFO:

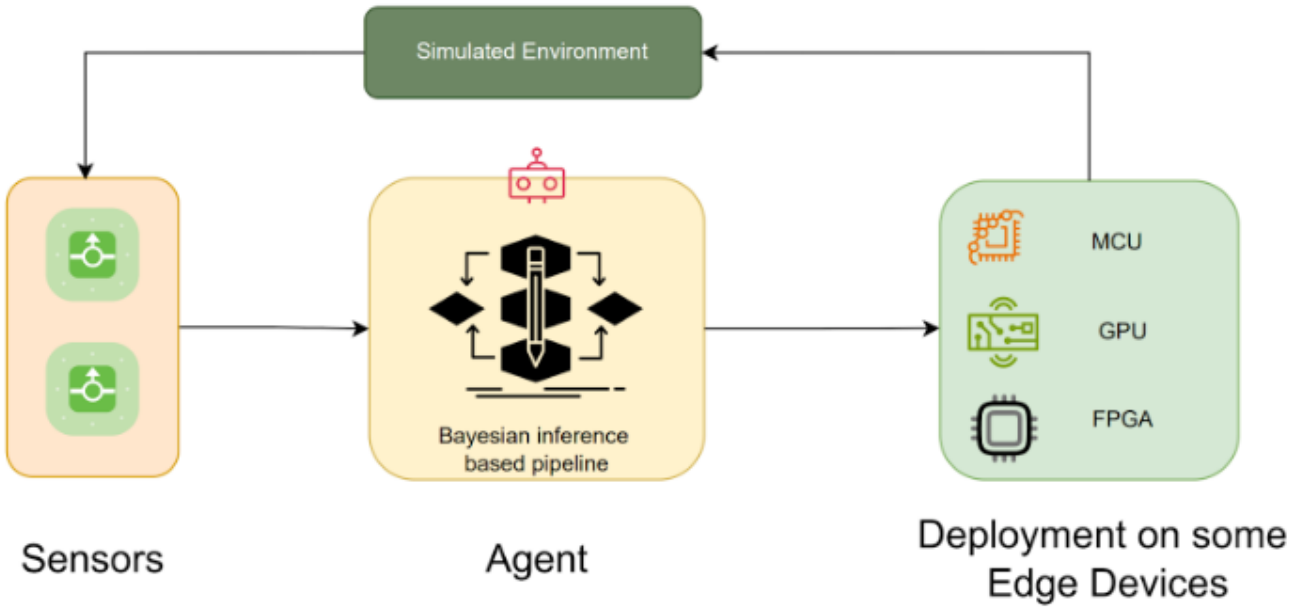
Research Papers:

- Sedlak, Boris, et al. "Active inference on the edge: A design study." 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops). IEEE, 2024.
- Hamburg, Sarah, et al. "Active Inference for Learning and Development in Embodied Neuromorphic Agents." Entropy 26.7 (2024): 582.
- Heins, Conor, et al. "pymdp: A Python library for active inference in discrete state spaces." arXiv preprint arXiv:2201.03904 (2022).
- Pezzato, Corrado, et al. "Active inference and behavior trees for reactive action planning and execution in robotics." IEEE Transactions on Robotics 39.2 (2023): 1050-1069.
- Wei, Ran, et al. "Navigation under uncertainty: Trajectory prediction and occlusion reasoning with switching dynamical systems." arXiv preprint arXiv:2410.10653 (2024).
- de Prado, Miguel, et al. "Automated design space exploration for optimized deployment of dnn on arm cortex-a cpus." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 40.11 (2020): 2293-2305.

Sample Implementations:

- Pymdp: https://github.com/infer-actively/pymdp
- https://github.com/mbaltieri/PIDcontrolActiveInference
- https://github.com/openai/gym
- https://cyberbotics.com/

DIAGRAM & TECHNICAL SPECS:



Input spec:	**Agent:**	**Output Spec:**
Input sensor data	Sense, reason	Action in the environment

REQUISITES:

- **Mandatory**:
  - Open-source code under permissive license (e.g., MIT, Apache 2.0)
  - Demonstration of algorithmic implementation and integration into simulated environment, e.g., Webots, Gazebo or NVIDIA Isaac.
  - Use of a Bayesian algorithm
  - Deployment on one of the HW types hardware introduced in the deployment section.
- **Preferential:**
  - Implementation using the Pymdp framework
  - Implementation in embedded C, Verilog/VHDL, or CUDA
  - Integration into an agentic system with action loop

EXPECTED RESULT (target KPIs):

- Latency reduction of at least 2x compared to baseline implementation
- Preservation of inference accuracy within 5% margin
- Energy efficiency improvement of at least 30%

CHALLENGE 5. Towards a Unified Semantic and Hyperspatial World Model: Enabling Smart Edge Systems through Domain-Aware Graphs and Agent Governance

SHORT DESCRIPTION:

As AI agents increasingly operate across cities, warehouses, and industrial environments, the lack of a shared semantic, hyperspatial, and governance infrastructure limits interoperability, traceability, and autonomous decision-making. This challenge proposes to develop an open-source, standards-aligned data infrastructure framework that enables the ingestion, contextualization, registration, and querying of multimodal data, agents, and environments in complex cyber-physical systems.

At the core of the solution is the development of a semantic infrastructure for representing entities, agents, policies, and processes using the IEEE Hyperspatial Modeling Language (HSML). Incoming data and autonomous agents are semantically described with type, provenance, behavioral constraints, and operational context. These descriptions enable agents to be dynamically discovered, reasoned about, and acted upon. Governance is enforced through the Hyperspatial Transaction Protocol (HSTP), which manages secure, policy-aware interactions, ensuring that agents operate within defined roles, permissions, and trust frameworks across distributed environments.

The project shall deliver infrastructure components for:

- Ingestion and semantic translation of heterogeneous data sources (e.g., IoT sensors, robot interfaces, edge logs) into HSML.
- Semantic registration and alignment of domain-specific entities, agents, and capabilities.
- Generation of graph-based representations using HSML-compliant schemas.
- Hyperspatial modelling to enable any data set to be normalized and queryable.
- Federated, policy-aware semantic search and agent coordination and governance across distributed edge environments.

Development or integration of agents (LLM-based, VERSES Genius, or PyMDP) operating in a simulated or real environment.

This work builds on the recently approved IEEE P2874 Spatial Web standards and serves as a foundation for secure, interoperable, and intelligent multi-agent systems at the edge.

EXPECTED DELIVERABLE:

Implementation:

4. An open-source ingestion pipeline for multimodal data, including telemetry, spatial data, logs, and operational interfaces.
5. Domain registration process for semantic entities and agents using HSML.
6. Federated query APIs supporting spatial, semantic, temporal, and policy-bound search.
7. Development or integration of available agents that can query its environment, understanding its objects, other agents, spaces, and the context. The agent shall operate and take action while adhering to the environment's rules and policies. Agents can be LLMs, robots or other kinds of agents that reason.
8. Continuous updating of the domain graph representations by agents that operate within it, contributing to a real-time representation of the environment queryable by all stakeholders.

