Advanced Research + Invention Agency ARIA

Scaling compute: AI at 1/1000th the cost
Technical Area 4: Test and Evaluation
Request for proposals (RFP)

12 June 2024

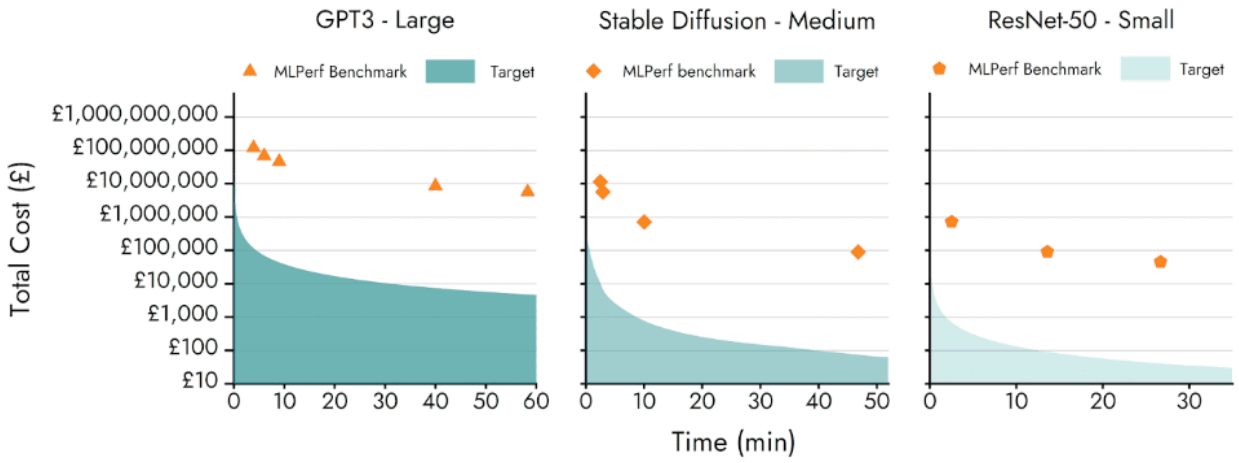
SECTION 1: PROGRAMME THESIS AND OVERVIEW

This Request For Proposals (RFP) is derived from the programme thesis Unlocking AI compute hardware at 1/1000th the cost and Nature computes better opportunity space.

As described in the solicitation for Technical Areas (TA) 1-3, the programme is designed to demonstrate that:

- It is possible to drop the hardware costs required to train large AI models by >1000x
- It is possible to do this *without* primarily relying on leading-edge fabrication facilities

All programme activities will be anchored around reducing the hardware costs required to train large-scale AI models. The initial **programme targets** are defined below, where we show the targeted time/cost pareto frontier to train three specific workloads from the MLPerf benchmark (to the quality level described in the benchmark).



During the delivery phase, **all programme activities in TA 1-3 will be evaluated based on ability to meet these targets.** Throughout the project duration, Creators from TA 1-3 will be asked to estimate the expected manufacturing and energy costs of their proposed solutions, and be tasked with justifying their estimates to ARIA.

In this RFP, we are looking for teams to submit proposals for TA 4 only, applications for TA 1-3 are now closed.

SECTION 2: TECHNICAL AREA OBJECTIVES

TA 4 represents the test and evaluation (T&E) component of the programme. Applicants for TA 4 will be asked to develop testing frameworks which can be used to evaluate Creator

outputs through the lifecycle of the programme. The programme targets defined statically above are, in reality, moving targets, and the primary task of the T&E team will be to continuously validate the figure above throughout the course of the programme and ensure that baseline workloads and targets stay relevant.

The primary objectives of the T&E team will be to:

- Qualitatively evaluate benchmark models on which programme goals are anchored.
- Define hardware cost models based on existing industry supply chains and reasonable sets of operating assumptions.
- Develop software infrastructure to benchmark industry-leading practices for training machine learning models.