Deliverable:

The final implementation shall contain the following infrastructure/testbed:

1. LLM or Active Inference agents for domain-specific environments deployed as a node on two different edge devices (e.g., Raspberry Pi, Nvidia Jetson, robot, drone).
2. Reference datasets with raw and HSML-annotated forms for each domain.
3. Queryable domains for semantic, data, and hyperspatial information.

4. Use cases to demonstrate the capabilities of the system developed.

Evaluation

1. Deployments in representative simulated or real-life domains: smart cities (traffic, sensors), warehouses (inventory robotics), factory floors (industrial control).
2. Evaluation metrics: semantic precision, ingestion throughput, agent discovery time, policy compliance, and ontology alignment.

Documentation

- User documentation for the reference implementation.
- Full pipeline technical report, including graph schema templates, ingestion flows, PI specifications for domain queries, agent discovery, transaction negotiation, and agent registration models.
- Open contribution guidelines for new domain ontologies, agent classes, and search/query extensions.

REFERENCE INFO:

Standards & Protocols:

- Spatial Web Foundation Web Site
- IEEE P2874 Working Group
- IEEE P2874 D3.1 Spatial Web Specification
- Introduction to Spatial Web Protocol and Governance
- The Future of Global AI Governance

Spatial Web Specifications

- Spatial Web Reference Implementation SDK
- HSML Validator
- SHACL – Semantic constraint validation
- HSTP Implementation Specification
- SWIDs and SWID documents
- DID SWID Method
- W3C DID – Decentralized Identifiers for domain identity

Agents

- VERSES Genius agent
- PyMDP agents

Sample Datasets:

- Energy grid and climate datasets (CityLearn, HVAC logs).
- Spatial Web SDK reference implementation providing example implementations and data (available Q3 2025)
- Smart home demo use case (available Q3 2025)
- Warehouse robotics operations demo use case. (available Q4 2025)

DIAGRAM & TECHNICAL SPECS:

Input Spec:

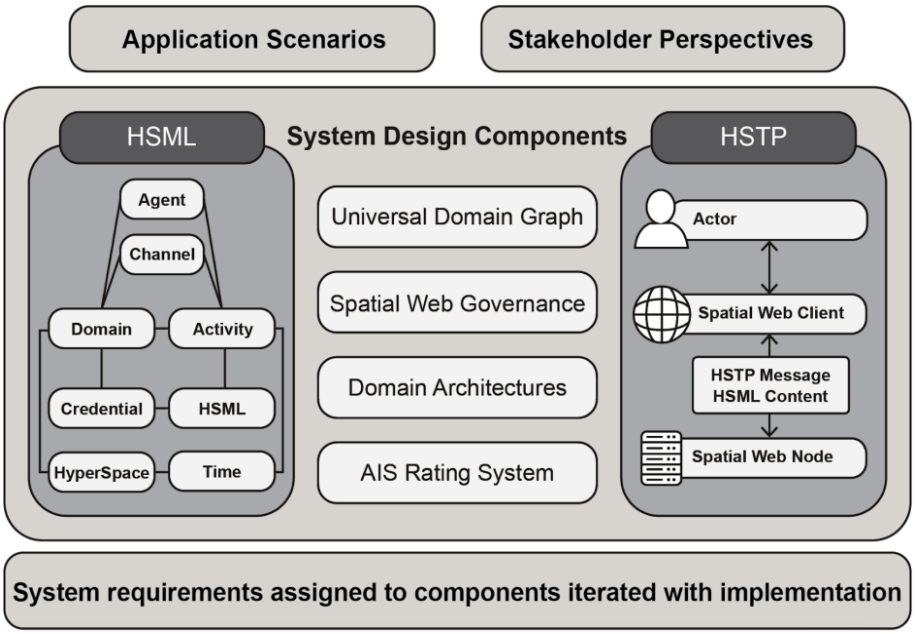
- Entity & agent models (types, activities, roles, permissions, behaviors)
- Time-series sensor telemetry (numeric, categorical)

- Structured logs: JSON, CSV, XML
- Metadata: device ID, location, capabilities

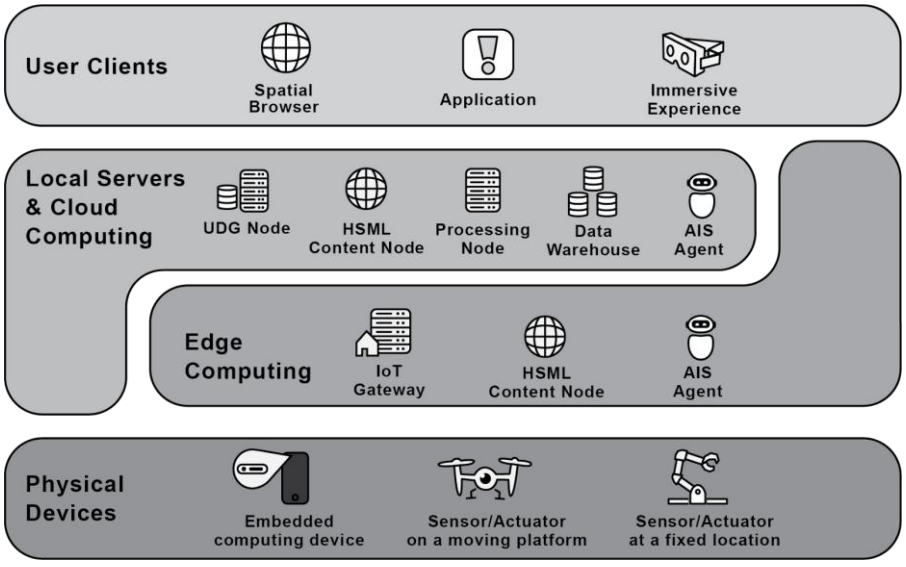
Output Spec:

- HSML-formatted representations of entities, environments, agents and hyperspaces
- Domain-registered graphs and federated query endpoints
- Policy-bound search and discovery APIs for multi-agent environments

System Overview



Computing Continuum Diagram



You can find an extended system description here

REQUISITES:

- **Mandatory**:
  - Open-source code under permissive license (e.g., MIT, Apache 2.0)
  - Demonstration of algorithmic implementation and integration into simulated environment, e.g., Webots, Gazebo or NVIDIA Isaac.
  - Use of a Bayesian algorithm
  - Deployment on one of the HW types hardware introduced in the deployment section.
- **Preferential:**
  - Implementation of an Active Inference based agent

EXPECTED RESULT:

- Ingest and semantically register <1000 data types across 2 environments
- Ingestion and successful querying of hyperspatial data
- Successful integration or implementation of at least two agent using governance framework
- Successful deployment of the same search API across at least 2 domains.
- Compliance with obligations imposed by the EU AI act, if applicable.

CHALLENGE 6. Automated Edge deployment, tuning and Performance Evaluation of Binary Neural Networks targeting FPGA enabled platform

Binarization is essentially a form of extreme (1-bit) quantization of neural networks values of weights, activations or both. This compression technique allows heavy floating point math operations to be replaced by lightweight bitwise XNOR and Bitcount operations. As a result, binary neural networks offer several hardware-friendly advantages that fits well with the constraints of embedded and edge platforms: reduced memory consumption, improved power efficiency, and significant speed enhancements.

The pioneering research results on BNN and XNOR-Net networks have demonstrated the effectiveness of binarization, achieving up to 32× memory savings and 58× speed improvements on CPUs for 1-bit convolution layers. This has sparked extensive research in computer vision and machine learning, with applications in tasks like image classification and detection. Additionally, binarization allows for easy validation of a layer's importance by switching between full-precision and 1-bit representations.

FPGA based hardware is natively suited for BNN implementation from resources perspective (i.e. configurable logic operators) however the design, implementation, validation and integration process is relatively lengthy and tedious.

CETIC is developing and edge middleware (DMWay) aiming at enhancing the productivity of edge AI computation developers by providing means to quickly configure and interface these computations, the associated In/out data streams and their set-up and execution parameters. It also enables a better automation of bechmarking and performance evaluation scenarios.

The objective of this challenge is to develop, prototype and evaluate a dedicated approach for the automation of BNN deployment, tuning and performance evaluation on Zynq FPGA platform, leveraging the DMWay middleware capabilities.

The challenge will proceed along the following steps:

4) Selection of a suitable target set of BNN models (and data sets) for the challenge based on the SotA literature on the subject, the availability of exploitable baseline or reference implementations for Zynq FPGA and the level of “tunability” of these models.
4) Define relevant, generic and reusable scenarios for the automation of performance evaluation and parameters exploration of these models execution on the FPGA targets, leveraging the Pynq environment (Python for Zynq) for Zynq FPGA
4) Set-up, develop and showcase the automation process implementation on different BNNs, different hardware targets, leveraging DMWay middleware capabilities. Direct and continuous support from the middleware team is guaranteed.
4) Based on the previous steps, define a reusable set of BNN related primitives and operations that could be automated, monitored and controlled by the middleware

EXPECTED DELIVERABLE:

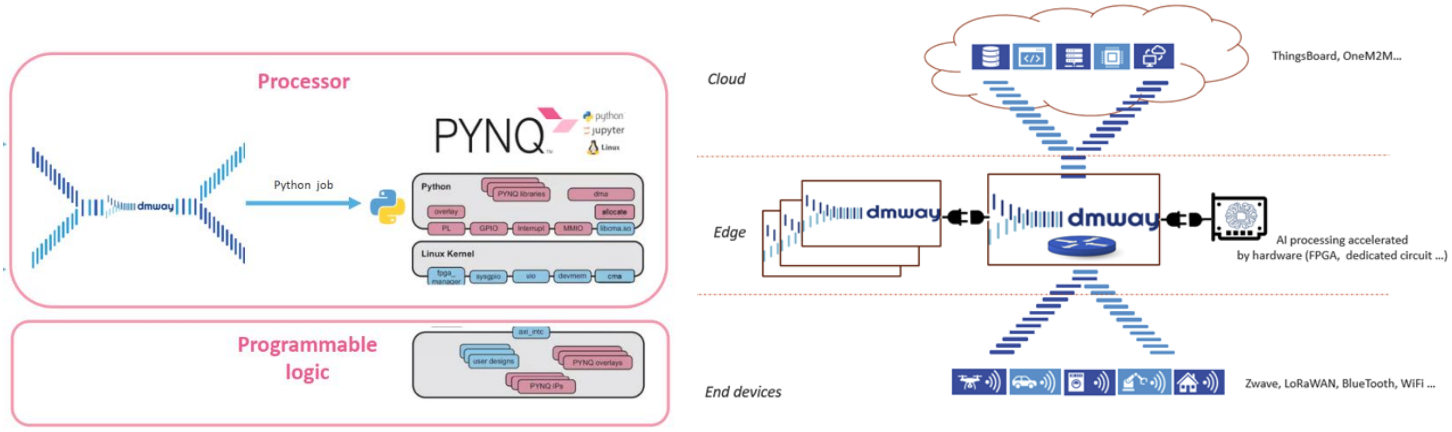
- Working hardware implementations of selected BNNs on Zynq FPGA
- Definition of automation scenarios for deployment and evaluation of BNNs on Zynq FPGA leveraging DMWay middleware
- Demonstrator showcasing the forementioned scenarios and the automation capabilities developed

REFERENCE INFO:

- BNNs : Survey (2020) Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, Nicu Sebe, Binary neural networks: A survey, Pattern Recognition,Volume 105,2020, 107281, ISSN 0031-3203,https://doi.org/10.1016/j.patcog.2020.107281. https://www.researchgate.net/publication/339411073_Binary_Neural_Networks_A_Survey
- Pynq framework for Zynq FPGA : https://www.pynq.io/
- DMWay middleware:

- L. Deru, A. Achour and L. Guedria, "Multi-protocols and Data Manager for IoT Gateways: A smart-building use-case demo," *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, Bordeaux, France, 2021, pp. 722-723.
- https://asset.cetic.be/en/dmway/

DIAGRAM & TECHNICAL SPEC:



REQUISITES:

- Python programming
- Decent practical FPGA/embedded computing experience
- AI & neural networks background

EXPECTED RESULT:

- Higher level of automation of BNNs implementation and qualification on FPGA Zynq platforms
- 3+ implemented and showcased tuning scenarios of BNNs on FPGA using DMWay middleware

CHALLENGE 7. Towards Automatic DMWay Configuration with Generative AI

DMWay is CETIC's middleware for data flow management. Its main objective is to facilitate the integration of IoT systems by providing implementations of various communication protocols for data sources, as well as seamless connectivity with user backend applications. Additionally, DMWay handles data semantics and supports task orchestration.

Building an IoT architecture with DMWay typically involves configuring the system through a graphical interface. The challenge proposed by CETIC is to enable the automatic generation of DMWay configurations using a Generative AI algorithm. This configuration could be created within a sandbox environment, allowing validation tests to be conducted in isolation before deployment.

Such an approach would represent a significant improvement to the DMWay middleware.

EXPECTED DELIVERABLE:

a. A first version of an executable tool, or script, or API which takes high-level user input and outputs a valid DMWay configuration
b. Environment for testing
c. Documentation : technical report, evaluation report

REFERENCE INFO:

Current version of DMway software

EXPECTED RESULT:

Simple configuration generator showing the core: 1 data source + 1 data field + 1 database

More complex configurations in terms of data sources, data model

Task orchestration configuration

a. Demonstrator on embedded platform

CHALLENGE 8. Health Monitoring AI for Edge Devices with Privacy-First

This challenge invites participants to design an AI application for edge devices—such as smartphones, wearables, or IoT health sensors—that enables real-time health monitoring by analyzing physiological and movement data like heart rate, physical activity, and sleep patterns. The primary goal is to detect early signs of health issues, such as stress or irregularities, offering users timely insights while maintaining strict data privacy and energy efficiency. Incorporating federated learning (FL) will enhance the system's adaptability and learning potential across multiple devices without compromising individual data privacy.