T&E teams will be asked to assess runtime and accuracy when training ML models using commercially available hardware accelerators. They will then be asked to combine this information with comprehensive hardware cost modelling to estimate overall manufacturing and operating costs. Finally, they will also be asked to obtain baseline cost estimates for technologies in the research phase being pursued by Programme Creators in TA 1-3, and to compare these costs with those of commercially available alternatives.

Initial benchmark workloads will come from MLPerf, which is an industry organisation chartered to maintain relevant data, integrity, and relevance.

We are looking to fund this TA with up to £1m (inclusive of VAT and all other costs including overheads). We expect to fund one to two awards in this TA. Applicants recognise and accept that it will be at ARIA's sole discretion as to which, if any, proposal is accepted.

SECTION 3: What are we looking for/what are we not looking for

We are looking for Suppliers who can build software that allows them to evaluate the hardware costs required to train a defined set of ML models. Suppliers will be asked to start with openly-available data and code from established benchmarks (MLPerf) and make use of existing commercial or open-source offerings to establish cost benchmarks. Suppliers are encouraged to use (and discouraged from re-creating) software/services designed to streamline ML training on commercial hardware including (but not limited to):

- Mosaic ML
- Together AI
- SkyPilot

SECTION 4: Project Milestones and Project Management

Activity for this technical area will run for the first 3 years of the Scaling Compute Programme. The primary milestones will come in the form of technical reports, to be disseminated widely throughout the ARIA Creator, and general R&D, community. Reports should be published every 6 months throughout the lifecycle of the programme.

Project Milestones

Milestones for this technical area will consist of reports provided at regular intervals. An example of the content/cadence is shown below.

- Month 6 → Commercial hardware cost modelling (v1)
- Month 12 → ML Model training validation and hardware cost estimation (v1)
- Month 18 → Commercial hardware cost modelling (v2)
- Month 24 → ML Model training validation and hardware cost estimation (v2)
- Month 30 → ARIA Creator hardware cost modelling (v1)
- Month 36 → Final report: Comparison of commercial hardware to ARIA-funded technologies

Success/pivot/closure criteria for each project will be determined by the applicant's ability to meet these agreed-upon milestones.

Programme & Project Management

Alongside our standard project management requirements, the ARIA Programme Director (PD) will also monitor progress of each project through a series of 1:1 calls, site visits, and Programme-wide meetings. Project status updates are expected to be shared at quarterly intervals between ARIA and each Supplier.

SECTION 5: Application process & Eligibility

Application Process

The application process for Technical Area 4 consists of one stage.

Applicants are invited to set out how they propose to deliver the services outlined within this RFP. The format below is set out as a guide and represents a maximum length response.

Applicants are required to submit a detailed proposal including:

Section 1: Technical proposal

- **Your Approach -** to help us gain a detailed understanding of your proposal. This should include:
    + A detailed explanation of the proposed solution, how it supports the technical objectives of the requirement. This should be supported by technical information, visual aids and/or data, where relevant.
    + A comprehensive list of the known technical risks/unknowns standing in the way of achieving the objectives of the requirement.
    + Description of the proposed activity of work, key metrics and milestones and any dependencies and assumptions.

- **Background and Expertise** - to help us learn more about the team who will be delivering the project and their expertise. This should include:
    + Details of the project team, including a clear demonstration of experience and expertise in delivering test and evaluation services similar to the requirements outlined in Section 2.
    + If you intend to collaborate with or rely on any third parties, sub contractors/grantees, who they are and which elements of the project they will support/deliver.
    + Provide examples of your ability to provide the operational and executional resources required to deliver.

How to format your proposal:

- Page count: 4 pages, (including diagrams) single line spacing, standard character spacing (neither expanded nor condensed)
- Font: Arial. Colour: black. Size: 11-point font or larger
- Margins: At least 0.5" margins all around
- File Type: PDF

Section 2: Commercial proposal and administrative questions

In completing your application you must also provide answers to the following questions. Answers to these questions are not included in the 4 page cap. You should complete these questions in the application portal so there is no need to format these in a specific way.