- Inference : Cardiac sensor + DMWay + RPI (on Edge with embedded AI)
- Training : Secure FL with RPi in the Edge + DMWay

Key elements for this challenge include:

- **On-Device Data Processing for Privacy**: The application must process all health data directly on the device (RPi, ...), without relying on cloud computing. This local data processing approach not only enhances user privacy by preventing data transfer to the cloud but also reduces the risk of sensitive information leaks, providing a secure solution suitable for handling personal health metrics.
- **Proactive Health Monitoring**: The AI should effectively analyze biosignals and detect early indicators of stress, health abnormalities, or other potential health risks. By providing early warnings or insights, the application can encourage proactive health management, empowering users to take action sooner and potentially preventing more serious health issues.
- **Energy Efficiency for Sustainable Use**: The application should be optimized for low power consumption. Participants are encouraged to explore efficient AI model designs, such as lightweight neural networks, model quantization, and energy-efficient algorithms, to ensure that the application can run sustainably on devices with constrained power resources.
- **Personalization and Adaptability**: The solution should adapt to each user's unique health patterns, allowing for personalized insights that accommodate normal variations in health metrics. This adaptability will improve the accuracy of detections and create a more meaningful user experience tailored to individual needs.
- **End-to-End Data Encryption:** Ensure all health data is encrypted both in transit and at rest on the device. Implementing strong encryption standards (such as AES-256 or ChaCha20) for stored data and TLS for data in motion can secure information from potential interception or tampering.
- **Federated Learning (FL) for Privacy-Enhanced Model Training:**
    - Local Model Training: FL allows the AI models to be trained locally on each device using user-specific health data, meaning sensitive information never leaves the device. After local training, only model updates (rather than raw data) are shared, ensuring that health metrics and patterns remain private.
    - Aggregation on Secure Server: These updates are sent to a central server where they are aggregated to improve the model across all devices without compromising individual privacy. This decentralized approach enhances the AI's ability to learn from multiple users while maintaining strict data privacy.
    - Differential Privacy and Secure Aggregation: Integrating differential privacy into FL can further obscure individual contributions by adding noise to the local updates. Secure aggregation methods can also ensure that model updates cannot be individually traced back to any single device.

References:

- DMWay - a platform for interconnecting heterogeneous IoT data https://www.cetic.be/DMWay-plateforme-d-interconnection-de-donnees-IoT-heterogenes
- RAMi: A New Real-Time Internet of Medical Things Architecture for Elderly Patient Monitoring https://www.mdpi.com/2078-2489/13/9/423
- Secure federated learning applied to medical imaging with fully homomorphic encryption, https://ieeexplore.ieee.org/document/10433836
- SecureBFL : a Blockchain-enhanced federated learning architecture with MPC https://www.cetic.be/SecureBFL-a-Blockchain-enhanced-federated-learning-architecture-with-MPC

CHALLENGE 9. Edge AI framework for web browsers empowering Federated Machine Learning applications addressing Classification problems over multiple format data (text and images

Federated learning (FL) is a decentralized approach to training machine learning models that gives advantages of privacy protection, data security, and access to heterogeneous data over the usual centralized machine learning approaches. We can obtain more accurate and generalizable models through FL without having the data leave the client devices. The three main strategies to perform FL are Centralized FL, Decentralized FL, and Heterogeneous FL with popular FL algorithms such as FedSGD, FedAvg, and FedDyn.

Centralized federated learning requires a central server. It coordinates the selection of client devices in the beginning and gathers the model updates during training. The communication happens only between the central server and individual edge devices. While this approach looks straightforward and generates accurate models, the central server poses a bottleneck problem—network failures can halt the complete process. Decentralized federated learning does not require a central server to coordinate the learning. Instead, the model updates are shared only among the interconnected edge devices. The final model is obtained on an edge device by aggregating the local updates of the connected edge devices. Heterogeneous federated learning involves having heterogeneous clients such as mobile phones, computers, or IoT (Internet of Things) devices. These devices may differ in terms of hardware, software, computation capabilities, and data types.

Federated ML Framework - The participants should develop a pioneering framework leveraging federated learning (FL), WebAssembly (Wasm), and WebGPU to facilitate training and deploying deep neural networks (DNNs) directly in web browsers. By combining FL's privacy-preserving distributed learning approach with Wasm and WebGPU's performance capabilities, we aim to create a powerful, accessible, and secure environment for real-time machine learning on client-side devices. This framework will enable decentralized model training without centralizing sensitive data, making it particularly suited for applications where data privacy and security are paramount.

Use case development - This framework can be valuable for building applications in areas such as: 1) *Healthcare Diagnostics:* Real-time image and signal analysis for remote health diagnostics without sharing patient data externally. 2) *Personalized E-Learning:* Adaptive learning systems that adjust in-browser educational content in real-time based on user interaction. 3) *Smart Cities:* Decentralized processing for edge devices, enhancing urban management services, such as traffic monitoring, without centralized data storage. 4) *IoT Device & Sensing Optimization*: Localized model training on IoT-enabled / networked sensor devices, ensuring data remains on-device while enhancing capabilities such as predictive maintenance, event forecasting and identification. 5) *Augmented Reality and Virtual Assistance:* Real-time personalization and contextual awareness in AR/VR environments based on user interaction patterns, all executed within the browser. One such test application will be developed (an existing use case can be adapted appropriately) that will leverage federated learning based on local user or other agent type interactions over the browser. Decentralized approach ensures that users retain full control of their data while benefiting from a truly personalized and contextual user experience that dynamically adapts to foster engagement and improve specific outcomes for each use case. FL framework built for computer vision applications will be utilised, combined with AI/NLP resources where needed.

Despite the ongoing research on scaling FL systems, certain limitations still need to be addressed. It's necessary to improve communication efficiency, protect data privacy, and incorporate the heterogeneity present at systems and statistical levels.

EXPECTED DELIVERABLE:

a. Report on the FL framework architecture
b. Software implementation of the enhanced browsers
c. Model Trained using the Federated ML framework
d. Development of an application use case utilizing the Federated ML framework

REFERENCE INFO :

[1] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, Mar. 2021, doi: 10.1016/j.knosys.2021.106775.

[2] S. Kakati and M. Brorsson, "WebAssembly Beyond the Web: A Review for the Edge-Cloud Continuum", 2023 3rd International Conference on Intelligent Technologies (CONIT), Karnataka, India. June 23-25, 2023

https://orbilu.uni.lu/bitstream/10993/57810/1/WebAssembly_Beyond_the_Web_A_Review_for_the_Edge-Cloud_Continuum.pdf

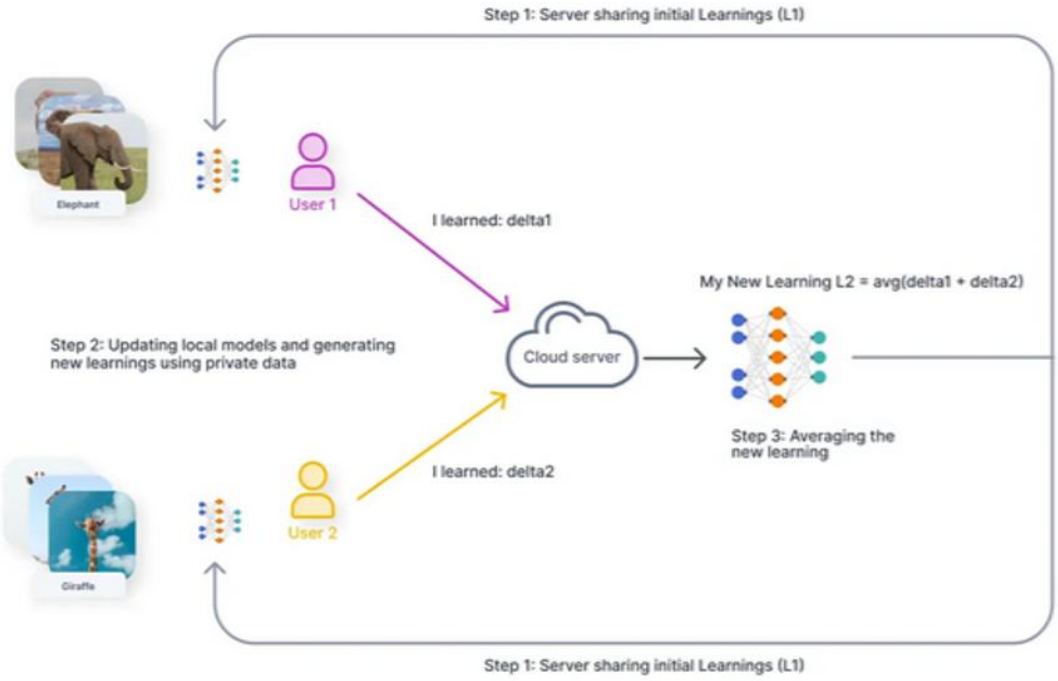
[3] ONNX Runtime Web (https://onnxruntime.ai/)

[4] Apache TVM for WebGPU (**https://tvm.apache.org/**): Apache TVM, a deep learning compiler, now supports WebGPU and WebAssembly, making it possible to achieve near-native GPU performance within browsers.

Datasets: https://universaldependencies.org/

DIAGRAM & TECHNICAL SPEC:

A generic baseline model is stored at the central server. The copies of this model are shared with the client devices, which then train the models based on the local data they generate. Over time, the models on individual devices become personalized and provide a better user experience.



In the next stage, the updates (model parameters) from the locally trained models are shared with the main model located at the central server using secure aggregation techniques. This model combines and averages different inputs to generate new learnings. Since the data is collected from diverse sources, there is greater scope for the model to become generalizable.

Once the central model has been re-trained on new parameters, it's shared with the client devices again for the next iteration. With every cycle, the models gather a varied amount of information and improve further without creating privacy breaches.

REQUISITES:

- Machine Learning
- NLP

EXPECTED RESULTS:

Benchmarking of the federated ML framework, in terms of:

- Flexibility in terms of use case development
- Performance
- Scalability
- Accuracy

CHALLENGE 10. Off-Chip Weights for Streaming Architectures

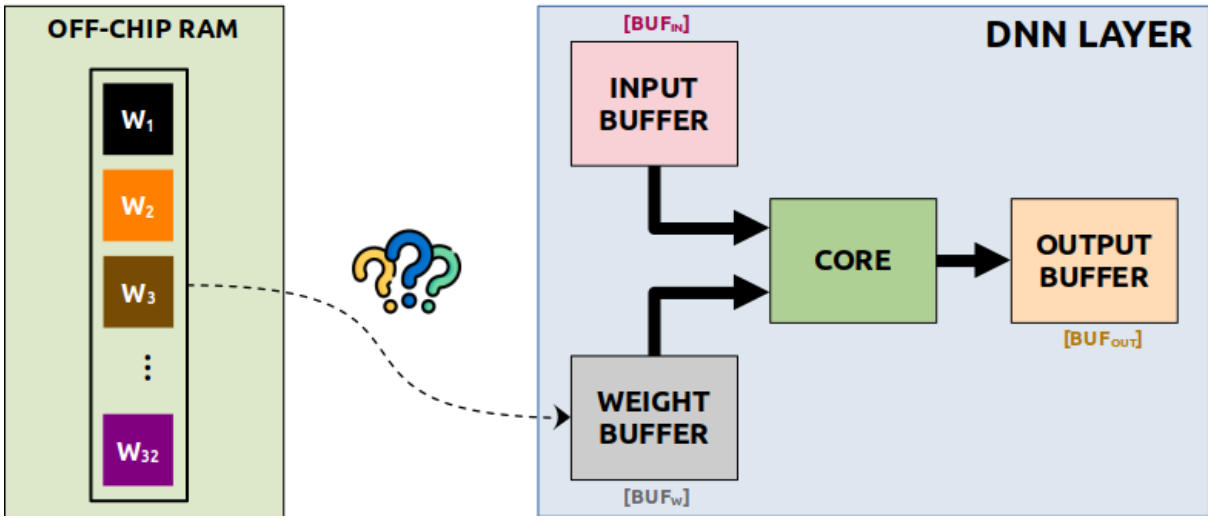
SHORT DESCRIPTION:

It is often challenging to fit large models directly on-chip due to their substantial memory requirements. System-on-Chip (SoC) architectures provide a flexible solution, enabling customized data handling and processing to accommodate such large networks. Streaming architectures, a popular approach in SoC designs, help manage data flow efficiently, allowing for continuous data processing rather than batch loading, which can minimize latency and optimize bandwidth usage. When the entire neural network cannot be stored on-chip, off-chip memory becomes essential for storing large sets of model weights. By strategically utilizing off-chip memory and transferring only the required weights to the on-chip processors when needed, streaming architectures ensure that data movement is minimized and managed efficiently. This approach maximizes processing throughput while keeping power consumption low, addressing the constraints of both hardware resources and real-time processing demands.

The **objective** of this challenge is to determine the optimum **size** and **order** of parameter chunks that should be transferred from the off-chip memory onto the on-chip fabric so that the latency is minimal, and the energy inefficiency is avoided.

EXPECTED DELIVERABLE:

- An algorithm that manages off-chip weight storage and dynamically loads only required weights onto on-chip memory, minimizing latency and optimizing memory usage.
- The algorithm should be benchmarked on common deep neural networks such as MobileNet, ResNet and AlexNet.
- Comprehensive documentation outlining the algorithm design, memory access strategy, setup instructions, and potential areas for further optimization.



CHALLENGE 11. 2D Vehicle Detection Network on PYNQ-Z1

This challenge aims to implement a real-time vehicle detection network on the PYNQ-Z1 board using a compact convolutional neural network (CNN) optimized for the resource constraints of the FPGA-based platform. Leveraging the flexibility of the PYNQ-Z1's System-on-Chip (SoC) architecture, the network will be partitioned to run both on the ARM Cortex-A9 processor and on the programmable logic (FPGA) for acceleration. The network could use off-chip memory efficiently to store model weights, ensuring seamless data transfer through DMA channels and optimizing detection speed for real-time applications.