This includes:

+ You should base your pricing on a time and materials basis
+ The proposed cost including a breakdown of costs. A short table is provided in the application portal (please ensure you account for VAT where applicable)
+ Any reliance on pre existing background IP (this includes third party data you may rely upon to deliver the service - please ensure any cost associated with accessing/purchasing this data is included in your cost breakdown)
+ Any other factors or restrictions that might impact your freedom to operate and deliver the project (such as conflicts of interest with the PD, import/export or security restrictions that you are aware of)
+ Details of any sub-contractors

Eligibility

We welcome applications from across the R&D ecosystem, including individuals, universities, research institutions, small, medium and large companies, charities and public sector research organisations.

This RFP is not seeking proposals for research projects, the requirement is to provide a service to the programme. Therefore, our review process of applications is different from the other TAs in the programme. For more information on the evaluation criteria for this RFP, see here.

Any resultant agreement from this RFP will be a contract for services.

The contract will be placed on mutually agreed terms and conditions (T&Cs) provided by ARIA to successful applicants. The proposed terms will include the following principles:

+ Either party shall have the right to terminate the contract or part of the contract for convenience upon ninety days' prior notice

- ARIA shall have the right to terminate the contract or part of the contract where the supplier fails to provide the service contracted. Upon thirty days' prior notice
- Any Intellectual Property generated in the performance of the contract shall vest in ARIA
- The supplier shall indemnify ARIA, its employees, officers and agents against the supplier's infringement of third party Intellectual Property Rights
- All information shared with the supplier shall be subject to confidentiality terms

Additionally, in order to carry out the assessment of the proposed technologies by the Programme Creators, you may be required to enter into confidentiality agreements with those Creators.

SECTION 6: Evaluation Criteria

Applicants will be evaluated on the comprehensiveness of the proposed methods to develop hardware cost models. This includes not just processor cores but interconnect subsystems, cooling, energy costs, and other critical aspects of AI hardware systems. As such, proposals will be evaluated against the following criteria:

- **Your Approach** - A clear articulation of what you see as our requirements as set out in the Technical Area Objectives section above, and how you would meet them including proposed technical solution and project plan
- **Background and Experience** - A clear demonstration of experience in delivering T&E projects
- **Demonstration of Resource** - Demonstrate that you have (or have access to) the operational and executional resources required to deliver
- **Commercial Proposal** - Commercial terms that demonstrate value for the tax-payer

SECTION 7: Timelines

This call for TA4 will be open for applications as follows (we may update timelines based on the volume of responses we receive):

Applications open	**12 June 2024**

Application submission deadline	**26 June 2024 (12:00 BST)**
Application review	**27 June - 11 July 2024**

As part of our review we may invite applicants to meet with the PD to discuss any critical questions/concerns prior to final selection — this discussion can happen virtually or we may seek clarification on certain aspects of your proposal via email.

Successful applicants notified	**12 July 2024**

At this stage you will be notified if you have or have not been selected for an award subject to due diligence and negotiation. If you have been selected for an award (subject to negotiations) we expect a 1 hour initial call to take place between ARIAs PD and your key team members within 10 working days of being notified.

We expect contract signature to be no later than 4 weeks from successful/ unsuccessful notifications. During this period the following activity will take place:

- Due diligence will be carried out
- The PD and the applicant will discuss, negotiate and agree the project activities, milestones and budget details
- Agreement to the set Terms and Conditions of the Contract.

SECTION 8: How to apply

Before submitting an application we strongly encourage you to read this call in full.

If you have any questions relating to the call, please submit your question to clarifications@aria.org.uk.

Clarification questions should be submitted no later than 3 days prior to the application deadline date. Clarification questions received after this date will not be reviewed. Any questions or responses containing information relevant to all applicants will be provided to everyone that has started a submission within the application portal.

Please read the portal instructions below and create your account before the application deadline. In case of any technical issues with the portal please contact clarifications@aria.org.uk.

Portal instructions

APPLY HERE