EXPECTED DELIVERABLE:

- A quantized (and pruned) vehicle detection CNN model (e.g., YOLO or MobileNet SSD), tailored for deployment on the PYNQ-Z1.
- A quantized hardware-accelerated CNN implemented as custom IP cores or using Xilinx's Vitis AI library deployed on the PYNQ-Z1.
- A complete software stack on PYNQ, including model partitioning, DMA for A functional real-time detection demo showcasing real-time vehicle detection on live video input or sample images, running on the PYNQ-Z1 board.
- Comprehensive documentation covering the setup, model optimization, hardware-software integration, and usage instructions for deploying the vehicle detection network on PYNQ-Z1.

Datasets

COCO, KITTI Vehicle Detection, or PASCAL VOC.

Image source: https://www.mouser.mx/blog/mystery-of-vehicle-detection

CHALLENGE 12. Real-Time Vision Transformer (ViT) on FPGA SoC for Image Classification

This challenge focuses on deploying a Vision Transformer (ViT) model on an FPGA SoC to perform image classification tasks. Vision Transformers leverage attention mechanisms instead of traditional convolutional layers, making them ideal for handling visual tasks that require global context. Implementing ViT on an FPGA will involve optimizing the model's matrix operations, attention layers, and patch embeddings for efficient FPGA execution. This approach is designed to accelerate processing while maintaining model accuracy, making it suitable for real-time image classification on edge devices.

EXPECTED DELIVERABLE:

- Optimized Transformer Model: A streamlined, quantized version of the Transformer model, with attention layers and feedforward networks tailored to the FPGA SoC architecture.
- Custom Hardware Modules: FPGA-accelerated modules for matrix multiplication, attention mechanisms, and softmax, designed as reusable IP cores.
- Efficient Memory Management: Implementation of an optimized memory transfer system (e.g., using DMA) to handle large token embeddings and model weights stored off-chip, with data streaming to the FPGA.
- Demo Application: A working demo showcasing the Transformer model performing tasks like text classification or summarization on a sample input dataset.
- Documentation: Detailed documentation on model optimization, hardware design, memory management strategy, and usage instructions for deploying the Transformer on FPGA SoC.

Datasets

CIFAR-10, ImageNet Subset, or Tiny ImageNet.

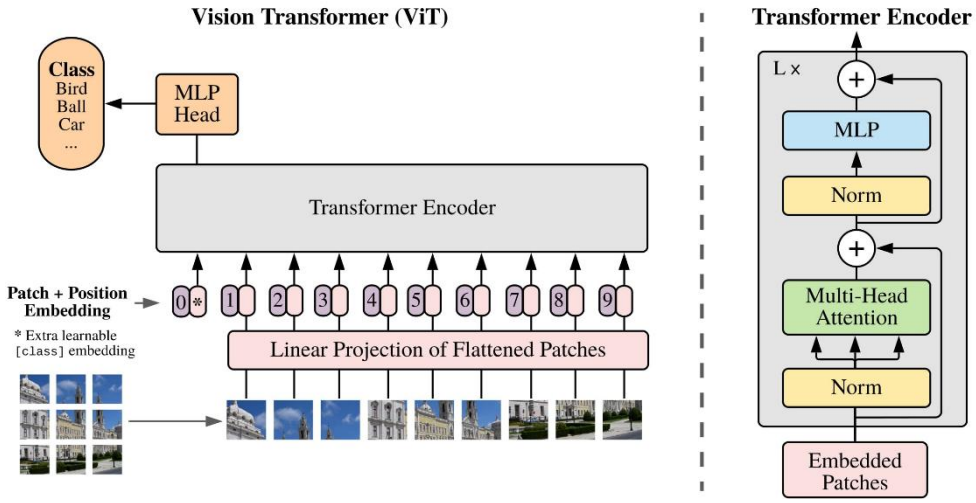


Image source: https://arxiv.org/pdf/2010.11929

CHALLENGE 13. Generative AI for object detection in Edge scenarios

In the context of computer vision, Generative AI has progressed to the point of generating synthetic but highly realistic images that can potentially be used for training (partially or fully) vision models. On the other hand, for the specific problem/task of object detection, there is always a background on which we have to detect the present (and location) of one or more objects of interest. If we consider an edge vision system (like a smart camera), we can capture an indefinite amount of those background images. In fact, each vision unit can be placed in different rooms, corridors, etc., and we can get indefinite amounts of background images for each position. This challenge aims at studying the generation/insertion of synthetically generated objects on those background images and assessing the effect on fine-tuning object detectors with the new 'semi-synthetic' images.

EXPECTED DELIVERABLE (Report, Software, Hardware, etc): The challenge involves the preparation of a brief survey on the most recent work related to the idea just described. Besides, software and models are expected at the end of the project, along with a descriptive report. The challenge proposer aims at publishing the results of the study with the beneficiary through co-authored publications in scientific conferences or journals.

REFERENCE INFO (datasets, AI/SW/HW models, benchmarks) : The task of object detection is a classic in computer vision and there are several public datasets available for multiple applications (generic object detection *á la* ImageNet, person detection, face detection, car detection, weapon detection, etc.). In any case, the images in the dataset chosen will need to have a single common background (or a few different backgrounds).

DIAGRAM & TECHNICAL SPEC (input spec incl. pre-processing/ output spec incl. post-processing, global picture of the challenge): The workflow for the experimental work should be based on a previously existing baseline detector. From some of the images of the dataset, we would generate semi-synthetic images containing the newly inserted examples of the object of interest. With these additional semi-synthetic images (they will have to be labelled, either manually or in some automatic form) we would fine-tune the baseline detector and will then get performance metrics of the new model, comparing them with those of the baseline detector.

REQUISITES:: A background on computer vision is required.

EXPECTED RESULT (target KPIs): The main KPI to consider will be the typical object detection metrics (like IoU), especially in relative terms to the pre-existing baseline detector. Several factors will have to be considered: methods used to generate the semi-synthetic images (considering their computational cost, unintended artifacts or changes introduced to the background, possibility of generating the object with different poses, etc.), ways to annotate the semi-synthetic images, etc.

The dAIEdge-VLab Description

The **dAIEdge-VLab** aims to implement a collaborative platform that enables Researchers and Developers to conduct experiments and research across various edge AI domains and edge AI devices. This will be achieved by sharing resources on a distributed virtual lab.

The dAIEdge-VLab is designed to help users who lack expertise in embedded programming or do not have direct access to specific hardware. It will enable them to conduct real-time AI experiments on a remote farm of embedded boards. Currently, it allows: (i) the launch of AI models benchmarking using randomly generated data; (ii) the launch of AI APPs benchmarking with preprocessed data; and (iii) the launch of on-device training experiments.

In terms of hardware compatibility, the dAIEdge-VLab is built to support a wide range of embedded boards, spanning from high-performance MPUs and GPGPUs to energy-efficient MCUs and specialized NPUs. On the software side, it accommodates both Linux-based and real-time operating systems, as well as bare-metal solutions. Additionally, the platform will be compatible with widely used vendor-agnostic inference runtimes along with proprietary AI engines.

A virtual lab for online benchmarking of edge AI applications typically consists of several architectural components designed to facilitate remote experimentation, testing, and optimization of AI models across different edge hardware platforms. Below is an outline of its overall structure together with the main functionalities:

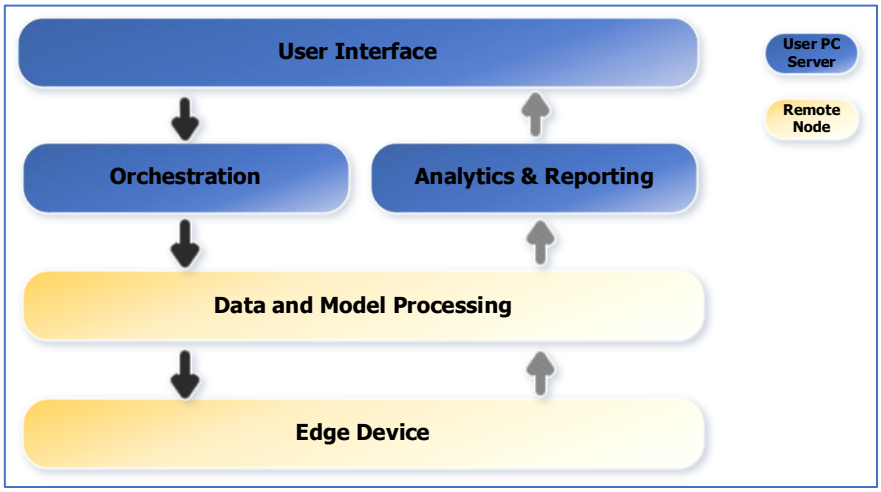


Figure 1. dAIEdge-VLab five-layer model

1. User Interface (UI) Layer

- **Web Interface/Dashboard**: The virtual lab provides a user-friendly web-based dashboard that allows data scientists and developers to interact with the system. This includes uploading AI models, selecting target hardware platforms, configuring benchmarking parameters, and monitoring results.
- **CLI API**: An API-based interface allows a seamless integration with custom AI pipelines.

2. Orchestration Layer

- **Resource Management and Scheduling**: This component handles the orchestration of resources across multiple platforms, ensuring efficient allocation of hardware for testing and benchmarking. It manages queues, schedules jobs, and optimizes resource usage based on platform availability.

- **Virtualization and Containerization**: Models and benchmarking tasks are often containerized (e.g., using Docker) to ensure compatibility across diverse hardware. This enables easy deployment across various edge devices, regardless of their underlying operating system or architecture. This layer deals with the management of containers on edge devices running general purpose operating systems.
- **Benchmark Configuration**: Users define test cases, including hardware configurations, datasets, and performance metrics such as latency, power consumption, and accuracy. The ultimate goal is to define the technical requirements to compare benchmarking results among different configurations.

3. Edge Device Layer (Board farm)

- **Multi-Platform Hardware Pool**: The system integrates a variety of edge devices, ranging from low-power IoT devices (like STM32 MCUs, RISC-V based platforms, etc.) to high-performance systems (such as Raspberry Pi or Jetson Orin Nano or AGX). This diversity ensures comprehensive testing across different edge environments.
- **Remote Access & Control**: The lab allows remote access to real, physical hardware. Developers can deploy AI models directly to these devices for testing under real-world conditions, avoiding the limitations of simulated environments.

4. Data and Model Processing Layer

- **Data Preprocessing and Inference**: This layer handles the preprocessing of input data and manages the inference execution on edge devices. It ensures that the models are efficiently adapted to the target hardware's limitations, such as memory and computational power.
- **Performance Metrics Collection**: During benchmarking, this layer monitors critical performance metrics like inference time, throughput, power consumption, and memory usage, which are fed back into the system for analysis.

5. Analytics & Reporting Layer

- **Results Analysis & Visualization**: The benchmarking results are processed and presented through interactive dashboards, allowing users to compare metrics across different hardware platforms. Detailed reports include insights into energy efficiency, processing latency, accuracy, and other relevant performance factors.

The proposed dAIEdge-VLab platform features a modular architecture that allows for easy integration of remote embedded boards.

Figure 2 illustrates the overall architecture of the dAIEdge-VLab. The key components of this architecture are:

- **Remote User**: A user initiating an AI experiment on a remote node submits a request via a web-based user interface or by integrating the CLI API in their own pipeline.
- **dAIEdge-VLab Server**: The dAIEdge-VLab Server receives the user request and forwards it to the corresponding Remote Host that is connected to the target Remote Node. All available Remote Nodes are registered to the dAIEdge-VLab Server.
- **Remote Host**: Located at the facility of the Remote Node owner, the Remote Host is responsible for executing node-specific scripts. It compiles and deploys the AI application to the Remote Node, retrieves the benchmarking results, and sends them back to the user through the dAIEdge-VLab Server.
- **Remote Node**: This refers to the embedded board where the AI experiments, such as model benchmarking, are executed. A single Remote Host can manage multiple Remote Nodes at the owner's site.

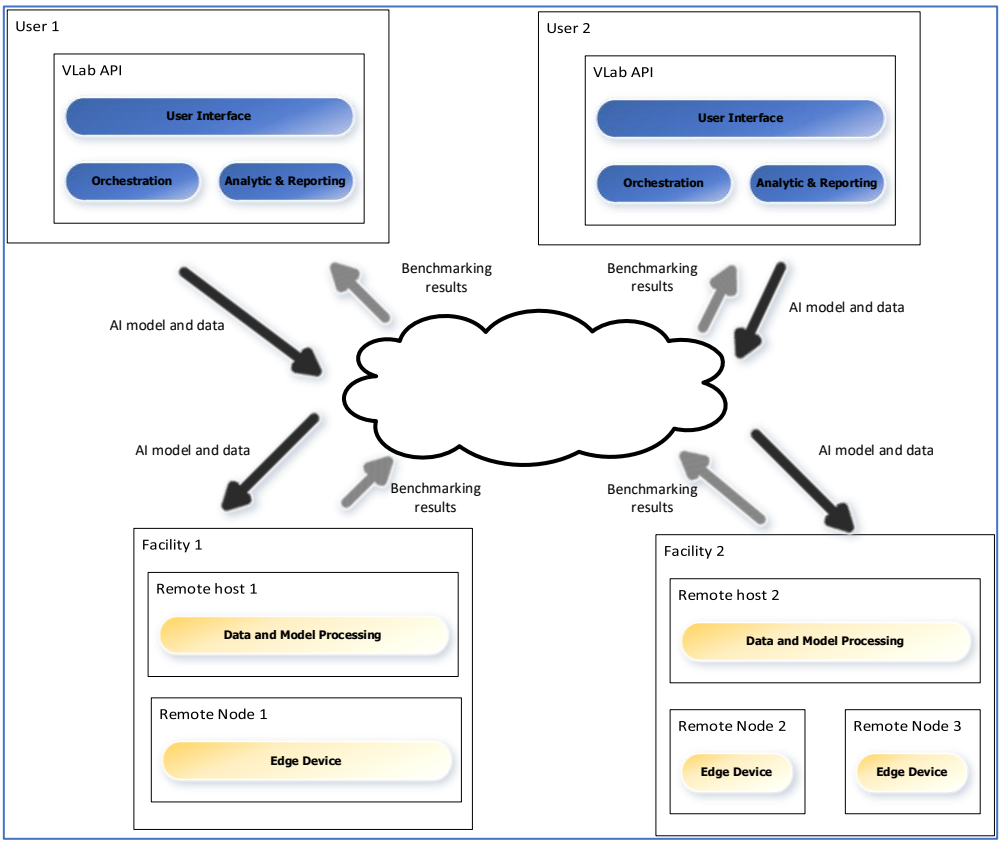


Figure 2. dAIEdge-VLab Architecture

Below are screenshots showcasing the web interface dashboards for both input selection (Figure 3) and visualization of benchmarking results (Figure 5). A history of previously executed experiments (Figure 4) is saved locally in the browser cache, allowing users to review and compare different benchmarking experiments in the future.

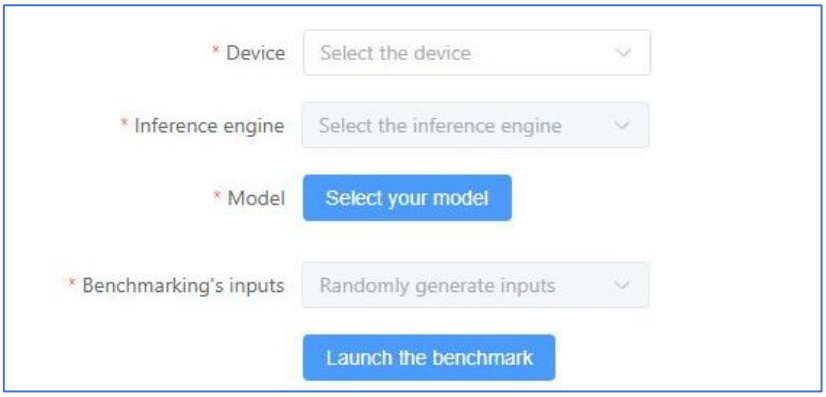


Figure 3. User Input Selection through Web Interface

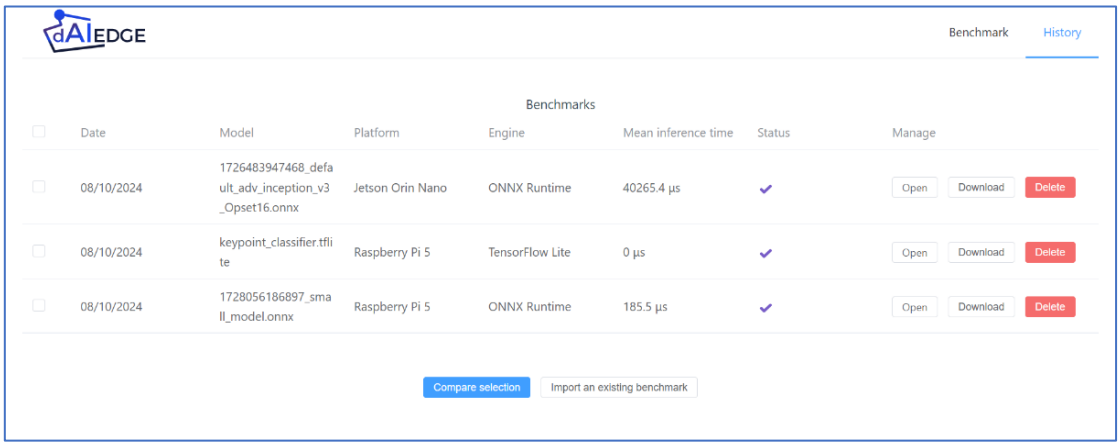


Figure 4. List of Benchmarking Results

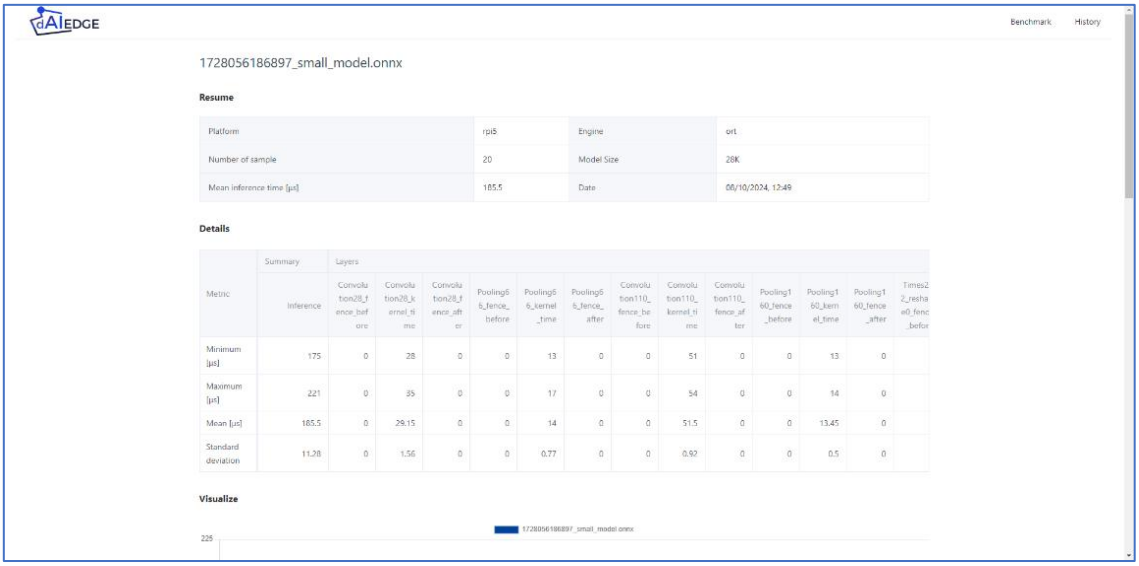


Figure 5. Visualisation of Model Benchmarking Results

The current implementation of **dAIEdge-VLab** supports a wide range of Linux-based MPUs, including Raspberry Pi 4B and Raspberry Pi 5, as well as GPGPUs such as Nvidia Jetson Orin Nano. It also accommodates MCU+NPU platforms like the STM32MP257 and the Qualcom RB3 Gen2, along with bare-metal MCUs such as STM32L4R9 and NXP LPC55S69. Additionally, it includes support for vendor-agnostic Machine Learning runtimes, such as ONNX Runtime, TensorFlow Lite, and TensorFlow Lite for microcontrollers, as well as vendor-specific Machine Learning runtimes like CUBE-AI and TensorRT.

dAIEdge-VLab currently assists users who may not have expertise in embedded programming or access to the target embedded board by enabling online model benchmarking on a remote embedded board (see above the list of supported boards).